

CHIRA 2022

6th International Conference on Computer-Human
Interaction Research and Applications

PROCEEDINGS

27 - 28 October, 2022

EDITORS

Hugo Plácido Silva
Jean Vanderdonckt
Andreas Holzinger
Larry Constantine

<https://chira.scitevents.org>

SPONSORED BY



PAPERS AVAILABLE AT



CHIRA 2022

Proceedings of the
6th International Conference on
Computer-Human Interaction Research and
Applications

Valletta - Malta

October 27 - 28, 2022

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

ACM In Cooperation
SIGMIS - ACM Special Interest Group on Management Information Systems

In Cooperation with
EUSSET - European Society for Socially Embedded Technologies

Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda.
All rights reserved

Edited by Hugo Plácido Silva, Jean Vanderdonckt, Andreas Holzinger and Larry Constantine

Printed in Portugal
ISSN: 2184-3244
ISBN: 978-989-758-609-5
DOI: 10.5220/0000166000003323
Depósito Legal: 505700/22

<https://chira.scitevents.org>
chira.secretariat@insticc.org

BRIEF CONTENTS

INVITED SPEAKERS	IV
ORGANIZING COMMITTEES	V
PROGRAM COMMITTEE	VI
SELECTED PAPERS BOOK	VI
FOREWORD	VII
CONTENTS	IX

INVITED SPEAKERS

Abigail Sellen

Microsoft Research Cambridge
United Kingdom

Karen Holtzblatt

InContext Design
United States

Alan Dix

Swansea University
United Kingdom

ORGANIZING COMMITTEES

CONFERENCE CO-CHAIRS

Andreas Holzinger, Medical University Graz, Austria
Larry Constantine, Madeira Interactive Technologies Institute, Portugal

PROGRAM CO-CHAIRS

Hugo Plácido Silva, IT- Instituto de Telecomunicações, Portugal
Jean Vanderdonckt, Université Catholique de Louvain, Belgium

SECRETARIAT

Andreia Pereira, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

Inês Teles, INSTICC, Portugal

WEBMASTER

João Francisco, INSTICC, Portugal
Carolina Ribeiro, INSTICC, Portugal

PROGRAM COMMITTEE

Iyad Abu Doush, American University of Kuwait, Kuwait

Christopher Anand, McMaster University, Canada

Nizar Banu P. K., Christ University, India

John Brooke, Independent Researcher, United Kingdom

Eric Castelli, LIG Grenoble, France

Christine Chauvin, Université de Bretagne Sud, France

Cesar Collazos, Universidad del Cauca, Colombia

Arzu Coltekin, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

Lizette de Wet, University of the Free State, South Africa

Joerg Doerr, Fraunhofer Institute for Experimental Software Engineering, Germany

Vania Estrela, Universidade Federal Fluminense, Brazil

Jesus Favela, Cicese, Mexico

Peter Forbrig, University of Rostock, Germany

Gheorghita Ghinea, Brunel University London, United Kingdom

Toni Granollers, University of Lleida, Spain

Martin Hitz, Alpen-Adria-Universität Klagenfurt, Austria

M.-Carmen Juan, Instituto Ai2, Universitat Politècnica de València, Spain

Adi Katz, SCE - Shamoon College of Engineering, Israel

Wen-Chieh Lin, National Chiao Tung University, Taiwan, Republic of China

Federico Manuri, Politecnico di Torino, Italy

Daniel Mestre, Aix-Marseille University / CNRS, France

Lia Morra, Politecnico di Torino, Italy

Max North, Kennesaw State University, United States

Francisco Rebelo, LITI/LARSys and CIAUD, University of Lisbon, Portugal

Laura Ripamonti, Università degli Studi di Milano, Italy

Carsten Röcker, Fraunhofer IOSB-INA, Germany

Andrea Sanna, Politecnico di Torino, Italy

Markku Turunen, Tampere University, Finland

Frédéric Vanderhaegen, University of Valenciennes, France

Gualtiero Volpe, Università degli Studi di Genova, Italy

Marcus Winter, University of Brighton, United Kingdom

Floriano Zini, Free University of Bozen-Bolzano, Italy

SELECTED PAPERS BOOK

A number of selected papers presented at CHIRA 2022 will be published by Springer in a CCIS Series book. This selection will be done by the Conference Co-chairs and Program Co-chairs, among the papers actually presented at the conference, based on a rigorous review by the CHIRA 2022 Program Committee members.

FOREWORD

This volume contains the proceedings of the 6th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2022), which was held in Valletta, Malta as a hybrid event, from 27 to 28 October.

CHIRA is sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC), and is held in cooperation with the ACM Special Interest Group on Management Information Systems and the European Society for Socially Embedded Technologies (EUSSET).

The purpose of CHIRA is to bring together professionals, academics and students who are interested in the advancement of research and practical applications of human-technology & human-computer interaction. Different aspects of Computer-Human Interaction were covered in four parallel tracks: Human Factors and Information Systems, Interactive Devices, Interaction Design, and Adaptive and Intelligent Systems. Human-Computer Interaction is getting renewed interest as human-ai interaction due to the increasing success of artificial intelligence and its applications.

In addition to paper presentations, CHIRA's program included three invited talks delivered by internationally distinguished speakers: Alan Dix (Swansea University, United Kingdom), Truth in an Age of Information; Karen Holtzblatt (InContext Design, United States), Contextual Design: Origins, Ethics, and Diverse Voices; and Abigail Sellen (Microsoft Research Cambridge, United Kingdom), Transforming the Future of Work through Human-Centred Design.

CHIRA received 36 paper submissions from 17 countries, of which 17% were accepted as full papers. The high quality of the papers received imposed difficult choices during the review process. To evaluate each submission, a double-blind paper review was performed by the Program Committee, whose members are highly qualified independent researchers in the CHIRA topic areas.

All presented papers will be submitted for indexation by Scopus, Google Scholar, DBLP Computer Science Bibliography, Semantic Scholar, EI (Elsevier Engineering Village Index), and Web of Science/Conference Proceedings Citation Index (Clarivate). A short list of papers presented at the conference will be selected for publication of extended and revised versions in the CCIS (Communications in Computer and Information Science) Series book by Springer. Another short list of papers will be invited for a post-conference special issue of the Springer Nature Computer Science Journal. All papers presented at this conference will be available at the SCITEPRESS Digital Library.

We express our thanks to all participants. First to all the authors, whose quality work is the essence of this conference; secondly to all members of the Program Committee and auxiliary reviewers, who helped us with their expertise and valuable time. We also deeply thank the invited speakers for excellent contributions in sharing their knowledge and vision.

Finally, we acknowledge the professional support of the CHIRA 2022 team for all organizational processes, especially given the needs of a hybrid event, in order to make it possible for CHIRA 2022 authors to present their work and share ideas with colleagues in spite of the logistic difficulties.

We wish you all an inspiring conference. We hope to meet you again next year for the 7th edition of CHIRA, details of which will soon be available at <http://www.chira.scitevents.org/>.

Hugo Plácido Silva

IT- Instituto de Telecomunicações, Portugal

Jean Vanderdonckt

Université Catholique de Louvain, Belgium

Andreas Holzinger

Medical University Graz, Austria

Larry Constantine
Madeira Interactive Technologies Institute, Portugal

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

Transforming the Future of Work through Human-centred Design 5
Abigail Sellen

Truth in an Age of Information 7
Alan Dix

INVITED LECTURE

Contextual Design: Origins, Ethics, and Diverse Voices 17
Karen Holtzblatt

PAPERS

FULL PAPERS

Insertion of Real Agents Behaviors in CARLA Autonomous Driving Simulator 23
Sergio Martín Serrano, David Fernández Llorca, Iván García Daza and Miguel Ángel Sotelo

Scale Perception in VR for Urban Scale Environments: 360° Photos versus Full 3D Virtual Reality 32
Claus B. Madsen, Nicolai Steinø, Andrei Iuliu Lucaci, Emil O. Sandkvist and Alexander Jonstrup

A-Eye: Driving with the Eyes of AI for Corner Case Generation 41
Kamil Kowol, Stefan Bracke and Hanno Gottschalk

An Integrated Neural Network and Structural Equation Modeling Approach for Modeling Activity Trackers Use 49
Ricardo Sol and Karolina Baras

Between the Buttons: Stress Assessment in Video Games using Players' Behavioural Data 59
Susanna Brambilla, Giuseppe Boccignone, N. Alberto Borghese and Laura A. Ripamonti

Augmented Reality-based Visualization of the Reach of a Collaborative Robot 70
Fabio Schwarz, Maxim Feinleb and Gerrit Meixner

SHORT PAPERS

Usability Evaluation of a Community-led Innovation Mobile App 81
André Castello Branco, Eveline Sacramento, Eliza Oliveira, Oksana Tymoshchuk, Maria Antunes, Margarida Almeida, Luís Pedro, Fernando Ramos and Daniel Carvalho

Relationship between Demographic Factors and Metacognition in Digital Library Interaction 89
Heesop Kim, Aluko Ademola and Yumi Kim

Effect of Environment Size on Spatial Perception in Virtual Reality 97
Morten Trier Bach, Poul Anker Jensen, Andrei Iuliu Lucaci, Damian Grzegorz Pupczyk and Claus Brøndgaard Madsen

An Analysis of Correlations between Empathy and Both EEG and HEG during Text Chat 105
Masaki Omata and Kana Watanabe

UX- for Smart-PSS: Towards a Context-aware Framework <i>Angela Carrera-Rivera, Felix Larrinaga, Ganix Lasa and Giovanna Martinez-Arellano</i>	113
A Music Tangible User Interface for the Cognitive and Motor Rehabilitation of Elderly People <i>Adriano Baratè, Helene Korsten and Luca A. Ludovico</i>	121
Context-aware Personalized Decision Support based on User Digital Life Model <i>Alexander Smirnov and Tatiana Levashova</i>	129
Providing Personalised Recommendations of Critical Incident Narratives in a Cross-platform Mobile Application <i>Tim Wenzel, Doris Fetscher, Wolfgang Golubski, Susanne Klein and Rainer Wasinger</i>	137
A Study of Search User Interface Design based on Hofstede's Six Cultural Dimensions <i>Karen Chessum, Haiming Liu and Ingo Frommholz</i>	145
Subjective Assessment of Commercially Common Input and Display Modalities in a Driving Simulator <i>Kian Motahari and Margarita Vinnikov</i>	155
Towards Evaluating e-Commerce Sites Purchase Intention using Affective Computing, a Preliminary Study <i>Hayri Yigit, Javier De Andrés, Martin González-Rodríguez, Daniel Fernández-Lanvin, Kathia Marçal de Oliveira and Emmanuelle Grislin-Le Strugeon</i>	163
Design Requirements for the Definition of Haptic Messages for Automated Driving Functionalities <i>Joseba Sarabia, Sergio Diaz, Asier Zubizarreta and Joshué Perez</i>	171
RetoñosApp: Work in Progress on a Platform to Support the Teaching of Programming in CS through the Automation and Customization of Learning Processes Guided by Artificial Intelligence <i>David Mauricio Valoyes-Porras, Juan Sebastián Rodríguez-Obregón, David Steven Salamanca-Sánchez and Miguel Alfonso Feijóo-García</i>	179
Popular Music Production Trend Analysis and Prediction Research <i>Mingyue Zheng and Jiarui Jin</i>	187
Fynex: Work in Progress on a Web-based Approach That Implements a Hybrid Recommendation System for Preventing and Treating Diseases based on Eating Disorders <i>Bríter Andrés González-Daza and Miguel Alfonso Feijóo-García</i>	193
User Empowering Design: Expanding the Users' Hierarchy of Needs <i>David Gallula, Hadar Ronen, Ido Shichel and Adi Katz</i>	201
Observing the Uncanny Valley: Gender Differences in Perceptions of Avatar Uncanniness <i>Jacqueline D. Bailey, Karen L. Blackmore and Robert King</i>	209
A Comparison of Date Selection Elements on Mobile Touch Devices in eCommerce Sites <i>Asta Romikaitytė, Stelian Adrian Stanci, Javier De Andrés, Daniel Fernández-Lanvin and Martín González-Rodríguez</i>	217
I Wandered Lonely in the Cloud: A Review of Loneliness, Social Isolation and Digital Footprint Data <i>Dominic Reedman-Flint, John Harvey, James Goulding and Gary Priestnall</i>	225
Low-power Machine Learning for Visitor Engagement in Museums <i>Marcus Winter, Lauren Sweeney, Katie Mason and Phil Blume</i>	236

Student-centered Development of an Online Software Tool to Provide Learning Support Feedback: A Design-study Approach <i>Gilbert Drzyzga and Thorleif Harder</i>	244
EETAS: A Process for Examining Ethical Trade-Offs in Autonomous Systems <i>Catherine Menon, Silvio Carta and Frank Foerster</i>	249
AUTHOR INDEX	257

INVITED SPEAKERS

KEYNOTE SPEAKERS

Transforming the Future of Work through Human-centred Design

Abigail Sellen

Microsoft Research Cambridge, United Kingdom

Abstract: What does the future of work look like in this post-pandemic world? What would we like it to look like? Using a large compendium of research carried out over the past two years on how work is transforming, I will speculate on where we are heading, and discuss how human-centred design can help invent a future more aligned with human aspiration. This includes exploring why we work, and how we measure whether organisations are successful in empowering their workers. I will also look at some of the fundamental changes that have occurred because of the shift to remote work, and discuss the future of work looking beyond office workers in the Global North. I will argue that this shift to a more inclusive, global view of the future of work is where we next need to focus our attention.

BRIEF BIOGRAPHY

Abigail Sellen is Deputy Director at Microsoft Research Cambridge in the UK. She also oversees the lab's portfolio of research exploring the Future of Work, taking an interdisciplinary approach to designing and developing new productivity tools that work in partnership with people. Abi has published on many topics in HCI including: intelligible AI, healthcare, computer input, help systems, reading, paper use in offices, videoconferencing design, search, photo use, gesture-based input, human error and computer support for human memory. This includes "The Myth of the Paperless Office" (with co-author Richard Harper), which won an IEEE award. She is a member of the CHI Academy and a Fellow of the ACM, the Royal Academy of Engineering, and the Royal Society for the Arts. She is an Honorary Professor at Lancaster University, the University of Nottingham, and University College London. In 2020, she was elected as an International Member of the US National Academy of Engineering, and in 2021 she was elected to the Royal Society.

Truth in an Age of Information

Alan Dix^{1,2} 

¹Professorial Fellow, Cardiff Metropolitan University, Wales U.K.

²Director of the Computational Foundry, Swansea University, Wales, U.K.
alan@hcibook.com, <https://alandix.com>

Keywords: Misinformation, Argumentation, Provenance, Qualitative–quantitative Reasoning, Fake News, Echo Chambers.

Abstract: Many of the issues in the modern world are complex and multifaceted: migration, banking, not to mention climate change and Covid. Furthermore, social-media, which at first seemed to offer more reliable 'on the ground' citizen journalism, has instead become a seedbed of dis-information. Trust in media has plummeted, just when it has become essential. This is a problem, but also an opportunity for research in HCI that can make a real difference in the world. The majority of work in this area, from various disciplines including data-science, AI and HCI, is focused on combatting misinformation – fighting back against bad actors. However, we should also think about doing better – helping good actors to curate, disseminate and comprehend information better. There is exciting work in this area, but much still to do.

1 INTRODUCTION

Falsehood flies, and truth comes limping after it.

Jonathan Swift, The Examiner No. 14,
Thursday, 9th November 1710

Politicians have always been 'economical with the truth' and newspapers have toed an editorial line. However, never in recent times does it seem that confidence in our media has been lower. From the Brexit battle bus in the UK to suspected Russian meddling in US elections, fake news to alternative facts – it seems impossible for the general public to make sense of the contradictory arguments and suspect evidence presented both in social media and traditional channels. Even seasoned journalists and editors seem unable to keep up with the pace and complexity of news.

These problems were highlighted during Covid when understanding of complex epidemiological data was essential for effective government policy and individual responses. As well as the difficulty of media (and often government) in understanding and communicating the complexity of the situation, various forms of misinformation caused confusion. There are obvious health impacts of this misinformation due to taking dangerous 'cures' (Nelson, 2020) and vaccination hesitancy (Lee, 2022a), as well as its role in encouraging violence against health workers (Mahase, 2022). In addition, a

meta-review of many studies of Covid misinformation identified mental health impacts as also significant (Rocha, 2021).

If democracy is to survive and nations coordinate to address global crises, we desperately need tools and methods to help ordinary people make sense of the extraordinary events around them: to sift fact from surmise, lies from mistakes, and reason from rhetoric. Similarly, journalists need the means to help them keep track of the surfeit of data and information so that the stories they tell us are rooted in solid evidence.

Crucially in increasingly politically fragmented societies, we need to help citizens explore their conflicts and disagreements, not so that they will necessarily agree, but so that they can more clearly understand their differences.

These are not easy problems and do not admit trite solutions. However, there is existing work that offers hope: tracking the provenance of press images (ICP, 2016), ways to expose the arguments in political debate (Carneiro, 2019), even using betting odds to track the influence of news on electoral opinion (Wall, 2017).

I hope that this paper will give hope that we can make a difference and offer challenges for future research.

2 THE B-MOVIE CAST OF MISINFORMATION

Deliberate misinformation is perhaps the most obvious problem we face. There is extensive data science studies by academics and data journalists attempting to understand the extent and modes of spread (e.g. Albright, 2016; Vosoughi, 2018). Crucially false information appears to spread more rapidly than true information; possibly because it is more novel (Vosoughi, 2018). Although there is considerable debate as to the sufficiency of their responses, both Facebook and Twitter are constantly adjusting algorithms and policies to attempt to prevent or discourage fake news (Dreyfuss, 2019; NPR, 2022; Twitter, 2022). Within the HCI community there has been considerable work exploring the human aspects around the spread of misinformation online (Flintham, 2018, Geeng, 2020; Varanasi, 2022), ways to visualise it (Lee, 2022b), tools for end-users to help identify it (Heuer, 2022) and CHI workshops (Gamage, 2022; Piccolo, 2021).

2.1 Bad Actors

Much of the focus on misinformation is on ‘bad actors’: extremist organisations, ‘foreign’ powers interfering in elections, or simply those aiming to make a fast buck. In the context of mis-information, ‘bad’ can mean two things:

1. They are intrinsically bad people, bad states, or bad media.
2. They use bad methods and/or spread bad information (including misinformation and hateful or violent content).

The first of these can be relative to clear criteria such as human rights or terrorism, but may simply mean those we disagree with; and, of course, the boundary between the two may often be unclear.

When the two forms of ‘bad’ agree the moral imperative is clear, even though implementation may be harder. Forced in part by government and popular pressure, social media platforms have extensive mechanisms both to attempt to suppress bad information and suspend accounts of those who promulgate it (Guardian, 2018).

Probably the most high-profile example of the later was Twitter’s suspension of @realDonaldTrump. This was both met with widespread relief, but also caution due to its potential impact on free speech (Noor, 2021), especially given Twitter’s arguments for why it was suspended when it was (Twitter, 2021).

Of course, sometimes bad actors may spread true (or even good) information.

In some cases this is simply because few are altogether bad. For example, those who believe and then promulgate Covid conspiracy theories; many will be well meaning, albeit deeply misguided, and some of the information may be accurate.

However, true information can also be cynically used to give credence to otherwise weak or misleading arguments; for example a recent study of cross-platform misinformation (Micallef, 2022) found a substantial proportion of cases where a YouTube video with true information about Covid was referenced by a tweet or post that in some way mis-interpreted the material or used it out of context. In addition, many Astroturfing accounts will distribute accurate information as a means to create trust before disseminating misinformation. It can be hard to distinguish these and it is not uncommon for politicians or other campaign groups to inadvertently re-tweet or quote true or at least defensible information that originated from very unsavoury groups, thus giving them credence.

2.2 When Good Actors Spread Bad Information

As we saw in the last example, those we regard as ‘good’ actors can also sometimes spread bad information.

Sometimes this is deliberate. An extreme case is during war when misinformation campaigns in an enemy country are regarded as a normal and indeed relatively benign form of warfare (Shaer, 2017). In peace time deliberate misinformation is likely to be less extreme and more often stretching or embroidering the truth, or selectively reporting.

It may also be accidental. For example, Figure 1 shows a “Q&A” (form of fact check) on the BBC news web site following a claim made by Boris Johnson in January 2018 regarding UK contributions to the EU budget. The overall thrust of the Q&A is correct, the net amount that was sent to the EU at that time was substantially less than the £350 million figure that Johnson claimed, but the actual figures are wrong, the Q&A suggested that around 2/3 of the gross figure was returned, when the actual figure was close to a half. This is probably because at some point a journalist lost track of which figure the half was referring to, but the overall effect was to create a substantially incorrect figure.



Figure 1: Q&A on BBC new web site January 2018 (BBC, 2018). Note this Q&A pop-up is no longer in the news item; instead there a link to a 'Reality Check' page which is correct, but with no explicit retraction.

In between are the subtle biases are simply assumptions of journalists that play out in the selection of which stories to report and also in the language used. For example, in crime or conflict reporting passive language may be used ("the assailant was shot", or "shells fell on") compared with active language ("AAA shot BBB" or "XXX fired shells on") depending on which side is doing the shooting or bombing.

Personally, while I may despair or be angry at the misinformation from those with whom I disagree, I am most upset when I see poor arguments from those with whom I agree. This is partly pride, wanting to be able to maintain a moral high ground, and partly pragmatic, if the arguments are poor then they can be refuted.

In an age of adversarial media, any mistakes, misrepresentation or hyperbole can be used to discredit otherwise well-meaning sources and promote alternatives that are either ill-informed or malicious. This was evident in the US during the 2016 presidential campaign when many moderate Republican supporters lost faith in the reputable national press in favour of highly partisan local papers; a trend which has intensified since (Gottfried, 2021; Meek, 2021)

3 SEEKING TRUTH

3.1 The Full Cast

We have already considered the 'B-movie' bad/good guy roles, of the producers and influencers, both of whom can mislead whether ill-intentioned or ill-informed. In reality even the 'bad' actors may be those with genuinely held, albeit unfounded, beliefs about 5G masts or communist take-over of US government. Of course, those of us who would consider ourselves 'good' actors, may still distort or

be selective in what we say albeit for the best of reasons.

In addition, those who receive misinformation and are confused or misled by it may differ in levels of culpability. It is easier to believe the things that make life easier, whether it is the student grasping at suggestions that the impact of Covid may be over exaggerated in order to justify a party, or the professional accepting climate change scepticism to justify buying that new fuel-hungry car.

Of course, the purveyors of news and information are under pressure, and may not be wholly free in what they say, or may run risks if they do. Even in the last year we have seen many journalists, bloggers and authors arrested, sanctioned, stabbed and shot.

Perhaps more subtle is the interplay within the ecology of information: journalists and social media modify what and how they present information in order to match the perceived opinions and abilities of their readership.

3.2 Two Paths

The greatest effort currently appears to be focused on fighting back against bad actors. This includes algorithms to detect and counter mis-information, such as Facebook's intentions to weed out anti-vaccination. These are predominantly aimed at the bad actors.

However, in addition we need to think about doing better, ways for the good actors to disseminate and understand information so that they are in a better position to evaluate sources of information and ensure that they do not inadvertently create bad information.

We'll look briefly at four areas where appropriate design could help us to do better:

- echo chambers and filter bubbles
- better argumentation
- data and provenance
- numeric data and qualitative-quantitative reasoning

These are not the only approaches, but I hope they will stimulate the reader to think of more.

3.3 Echo Chambers and Breaking Filter Bubbles

Social media was initially seen as a way to democratise news and information sharing and to allow those in the 'long-tail' of small interest groups to find like-minded people in the global internet. However, we now all realise that an outcome of this has been the creation of echo chambers, where we increasingly only hear views that agree with our own.

In some ways this has always been the case, both in choices of friendship groups for informal communication and the audiences of different newspapers. However, social media and the personalisation of digital media has both intensified the effect and made it less obvious – you know that a newspaper has a particular editorial line, but do not necessarily recognize that web search results have been tuned to your existing prejudice.

This is now a well-studied area with extensive work analysing social media to detect filter bubbles and understand the patterns of communication and networks that give rise to them (Terren, 2021, Garimella, 2018; Cinelli, 2021). Notably, one of these studies (Garimella, 2018) highlighted the role of ‘gatekeeper’, people who consume a broad range of content, but then select from this to create partisan streams. Perhaps more sadly, the same study notes that those who try to break down partisan barriers pay a “price of bipartisanship” in that balanced approaches or multiple viewpoints are not generally appreciated by their audiences.

In addition, there has been work on designing systems that in different ways attempt to help people see beyond their own filter bubbles (e.g. Foth, 2016; Jeon, 2021), but on the whole this has been less successful, especially in actual deployment. Indeed, attempts to present opposite arguments can end up deepening divides if they are too different and too soon.

3.4 Argumentation

It is easy to see the flaws in arguments with which we disagree, we know it is wrong and can thus hunt for the faults – the places where our intuitions and the argument disagree are precisely the places where we are expecting holes in the reasoning. Of course, we all create bad arguments. It is very hard to notice the gaps in one’s own reasoning, but also the fallacious arguments of others when one agrees with their final conclusions.

Of course, those who disagree with us will notice the gaps in our arguments, thus increasing their own confidence and leading them to discount our opinions!

It is crucial therefore to have tools that both help the public to interrogate the arguments of politicians and influencers, and also to help those who are aiming to create solid evidence-based work (including academics) to ensure valid arguments.

There is of course long-standing work on argumentation systems, such as IBIS (Noble, 1988) and work in the NLP community to automatically

analyse arguments. Much of this is targeted towards more professional audiences, but there are also steps to help the general public engage with media, such as the Deb8 system (Carneiro, 2019) developed at St Andrews, an accessible argumentation system that allows viewers of a speech or debate to collaboratively link assertions in the video to evidence from the web.

This is an area which seems to have many opportunities for research and practical systems aimed at different audiences including the general public, journalists, politicians, academics, and fact checkers. This could include broad advice, for example, ensuring that fact checkers clearly state their interpretation of a statement before checking it to avoid inadvertently debunking a strawman misinterpretation. Similarly, we could imagine templates for arguments, for example, given an implication of the form “if A then B”, it is important to keep track of the assumptions. In particular, while more formal logics and some forms of argumentation schemes focus on low-level argumentation, it seems that the tools needed perhaps need to focus on the higher-level argumentation, the information and assumptions that underly a statement, more than the precise logic of the inference.

In addition, in the AI community there are now a variety of tools to help automatically detect possible bias in data or machine learning algorithms. Maybe some of these could be borrowed to help human reasoning, for example shuffling aspects of situations (e.g. gender, political party or ethnicity), to help us assess to what extent our view is shaped by these factors.

3.5 Data and Provenance

One of the forms of misinformation is the deliberate or accidental use of true information or accurate data divorced from its context. For the spoken word or text, this might be a quotation, for photographs or video the choice of a still, segment or even parts edited together that give a misleading impression. Indeed the potential for digital media to be compromised in different ways lead some to look for technology such as blockchains to prevent tampering, or the use of analogue or physical representations (Haliburton, 2021).

One example of work addressing this issue was the FourCorners project (ICP, 2016), a collaboration between OpenLab Newcastle, the International Centre for Photography and the World Press Photo Foundation, which embeds provenance into photographs allowing interrogation such as “what are

the frames before and after this photograph?", "are there other photos at the same time and place?". One can imagine similar things for textual quotes, in the manner of Ted Nelson's vision of *transclusion* (Nelson, 1981), where segments quoted from one document retain their connection back to the original.

This is an area I've worked on personally in the past with the Snip!t system, originally developed in 2003 following a study of user bookmarking practice (Dix, 2003). Snip!t allowed users to 'bookmark' portions of a web page and automatically kept track not just of the quoted text, but where it came from (Dix, 2010). Later work in this area by others has included both commercial systems such as Evernote, and academic research, such as Information Scraps (Bernstein, 2008). Currently there is an explosion of personal knowledge management (PKM) apps, some of which, such as Readwise (readwise.io) and Instapaper (instapaper.com), help with the process of annotating documents. However, these systems are mostly focused on retaining the context of captured notes and quotes; we desperately need better ways to retain this once the quote is embedded in another document or web page.

This connection to sources is also important for data. In the example from the BBC in Figure 1, the journalist had clearly lost track of the original data on UK/EU funding and so misremembered aspects.

Can we imagine tools for journalists that would help them keep track of the sources for data and images. Indeed, it would be transformative if everyday office tools such as word processors and presentation software made it easy to keep references to imported images. In work with humanities and heritage, we have noted how file systems have barely altered since the 1970s (Dix, 2022) – the folder structures allow us to store and roughly classify, but there is virtually no support for talking about documents and about their relationships to one another. Semantic desktop research (Sauermann, 2005), which seemed promising at the time, has never found its way into actual operating systems.

Happily there are projects, such as Data Stories (2022) that are helping communities to use data to tell their own stories, so that the online world can allow open discourse and interpretation, whilst connecting to the underlying data on which it is based. Furthermore, one of the popular PKM apps Obsidian (obsidian.md) supports semi-structured meta-data for every note.

3.6 Numeric Data and Qualitative–quantitative Reasoning

Going back to the example in figure 1, part of the problem here may well simply be that journalists are often more adept with words than numbers. We are in a world where data and numerical arguments are critical. This was true of Covid where the understanding of exponential growth and probabilistic behaviour was crucial, but equally so for issues such as climate change.

One of the arguments put forward by climate change sceptics, is that it is hard to believe in long-term climate models given forecasters sometimes struggle to predict whether it is going to rain next week. This, at first sight, is not an unreasonable argument; although anyone who has dealt with stochastic phenomena knows that it is often easier to predict long-term trends than short-term behaviour. Indeed, it is also relatively easy to communicate this – we can all say with a degree of reliability that a British winter will be wetter and colder than the summer, even though we'll struggle to know the weather from day to day.

This form of argument is not about exact numerical calculation, nor about abstract mathematics, but something else – informal reasoning about numerical phenomena. Elsewhere I've called this qualitative–quantitative reasoning (Dix, 2021a, 2021b) and seems to be a critical, but largely missing, aspect for universal education. Again this is an area that is open for radical contributions, for example, iVolver (Nacenta, 2017) allows users to extract numerical and other data from visualisations, such as pie charts, in published media. My own work has included producing table recognisers in commercial intelligent internet system OnCue in the dot-com years (Dix, 2000) and more recently investigating ways to leverage some of the accessibility of spreadsheet-like interfaces and simple ways to allow users to combine their own data (Dix, 2016).

4 CALL TO ACTION

We are at a crucial time in a world where information is everywhere and yet we can struggle to see the truth amongst the poorly sourced, weakly argued, deliberately manipulated or simply irrelevant. However, there are clear signs of hope in work that is being done and also opportunities for research that can make a real difference.

Of course, as academics we are also in the midst of a flood of scholarly publication, some more

scholarly than others! There are calls for us to ‘clean up our own act’ too including rigour of academic argumentation (Basbøll, 2018) and transparency of data and materials (Wacharamanotham, 2020). As well as being a problem we need to deal with within academia, it is also an opportunity to use our own academic community as a testbed for tools and techniques that could be used more widely.

REFERENCES

- Albright, J. (2016). *The #Election2016 Micro-Propaganda Machine*. Medium. Nov 18, 2016. <https://d1gi.medium.com/the-election2016-micro-propaganda-machine-383449cc1fba>
- Basbøll, T. (2018). A scientific paper shouldn’t tell a good story but present a strong argument. *LSE Impact Blog*. June 1st, 2018. <https://blogs.lse.ac.uk/impactofsocialsciences/2018/06/01/a-scientific-paper-shouldnt-tell-a-good-story-but-present-a-strong-argument/>
- Bernstein, M., van Kleek, M., Karger, D., schraefel, mc.. 2008. Information scraps: How and why information eludes our personal information management tools. *ACM Trans. Inf. Syst.* 26(4), Article 24 (September 2008), 46 pages. DOI:10.1145/1402256.1402263
- BBC (2018). £350m Brexit claim was ‘too low’, says Boris Johnson. *BBC News*, 16 January 2018. <https://www.bbc.co.uk/news/uk-42698981>
- Carneiro, G., Nacenta, M., Toniolo, A., Mendez, G., Quigley, A. (2019). Deb8: A Tool for Collaborative Analysis of Video. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19)*. ACM, NY, USA, 47–58. DOI: 10.1145/3317697.3323358
- Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), e2023301118.
- Data Stories (2022). *Data Stories: engaging with data in a post-truth environment*. Accessed 29/9/2022. <http://datastories.co.uk/project/>
- Dix, A., Marshall, J. (2003). At the right time: when to sort web history and bookmarks. In *Volume 1 of Proceedings of HCI International 2003*. J. Jacko and C. Stephandis (ed.). Lawrence Erlbaum Associates, 2003. pp. 758-762
- Dix, A., Beale, R., Wood, A. (2000). Architectures to make Simple Visualisations using Simple Systems. *Proceedings of Advanced Visual Interfaces - AVI2000*, ACM Press, pp. 51-60. <https://www.alandix.com/academic/papers/avi2000/>
- Dix, A., Lepouras, G., Katifori, A., Vassilakis, C., Catarci, T., Poggi, A., Ioannidis, Y., Mora, M., Daradimos, I., Akim, N.M., Humayoun, S.R. (2010). From the web of data to a world of action. *Journal of Web Semantics*, 8(4), pp.394-408.
- Dix, A. (2016). The Leaves are Golden - putting the periphery at the centre of information design. Keynote at *HCI2016*, July 2016, Bournemouth, UK. <https://www.alandix.com/academic/talks/HCI2016-the-leaves-are-golden/>
- Dix, A. (2021a). Qualitative–Quantitative Reasoning - Understanding and managing the behaviour of numeric phenomena. In *Emergent Interaction Complexity, Dynamics, and Enaction in HCI, A CHI 2021 Workshop*, 15th May, 2021. <https://www.alandix.com/academic/papers/QQ-Emergent2021/>
- Dix, A. (2021b). Qualitative–Quantitative Reasoning: thinking informally about formal things, Keynote at *ICTAC 2021: 18th International Colloquium on Theoretical Aspects of Computing*, Nazarbayev University, Nur-Sultan, Kazakhstan, September 6-10, 2021. <https://alandix.com/academic/papers/ICTCS-QQ-2021/>
- Dix, A. Jones, E., Neads, C., Davies, V., Cowgill, R., Armstrong, C., Ridgwell, R., Downie, J.S., Twidale, M., Reagan, M., Bashford, C. (2022). Tools and technology to support rich community heritage. In *Proceedings of British HCI Conference (BHCI2022)*, Keele, UK. 11-13 July 2022.
- Dreyfuss, E., Lapowsky, I. (2019). Facebook Is Changing News Feed (Again) to Stop Fake News. *Wired*, Apr 10, 2019. <https://www.wired.com/story/facebook-click-gap-news-feed-changes/>
- Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., Moran, S. (2018). Falling for Fake News: Investigating the Consumption of News via Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, NY, USA, Paper 376, 1–10. DOI:10.1145/3173574.3173950
- Foth, M., Tomitsch, M., Forlano, L., Haeusler, M.H., Christine Satchell, C. (2016). Citizens breaking out of filter bubbles: urban screens as civic media. In *Proceedings of the 5th ACM International Symposium on Pervasive Displays (PerDis '16)*. ACM, NY, USA, 140–147. DOI:10.1145/2914920.2915010
- Gamage, D., Stomber, J., Jahanbakhsh, F., Skeet, B., Kishore Shahi, G. (2022). Designing Credibility Tools To Combat Mis/Disinformation: A Human-Centered Approach. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, NY, USA, Article 107, 1–4. DOI:10.1145/3491101.3503700
- Garimella, K., de Francisci Morales, G., Gionis, A., Mathioudakis, M. (2018). Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 913–922. DOI:10.1145/3178876.3186139
- Geeng, C., Yee, S., Roesner, F. (2020). Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*

- (CHI '20). ACM, NY, USA, 1–14. DOI:10.1145/3313831.3376784
- Gottfried, J. (2021) *Republicans less likely to trust their main news source if they see it as 'mainstream'; Democrats more likely*. Pew Research Centre, July 1, 2021. <https://www.pewresearch.org/fact-tank/2021/07/01/republicans-less-likely-to-trust-their-main-news-source-if-they-see-it-as-mainstream-democrats-more-likely/>
- Guardian (2018). Twitter bans 270,000 accounts for 'promoting terrorism'. *The Guardian*, 5 Apr 2018. <https://www.theguardian.com/technology/2018/apr/05/twitter-bans-270000-accounts-to-counter-terrorism-advocacy>
- Haliburton, L., Hoppe, M., Schmidt, A., Kosch, T. (2021). Quick, Print This Page! The Value of Analogue Media in a Digital World. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. ACM, NY, USA, Article 33, 1–7. DOI:10.1145/3411763.3450375
- Heuer, H., Glassman, E. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, NY, USA, Article 241, 1–21. DOI:10.1145/3491102.3517717
-] ICP (2016). *ICP, World Press Photo Foundation, and Newcastle University's Open Lab Launch Image Authoring Initiative At World Press Photo Awards Days*. Media Release, International Centre for Photography, New York, NY. April 20, 2016. https://www.icp.org/files/ICP_FourCorners_press-release.pdf
- Jeon, Y., Kim, B., Xiong, A., Lee, D., Kyungsik Han, K. (2021). ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System. *Proc. ACM Human-Computer Interaction*, Vol 5 (CSCW2), Article 472 (October 2021), 26 pages. DOI:10.1145/3479859
- Lee, S.K., Sun, J., Jang, S. et al. (2022a). Misinformation of COVID-19 vaccines and vaccine hesitancy. *Nature Science Reports* 12, 13681 (2022). DOI:10.1038/s41598-022-17430-6
- Lee, S., Afroz, S., Park, H., Wang, Z., Shaikh, O., Sehgal, V., Peshin, A., Chau, D. (2022b). MisVis: Explaining Web Misinformation Connections via Visual Summary. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, NY, USA, Article 228, 1–6. DOI:10.1145/3491101.3519711
- Mahase E. (2022). Nearly 1500 health workers were attacked or arrested in 2021, report finds. *BMJ* 2022; 377 :o1315 doi:10.1136/bmj.o1315
- Meek, A. (2021). Republicans Are Abandoning The National Mainstream Media In Doves. *Forbes*, Sept 6, 2021. <https://www.forbes.com/sites/andymeek/2021/09/06/republicans-are-abandoning-the-national-mainstream-media-in-doves-because-they-dont-trust-it/?sh=6586c5e35889>
- Micallef, N., Sandoval-Castanñeda, M., Cohen, A., Ahamad, M., Kumar, S., Memon, N. (2022). Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022.
- Nacenta, M., and Méndez, G. (2017). iVolver: A visual language for constructing visualizations from in-the-wild data. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*. (pp. 438-441).
- Nelson, T. H. (1981). *Literary Machines: The Report On, And Of, Project Xanadu, Concerning Word Processing, Electronic Publishing, Hypertext, Thinkertoy, Tomorrow's Intellectual Revolution, And Certain Other Topics Including Knowledge, Education And Freedom*. Mindful Press, Sausalito, CA, USA.
- Nelson T, Kagan N, Critchlow C, Hillard A, Hsu A. (2020). The Danger of Misinformation in the COVID-19 Crisis. *Missouri Medicine*. 117(6):510-512. PMID: PMC7721433.
- Noble, D., Rittel, H. W. (1988). Issue-based information systems for design. *Computing in Design Education*, In *ACADIA Conference Proceedings*, Ann Arbor (Michigan / USA) 28-30 October 1988, pp. 275–286. DOI:10.52842/conf.acadia.1988.275
- Noor, P. (2021). Should we celebrate Trump's Twitter ban? Five free speech experts weigh in. *The Guardian*. 17th January 2021. <https://www.theguardian.com/us-news/2021/jan/17/trump-twitter-ban-five-free-speech-experts-weigh-in>
- NPR (2022), Twitter aims to crack down on misinformation, including misleading posts about Ukraine. *NPR Technology*. May 19, 2022. <https://www.npr.org/2022/05/19/1100100329/twitter-misinformation-policy-ukraine>
- Piccolo, L., Bertel, D., Farrell, T., Troullinou, P. (2021). Opinions, Intentions, Freedom of Expression, ... , and Other Human Aspects of Misinformation Online. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. ACM, NY, USA, Article 84, 1–5. DOI:10.1145/3411763.3441345
- Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., de Figueiredo Nicolette, L. D. (2021). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health*, 22(5):1-10. DOI: 10.1007/s10389-021-01658-z
- Sauermann, L., Bernardi, A., Dengel, A. (2005). Overview and Outlook on the Semantic Desktop. In *Proceedings of the 2005 International Conference on Semantic Desktop Workshop: Next Generation Information Management and Collaboration Infrastructure*, CEUR Workshop Proceedings, Vol. 175, pp. 74–91
- Shaer, M. (2017) Fighting the Nazis With Fake News. *Smithsonian Magazine*, April 2017. <https://www.smithsonianmag.com/history/fighting-nazis-fake-news-180962481/>
- Terren, L., Borge-Bravo, R. (2021). Echo Chambers on Social Media: A Systematic Review of the Literature.

- Review of Communication Research (Communication and Media Technologies)*, 9: 99–118. <https://rcommunicationr.org/index.php/rcr/article/view/94>
- Twitter Inc. (2021). *Permanent suspension of @realDonaldTrump*. Twitter Blog, 8th January 2021. https://blog.twitter.com/en_us/topics/company/2020/suspension.html
- Twitter Inc. (2022). *How we address misinformation on Twitter*. Twitter Help Centre (accessed 19/9/2022). <https://help.twitter.com/en/resources/addressing-misleading-info>
- Varanasi, R., Pal, J., Vashistha, A., (2022). Accost, Accede, or Amplify: Attitudes towards COVID-19 Misinformation on WhatsApp in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, NY, USA, Article 256, 1–17. DOI:10.1145/3491102.3517588
- Vosoughi, S., Roy D., Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151. DOI:10.1126/science.aap9559
- Wacharamanotham, C., Eisenring, L., Haroz, S., Echtler, F. (2020). Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, NY, USA, 1–14. DOI:10.1145/3313831.3376448
- Wall, M., Costello, R., Lindsay, S. (2017). The miracle of the markets: Identifying key campaign events in the Scottish independence referendum using betting odds. *Electoral Studies*, 46, 39–47.

INVITED LECTURE

Contextual Design: Origins, Ethics, and Diverse Voices

Karen Holtzblatt

InContext Design, United States

Abstract: Contextual Design (CD) is a user-centered requirements and design process used by teams and taught in universities worldwide. Introduced by Karen Holtzblatt in 1991, Contextual Inquiry—the field interviewing technique—and the other CD techniques help teams use customer data to drive product concept creation and design. These processes were honed by working with teams developing products, business applications, consumer goods, services, websites, apps, and more. In this talk, Karen returns to the origins of these CD processes, revealing the ethical considerations and concern for ensuring that all voices are heard on diverse teams that informed the process. Contextual Inquiry (CI) was a direct reaction to how experimental psychology studies treated “subjects,” especially children. Re-framing interactions with participants to ensure a more equal power relationship is at the core of this practice. Equal power depends on active listening and accountability on the part of the researcher. This accountability recognizes that researchers bring their unarticulated assumptions into the interaction and the interpretation of participants’ behavior and experience. It calls upon researchers to share and check their interpretations with the person, promoting a co-interpretation relationship. Karen explains how the principles of CI teach how to gather data from users—or anyone being interviewed—to ensure that they are respected, heard, and feel understood while getting the necessary information needed for the project. As Karen worked with product teams, ensuring that the people on the team worked well together became a concern. The voices of the customers cannot influence the product direction if the team does not know how to use the data. But innovation is not served if the practices of using the data do not allow all voices to be heard and influence the design direction. Extensive research now supports the link between developing innovative products and having a diverse team. As teams became more diverse in job roles, gender, and ethnicity, the need to create team practices to equalize power and participation was essential. To ensure that the voice of the customer actually influences the product, CD needed to build in practices so diverse teams worked well together. Informed by social psychology, quality practices, and good team techniques, CD techniques were iterated in the daily work with diverse teams. Karen walks through some of the CD practices revealing how they were designed to support ethical and responsible interactions with users and between team members so that all diverse voices are heard. Karen also shares how these techniques can be used for a wide range of inquiries and process design as she has done for her latest work on retaining women in tech.

BRIEF BIOGRAPHY

Karen Holtzblatt is CEO of InContext Design, a thought leader, industry speaker, and author. A recognized leader in requirements and design, Karen has developed transformative design approaches throughout her career. She introduced Contextual Design, the industry standard for understanding the customer and organizing that data to drive innovative product and service concepts. Her book *Contextual Design 2nd Edition Design for Life* is used by companies and universities worldwide. In recognition of her impact on the field, Karen was awarded the first Lifetime Award for Practice by ACM SIGCHI (The Association of Computing Machinery’s Special Interest Group on Computer-Human Interaction). Karen is also the driving force behind the Women in Tech Retention Project

housed at witops.org. WITops research explores why women in technology professions leave the field and creates tested interventions to help women thrive and succeed. Her book with Nicola Marsden, *Retaining Women in Tech: Shifting the Paradigm* describes this work.

PAPERS

FULL PAPERS

Insertion of Real Agents Behaviors in CARLA Autonomous Driving Simulator

Sergio Martín Serrano¹^a, David Fernández Llorca^{1,2}^b, Iván García Daza¹^c
and Miguel Ángel Sotelo¹^d

¹Computer Engineering Department, University of Alcalá, Alcalá de Henares, Spain

²European Commission, Joint Research Centre (JRC), Seville, Spain

{sergio.martin, david.fernandezl, ivan.garciad, miguel.sotelo}@uah.es

Keywords: Automated Driving, Autonomous Vehicles, Predictive Perception, Behavioural Modelling, Simulators, Virtual Reality, Presence.

Abstract: The role of simulation in autonomous driving is becoming increasingly important due to the need for rapid prototyping and extensive testing. The use of physics-based simulation involves multiple benefits and advantages at a reasonable cost while eliminating risks to prototypes, drivers and vulnerable road users. However, there are two main limitations. First, the well-known *reality gap* which refers to the discrepancy between reality and simulation that prevents simulated autonomous driving experience from enabling effective real-world performance. Second, the lack of empirical knowledge about the *behavior of real agents*, including backup drivers or passengers and other road users such as vehicles, pedestrians or cyclists. Agent simulation is usually pre-programmed deterministically, randomized probabilistically or generated based on real data, but it does not represent behaviors from real agents interacting with the specific simulated scenario. In this paper we present a preliminary framework to enable real-time interaction between real agents and the simulated environment (including autonomous vehicles) and generate synthetic sequences from simulated sensor data from multiple views that can be used for training predictive systems that rely on behavioral models. Our approach integrates immersive virtual reality and human motion capture systems with the CARLA simulator for autonomous driving. We describe the proposed hardware and software architecture, and discuss about the so-called *behavioural gap* or *presence*. We present preliminary, but promising, results that support the potential of this methodology and discuss about future steps.

1 INTRODUCTION

The recent explosion in the use of simulators for rapid prototyping and testing of autonomous systems has occurred primarily in the field of autonomous vehicles. This is mostly due to the compelling need for extensive validation, as real test-driving alone cannot provide sufficient evidence for demonstrating their safety (Kalra and Paddock, 2016). The relative ease of generating massive data for training and testing, especially edge cases, having full control over all the variables under study (e.g., street layout, lighting conditions, traffic scenarios), makes its use very attractive. In addition, the generated data is annotated by

design including semantic information.

But the use of virtual-world data also brings with it some challenges. The realism of the simulated sensor data (i.e., radar, LiDAR, cameras) and the physical models are critical aspects to minimize the reality gap. Advances in this area are very remarkable, and highly realistic synthetic data has been shown to be useful for improving different perception tasks such as pedestrian detection (Vázquez et al., 2014), semantic segmentation (Ros et al., 2016) or vehicle speed detection (Martinez et al., 2021). However, despite efforts to generate realistic synthetic behaviors of road agents (e.g., to address trajectory forecasting (Weng et al., 2021)), simulation lacks empirical knowledge about the behavior of other road users (e.g., vehicles, pedestrians, cyclists), including behavior and motion prediction, communication, and human-vehicle interaction (Eady, 2019).

In this paper, we present a new approach to in-

^a <https://orcid.org/0000-0002-8029-1973>

^b <https://orcid.org/0000-0003-2433-7110>

^c <https://orcid.org/0000-0001-8940-6434>

^d <https://orcid.org/0000-0001-8809-2103>

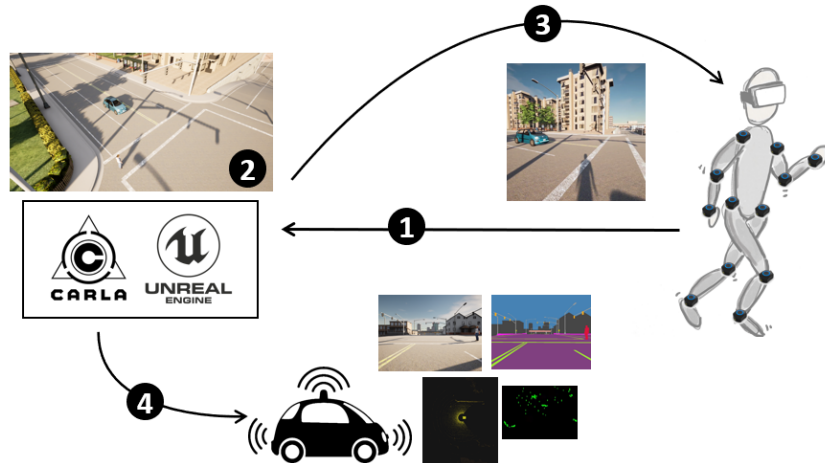


Figure 1: Overview of the presented approach. (1) CARLA-Unreal Engine is provided with the head (VR headset) and body (motion capture system) pose. (2) The scenario is generated, including the autonomous vehicles and the digitized pedestrian. (3) The environment is provided to the pedestrian (through VR headset). (4) Autonomous vehicle sensors perceive the environment, including the pedestrian.

corporate real agents behaviors and interactions in CARLA autonomous driving simulator (Dosovitskiy et al., 2017) by using immersive virtual reality and human motion capture systems. The idea is to integrate a subject in the simulated scenarios using CARLA and Unreal Engine 4 (UE4), with real time feedback of the pose of his head and body, and including positional sound, trying to generate an experience realistic enough for the subject to feel physically present in the virtual world and accept it at a subconscious level (i.e., maximize virtual reality presence). At the same time, the captured pose and motion of the subject is integrated into the virtual scenario by means of an avatar, so that the simulated sensors of the autonomous vehicles (i.e., radar, LiDAR, cameras) can be aware of their presence. This is very interesting because it allows, on the one hand, to obtain synthetic sequences from multiple points of view based on the behavior of real subjects, which can be used to train and test predictive perception models. And on the other hand, they also allow to address different types of interaction studies between autonomous vehicles (i.e., software in the loop) and real subjects, including external human-machine interfaces (eHMI), in fully controlled conditions and with total safety. To our knowledge, this is the first attempt to integrate the behaviors of real road agents into an autonomous vehicle simulator for these purposes.

2 RELATED WORK

In the following, we summarize the most relevant works that have proposed the use of virtual reality and

simulators for vulnerable road users (mainly pedestrians and cyclists) in the context of automated driving. Since the visual realism of simulators has improved greatly in recent years, we focus on relatively recent work, as the conclusions obtained with less realistic simulators are not entirely applicable to the current context.

Most of the previous works proposed the use of virtual reality and simulators to study pedestrian behavior in scenarios where they have to interact with an autonomous vehicle (high levels of automation (Fernández-Llorca and Gómez, 2021)) including the perception of distances and the intention to cross (Doric et al., 2016), (Deb et al., 2017), (Iryo-Asano et al., 2018), (Bhagavathula et al., 2018), (Farooq et al., 2018), (Nuñez Velasco et al., 2019).

Additionally, these approaches have recently been used to study the communication between autonomous vehicles and pedestrians, including acceptance and behavioral analysis of various eHMI systems (Chang et al., 2017), (Holländer et al., 2019), (Deb et al., 2020), (Gruenefeld et al., 2019), (Löcken et al., 2019), (Nguyen et al., 2019). Virtual environments allow experimentation with multiple types of eHMI systems, and with multiple types of vehicles in a relatively straightforward fashion.

We can find similar studies for the case of cyclists, based on simulation platforms and immersive virtual reality (Ullmann et al., 2020), usually to analyze the perceptions and behavior of cyclists to various environmental variables (Xu et al., 2017), (Nuñez Velasco et al., 2021), (Nazemi et al., 2021).

The aforementioned studies have not considered the use of motion capture systems to integrate visual

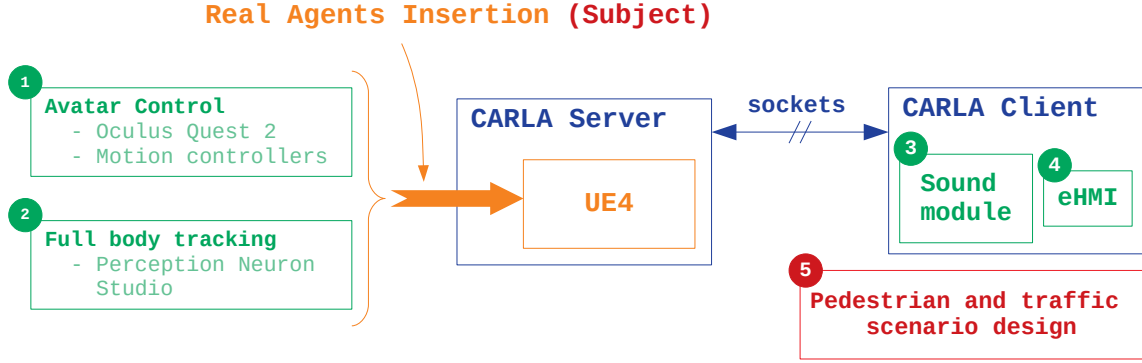


Figure 2: System Block Diagram.

body feedback into the simulation. This is an important factor for two main reasons. First, a fully-articulated visual representation of the users in the immersive virtual environment can enhance their subjective sense of feeling present in the virtual world (Bruder et al., 2009). Second, and more important for our purpose, the articulated representation of the pedestrian pose (body, head, and even hands) within the simulator allows, on the one hand, to integrate the user within the simulated scenario and to be detected by the sensors of the autonomous vehicles, allowing the generation of data sequences that can be very useful for training and testing predictive models. On the other hand, it allows real-time interaction between the autonomous vehicle and the pedestrian (i.e., software in the loop), which can be very useful to test edge cases.

In some studies, user-autonomous vehicle interaction is not possible, as they rely on the projection of prerecorded 360° videos (i.e., 360° video-based virtual reality (Nuñez Velasco et al., 2019), (Nuñez Velasco et al., 2021)). While this approach certainly enhances visual realism, it does not enable interaction. But even with the use of simulators, we have found only one paper (Gruenefeld et al., 2019) that attempts some kind of interaction between the pedestrian and the autonomous vehicle via hand gestures (i.e. stop gesture). However, the gesture detection process of the autonomous vehicle in the simulator does not seem to be related to the integration of the user in the virtual scenario.

Finally, regarding open source autonomous driving simulators we find several options such as CARLA (Dosovitskiy et al., 2017), AirSim (Shah et al., 2018) or DeepDrive (DeepDrive, 2018), with CARLA and AirSim being probably the ones that are receiving more attention and have experienced the largest growth in recent years. Although AirSim works natively with virtual reality (originally intended as flight simulator), CARLA does not support

virtual reality natively. The introduction of immersive virtual reality with CARLA has only recently been proposed for the driver use case (Silvera et al., 2022). To the best of our knowledge, this is the first work that provides a solution to integrate subjects into the CARLA simulator using immersive virtual reality for the vulnerable road users case, including motion capture, visual body feedback and integration of the articulated human pose into the simulator.

3 VR IMMERSION FEATURES

This section describes the features of the immersive virtual reality system in the CARLA simulator for autonomous driving. Full pedestrian immersion is achieved by making use of the functionality provided by UE4 and external hardware, such as VR glasses and a set of motion sensors, for behaviour and interaction research.

The CARLA open-source simulator is implemented over UE4, which provides high rendering quality, realistic physics and an ecosystem of interoperable plugins. CARLA simulates dynamic traffic scenarios and provides an interface between the virtual world created with UE4 and the road agents operating within the scenario. CARLA is designed as a server-client system to make this possible, where the server runs the simulation and renders the scene. Communication between the client and the server is done via sockets.

The main features for the insertion of real agent behaviours in the simulation are based on five points (depicted in Fig. 2): 1) **Avatar control**: from the CARLA's blueprint library that collects the architecture of all its actors and attributes, we modify the pedestrian blueprints to create an immersive and maneuverable VR interface between the person and the virtual world; 2) **Body tracking**: we use a set of inertial sensors and proprietary external software to cap-

ture the subject's motion through the real scene, and **motion perception**: we integrate the avatar's motion into the simulator via *.bvh* files; 3) **Sound design**: since CARLA is an audio-less simulator, we introduce positional sound into the environment to enhance the immersion of the person; 4) **eHMI integration**: enabling the communication of state and intent information from autonomous vehicles to address interaction studies; 5) **Scenario simulation**: we design traffic scenarios within CARLA client, controlling the behaviour of vehicles and pedestrians.

3.1 Avatar Control

CARLA's blueprints (that include sensors, static actors, vehicles and walkers) have been specifically designed to be managed through the Python client API. Vehicles are actors that incorporate special internal components that simulate the physics of wheeled vehicles and can be driven by functions that provide driving commands (such as throttle, steering or braking). Walkers are operated in a similar way and its behavior can be directed by a controller. However, they are not designed to adopt real behaviors and therefore do not include wide range of unexpected movements and an immersive interface.

We modify a walker blueprint to make an IK setup (inverse kinematics) for VR (full body room-scale) to support this. The tools available to capture the real human movement are the following: *a) Oculus Quest 2* (for head tracking and user position control), and *b) Motion controllers* (for both hands tracking). Oculus Quest 2 already has a safety distance system, delimiting the playing area through which the performer can move freely. The goal is to allow the performer to move within the safety zone that s/he has established, corresponding to a specific area of CARLA's scenario.

The first step to achieve immersion is to modify the blueprint of the walker by attaching a virtual camera to its head to get a first-person feeling. The camera functions as the spectator's vision, while the skeletal mesh defines the appearance the walker adopts. The displacement and perspective of the walker are activated, from certain minimum thresholds, with the translation and rotation of the VR headset, which are applied to the entire skeletal mesh. In addition, the VR glasses and both motion controllers adapt the pose of the neck and hand in real time.

VR immersion for pedestrians is provided by implementing a head-mounted display (HMD) and creating an avatar in UE4. The subject wearing the VR glasses and controlling the avatar has complete freedom of movement within the preset area where we

conduct the experiments.

3.2 Full-body Tracking

To control the avatar (i.e., to represent the movement and pose of the subject inside the simulator), head (VR headset) and hand (motion controllers) tracking are not enough. Some kind of motion capture (MoCap) system is needed. There are multiple options, including vision-based systems with multiple cameras and inertial measurement units (Menolotto et al., 2020). In our case, an inertial wireless sensor system has been chosen, namely the Perception Neuron Studio (PNS) motion capture system (Noitom, 2022), as a compromise solution between accuracy and usability. Each MoCap system includes a set of inertial sensors and straps that can be put on the joints easily, as well as a software for calibrating and capturing precise motion data. It also allows integration with other 3D rendering and animation software, such as iClone, Blender, Unity or UE4.

3.3 Sound Design

Other techniques for total immersion, such as sound design and real-world isolation, are other important aspects of behavioral research. World audio is absent in the CARLA simulator, but it is essential for interaction with the environment, as humans use spatial sound cues to track the location of other actors and predict their intentions.

We incorporate positional audio into the simulator, attaching a sound to each agent participating in the scene. Walking and talking sounds are assigned for pedestrians, and engine sounds parameterized by throttle and brake for vehicles. In addition, we add environmental sounds such as birds singing, wind and traffic sounds.

3.4 External Human-Machine Interfaces (eHMI)

Communication between road users is an essential factor in traffic environments. In our experiments we include external human-machine interfaces (eHMI) for autonomous vehicles to communicate their status and intentions to the actual road user. The proposed eHMI design consists of a light strip along the entire front of the car, as depicted in Fig. 3. This allows studying the influence of the interface on decision making when the pedestrian's trajectory converges with the one followed by the vehicle in the virtual scenario.



Figure 3: Left: vehicle without eHMI. Right: vehicle with eHMI activated.

3.5 Traffic Scenario Simulation

CARLA offers different options to simulate traffic and specific traffic scenarios. We use the Traffic Manager module to populate a simulation with realistic urban traffic conditions. The control of each vehicle is carried out in a specific thread. Communication with other layers is managed via synchronous messaging.

We can control traffic flow by setting parameters that enforce specific behaviors. For example, cars can be allowed to exceed speed limits, ignore traffic light conditions, ignore pedestrians, or force lane changes.

The subject is integrated into the simulator on a map that includes a 3D model of a city. Each map is based on an OpenDRIVE file that describes the fully annotated road layout. This feature allows us to design our own maps and reproduce the same traffic scenarios in real and virtual environments, to evaluate the integration of real behavior in the simulator and to be able to carry out presence studies by comparing the results of the interaction.

3.6 Recording, Playback and Motion Perception

When running experiments, simulation data must be recorded and replayed for later analysis. CARLA has a native record and playback system that serializes the world information in each simulator tick for post-simulation recreation. However, this is only intended for tracking actors managed by the Python API and does not include the subject avatar or motion sensors.

Along with the recording of the state of the CARLA world, in our case the recording and playback of the actual full-body motion of the subject is essential. In our approach we use the Axis Studio software provided by Perception Neuron to record the body motion during experiments where the subject has to interact with the traffic scene in virtual reality. The recording is then exported in a *.bvh* file which is subsequently integrated into the UE4 editor.

Once the interaction is recorded, the simulation is played back so that the simulated sensors of the autonomous vehicles (i.e., radar, LiDAR, cameras) not only perceive the avatar's skeletal mesh and its dis-

placement, but also the specific pose of all its joints (i.e., body language).

4 SYSTEM IMPLEMENTATION AND RESULTS

The overall scheme of the system is depicted in Fig. 4. In the following we describe the hardware and software configurations and present preliminary results.

4.1 Hardware Setup

The overall hardware setup is shown in Fig. 5. During experiments, we use the Oculus Quest 2 as our head-mounted device (HMD), created by Meta, which has 6GB RAM processor, two adjustable 1832 x 1920 lenses, 90Hz refresh rate and an internal memory of 256 GB. Quest 2 features WiFi 6, Bluetooth 5.1, and USB Type-C connectivity, SteamVR support and 3D speakers. For full-body tracking we use PNS packaged solution with inertial trackers. The kit includes standalone VR headset, 2 motion controllers, 17 Studio Inertial Body sensors, 14 set of straps, 1 charging case and 1 Studio Transceiver.

To conduct the experiments, we reserved a preset area wide enough and free of obstacles where the subject can act as a real pedestrian inside the simulator. Quest 2 and motion controllers are connected to PC via Oculus link or WiFi as follows:

- Wired connection: via the Oculus Link cable or other similar high quality USB 3.
- Wireless connection: via WiFi by enabling Air Link from the Meta application, or using Virtual Desktop and SteamVR.

The subject puts on the straps of the appropriate length and places the PN Studio sensors into the bases. The transceiver is attached to the PC via USB. After launching the build version of CARLA for Windows, Quest 2 enables "VR Preview" in the UE4 editor.

4.2 Software Setup

VR Immersion System is currently dependent on UE4.24 and Windows 10 OS due to CARLA build, and Quest 2 Windows-only dependencies. Using TCP socket plugin, all the actor locations and other useful parameters for the editor are sent from the Python API to integrate, for example, the sound of each actor or the eHMI of the autonomous vehicle. "VR Preview" launches the game on the HMD.

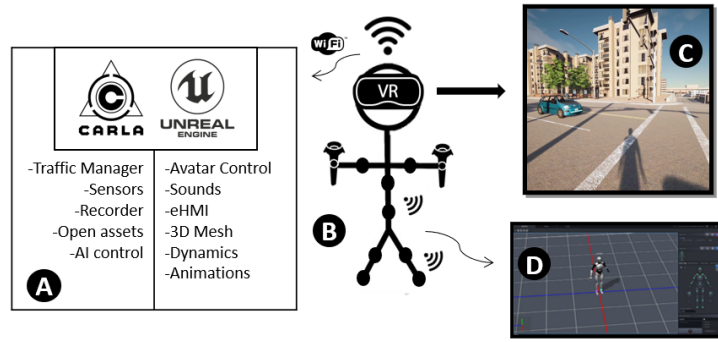


Figure 4: System Schematics. (A) Simulator CARLA-UE4. (B) VR headset, motion controllers and PN Studio sensors. (C) Spectator View in Virtual Reality. (D) Full-body tracking in Axis Studio.

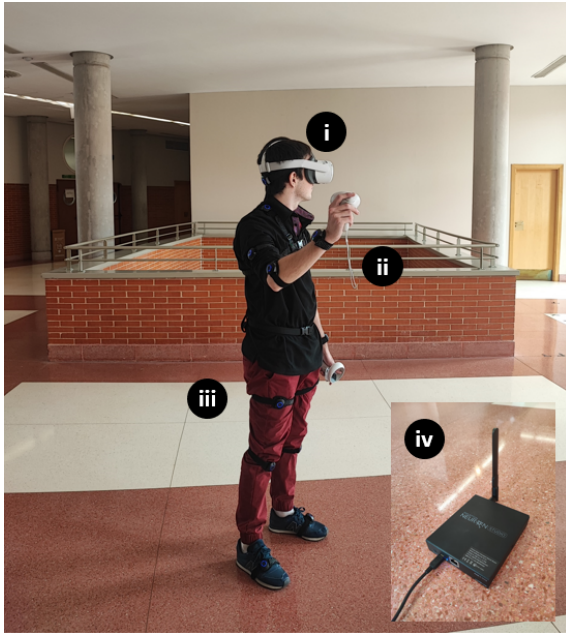


Figure 5: Hardware setup. (i) VR headset (Quest 2): transfer the image from the environment to the performer. (ii) Motion controllers: allow control of the avatar's hands. (iii) PN Studio sensors: provide body tracking withstanding magnetic interference. (iv) Studio Transceiver: receives sensors data wirelessly by 2.4GHz.

Perception Neuron Studio works with Axis Studio which supports up to 3 subjects at a time while managing up to 23 sensors for the body and fingers.

4.3 Results

The safety of automated driving functions depends highly on reliable corner cases detection. A recent research work describes corner cases (CC) as data that occur infrequently, represent a dangerous situation and are sparse in the datasets (Daniel Bogdoll, 2021). We aim to integrate behaviours of real subjects in the generation of wide variety of scenarios with pedestri-

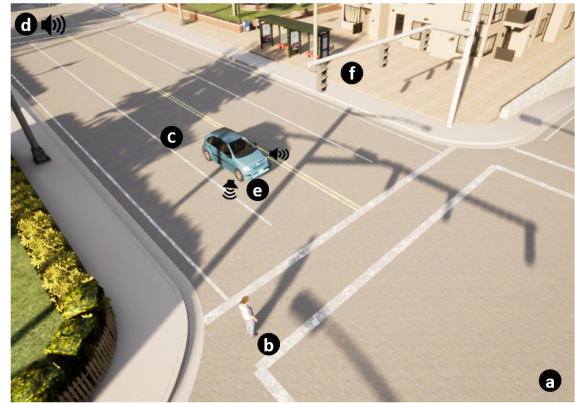


Figure 6: Simulation of Interactive traffic situations. (a) 3D world design. (b) Pedestrian matches the performer avatar. (c) Autonomous vehicle. (d) Environment sounds and agents sounds. (e) eHMI. (f) Traffic lights and traffic signs.

ans involved, including CC, in order to provide training data with real-time interaction. The preliminary scenario designed includes the interaction between a single autonomous vehicle and a single pedestrian. The vehicle circulates on the road when it reaches a pedestrian crossing where the pedestrian shows clear intentions to cross. The pedestrian receives information on status and vehicle intentions through an eHMI and can track its location by listening to the engine sound. Lighting and weather conditions are favorable (sunny day).

It is possible to spot the vehicle a few metres ahead of the pedestrian crossing before starting the action. When the pedestrian decides whether or not to wait for the vehicle to come to a complete stop, s/he pays attention to the speed and deceleration of the vehicle before it reaches the intersection. Sensors attached to the vehicle capture the image of the scene as shown in Fig. 7, and detect the pedestrian as shown in LiDAR point cloud in Fig. 8.

As a preliminary usability assessment, six real subjects completed a 15-item presence scale consist-

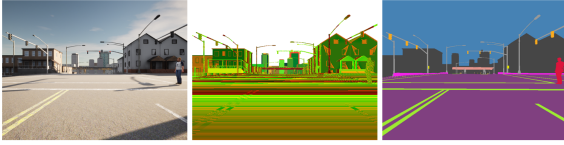


Figure 7: Cameras RGB / Depth / Segmentation.

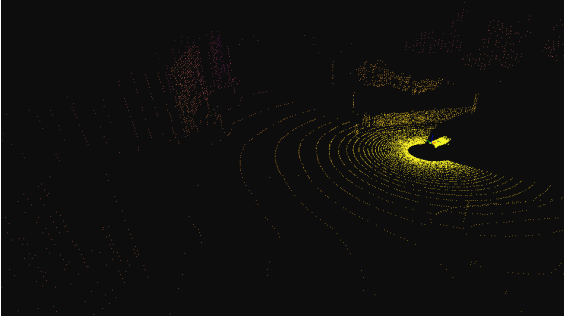


Figure 8: LiDAR point cloud (ray-casting).

ing in five items for self-presence, five items for autonomous vehicle presence and five items for environmental presence. The questions raised are depicted in Appendix A. In addition, we request subjective feedback about system usability and user experience.

Most of the participants felt a strong self-presence ($M=3.97$, $SD=0.528$), fitting into the avatar's body. Nevertheless, some of them were skeptical of the damage the avatar was exposed to (such as being run over) and did not feel under any real threat. Regarding autonomous vehicle presence ($M=3.60$, $SD=0.657$), the engine's sound composed the point of greatest disagreement among the participants, since some found it very useful to locate the vehicle, while others did not even notice or found it annoying. The image of the vehicle did not look completely clear-cut until the vehicle was close enough to the avatar, and its stop in front of the pedestrian cross did not feel threatening in the sense that its braking was appreciated too conservative. The environmental presence ($M=4.63$, $SD=0.266$) was felt very strongly by all the participants that felt really involved in an urban environment and commented that adding other agents or more traffic on the road would enhance the experience.

VR Immersion allows to integrate real participants in interactive situations on the road and to use objective metrics (such as collisions and reaction times) in order to measure interaction quality without real danger. CARLA enables custom traffic scenarios while UE4 provides the immersive and interactive interface into the same simulator.

5 CONCLUSIONS

In this paper, we have presented a preliminary framework to enable real-time interaction between real agents and the simulated environment. At this level, more results are expected from testing interactive traffic scenes with multiple real participants.

Regarding the immersive sensation, the virtual environment displayed from the glasses varies between 15 and 20 fps, showing a stable image to the performer that allows him/her to interact with the world of CARLA. The wireless connection is prioritized over the wired connection, since it allows a greater degree of freedom of movement. The performer pose within the virtual environment is registered by Perception Neuron, generating useful sequences to train and validate predictive models to, for example, predict future actions and trajectories of traffic agents.

The scenario shown in this paper serves as a basis for a preliminary presentation of the proposed methodology. From here, as future work, it is intended, on the one hand, to carry out a study of presence, interaction and communication with equivalent real and virtual environments. And on the other hand, to develop a data recording campaign with multiple subjects, different traffic scenarios and varied environmental conditions, in order to generate a dataset that can be used to train action and trajectory prediction models.

ACKNOWLEDGEMENTS

This work was mainly supported by the Spanish Ministry of Science and Innovation (Research Grant PID2020-114924RB-I00) and in part by the Community Region of Madrid (Research Grant 2018/EMT-4362 SEGVAUTO 4.0-CM).

REFERENCES

- Bhagavathula, R., Williams, B., Owens, J., and Gibbons, R. (2018). The reality of virtual reality: A comparison of pedestrian behavior in real and virtual environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1):2056–2060.
- Bruder, G., Steinicke, F., Rothaus, K., and Hinrichs, K. (2009). Enhancing presence in head-mounted display environments by visual body feedback using head-mounted cameras. In *2009 International Conference on CyberWorlds*, pages 43–50.
- Chang, C.-M., Toda, K., Sakamoto, D., and Igarashi, T. (2017). Eyes on a car: an interface design for communication between an autonomous car and a pedestrian.

- In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'17)*, pages 65–73.
- Daniel Bogdoll, Jasmin Breitenstein, F. H. M. B. B. S. T. F. J. M. Z. (2021). Description of corner cases in automated driving: Goals and challenges. *arXiv*.
- Deb, S., Carruth, D. W., Fuad, M., Stanley, L. M., and Frey, D. (2020). Comparison of child and adult pedestrian perspectives of external features on autonomous vehicles using virtual reality experiment. In Stanton, N., editor, *Advances in Human Factors of Transportation*, pages 145–156. Springer International Publishing.
- Deb, S., Carruth, D. W., Sween, R., Strawderman, L., and Garrison, T. M. (2017). Efficacy of virtual reality in pedestrian safety research. *Applied Ergonomics*, 65:449–460.
- DeepDrive (2018). Deepdrive.
- Doric, I., Frison, A.-K., Wintersberger, P., Riener, A., Wittmann, S., Zimmermann, M., and Brandmeier, T. (2016). A novel approach for researching crossing behavior and risk acceptance: The pedestrian simulator. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'16)*, pages 39–44.
- Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16.
- Eady, T. (2019). Simulations can't solve autonomous driving because they lack important knowledge about the real world – large-scale real world data is the only way.
- Farooq, B., Cherchi, E., and Sobhani, A. (2018). Virtual immersive reality for stated preference travel behavior experiments: A case study of autonomous vehicles on urban roads. *Transportation Research Record*, 2672(50):35–45.
- Fernández-Llorca, D. and Gómez, E. (2021). Trustworthy autonomous vehicles. *Publications Office of the European Union, Luxembourg*, EUR 30942 EN, JRC127051.
- Gruenefeld, U., Weiß, S., Löcken, A., Virgilio, I., Kun, A. L., and Boll, S. (2019). Vroad: Gesture-based interaction between pedestrians and automated vehicles in virtual reality. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, AutomotiveUI '19, page 399–404.
- Holländer, K., Wintersberger, P., and Butz, A. (2019). Overtrust in external cues of automated vehicles: An experimental investigation. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, page 211–221, New York, NY, USA. Association for Computing Machinery.
- Iryo-Asano, M., Hasegawa, Y., and Dias, C. (2018). Applicability of virtual reality systems for evaluating pedestrians' perception and behavior. *Transportation Research Procedia*, 34:67–74.
- Kalra, N. and Paddock, S. M. (2016). Driving to safety: how many miles of driving would it take to demonstrate autonomous vehicle reliability? Research report, RAND Corporation.
- Löcken, A., Golling, C., and Riener, A. (2019). How should automated vehicles interact with pedestrians? a comparative analysis of interaction concepts in virtual reality. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '19, page 262–274.
- Martinez, A. H., Díaz, J. L., Daza, I. G., and Llorca, D. F. (2021). Data-driven vehicle speed detection from synthetic driving simulator images. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2617–2622.
- Menolotto, M., Komaris, D.-S., Tedesco, S., O'Flynn, B., and Walsh, M. (2020). Motion capture technology in industrial applications: A systematic review. *Sensors*, 20(19).
- Nazemi, M., van Eggermond, M., Erath, A., Schaffner, D., Joos, M., and Axhausen, K. W. (2021). Studying bicyclists' perceived level of safety using a bicycle simulator combined with immersive virtual reality. *Accident Analysis & Prevention*, 151:105943.
- Nguyen, T. T., Holländer, K., Hoggenmueller, M., Parker, C., and Tomitsch, M. (2019). Designing for projection-based communication between autonomous vehicles and pedestrians. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '19, page 284–294. Association for Computing Machinery.
- Noitom (2022). Perception neuron studio system.
- Núñez Velasco, J. P., de Vries, A., Farah, H., van Arem, B., and Hagenzieker, M. P. (2021). Cyclists' crossing intentions when interacting with automated vehicles: A virtual reality study. *Information*, 12(1).
- Núñez Velasco, J. P., Farah, H., van Arem, B., and Hagenzieker, M. P. (2019). Studying pedestrians' crossing behavior when interacting with automated vehicles using virtual reality. *Transportation Research Part F: Traffic Psychology and Behaviour*, 66:1–14.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243.
- Shah, S., Dey, D., Lovett, C., and Kapoor, A. (2018). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Hutter, M. and Siegwart, R., editors, *Field and Service Robotics*, pages 621–635. Springer International Publishing.
- Silvera, G., Biswas, A., and Admoni, H. (2022). Dreyevr: Democratizing virtual reality driving simulation for behavioural & interaction research.
- Ullmann, D., Kreimeier, J., Götzmann, T., and Kipke, H. (2020). Bikevr: A virtual reality bicycle simulator towards sustainable urban space and traffic planning. In

- Proceedings of the Conference on Mensch Und Computer, MuC '20*, page 511–514. Association for Computing Machinery.
- Vázquez, D., López, A. M., Marín, J., Ponsa, D., and Gerónimo, D. (2014). Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):797–809.
- Weng, X., Man, Y., Park, J., Yuan, Y., Cheng, D., O'Toole, M., and Kitani, K. (2021). All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. *arXiv*.
- Xu, J., Lin, Y., and Schmidt, D. (2017). Exploring the influence of simulated road environments on cyclist behavior. *The International Journal of Virtual Reality*, 17(3):15–26.
4. The urban environment seemed like the real world.
 5. It could reach out and touch the objects in the urban environment.

APPENDIX A

Self-presence Scale Items

To what extent did you feel that... (1= not at all – 5 very strongly)

1. You could move the avatar's hands.
2. The avatar's displacement was your own displacement.
3. The avatar's body was your own body.
4. If something happened to the avatar, it was happening to you.
5. The avatar was you.

Autonomous Vehicle Presence Scale Items

To what extent did you feel that... (1= not at all – 5 very strongly)

1. The vehicle was present.
2. The vehicle dynamics and its movement were natural.
3. The sound of the vehicle helped you to locate it.
4. The vehicle was aware of your presence.
5. The vehicle was real.

Environmental Presence Scale Items

To what extent did you feel that... (1= not at all – 5 very strongly)

1. You were really in front of a pedestrian crossing.
2. The road signs and traffic lights were real.
3. You really crossed the pedestrian crossing.

Scale Perception in VR for Urban Scale Environments: 360° Photos versus Full 3D Virtual Reality

Claus B. Madsen¹^a, Nicolai Steinø²^b, Andrei Iuliu Lucaci¹, Emil O. Sandkvist¹
and Alexander Jonstrup¹

¹Computer Graphics Group, Aalborg University, Aalborg, Denmark

²Urban Design Group, Aalborg University, Aalborg, Denmark
{cbm, nic}@create.aau.dk

Keywords: Virtual Reality, Perception, Distance Estimation, Architectural Visualization, Depth Cues.

Abstract: The paper investigates the accuracy of scale perception in Virtual Reality (VR) for visualization of urban scale environments. Specifically, we evaluate users' scale estimation accuracy by subjecting them to an urban scale environment using two different viewing modes: 360° monoscopic panoramas viewed in a VR headset, versus a full stereoscopic 3D VR representation, also in a VR headset, allowing participants to move around. The paper explores various aspects of this, including both ego- and exo-centric distance estimation, perception of own height, and photographic realism of 360° modes. The main experimental result is that, somewhat surprisingly, user distance estimation accuracy is higher in the monoscopic 360° viewing modes than in 3D VR; in 3D VR participants on average underestimate distances by around 20%. Nevertheless, participants on average feel significantly taller than normal in the 360° modes, whereas they feel normal height in 3D VR mode. We conclude that more work is needed in order to properly understand the perceptual and cognitive mechanisms behind scale perception in VR.

1 INTRODUCTION

Proposals for architectural projects in urban space are often subject to public debate. Therefore, architects, and their clients through them, put effort into visualizing such projects in the most favourable way. Hence, architectural visualizations become important artifacts of such debates. The evaluation of architectural visualisations may ultimately determine the destiny of proposals, as they may impact decision-making. Thus, it may be argued that they not only influence the destiny of architectural projects, but also the democratic processes leading to their approval or dismissal.

With the advancements in VR in recent years, architectural VR is rapidly becoming a feasible supplement to traditional architectural visualization in the form of still images and animations. Due to the immersive nature of VR, it is potentially more persuasive than traditional forms of architectural visualization. Therefore, faithful scale representation is of the essence. To faithfully represent scale in VR, nonethe-

less, is not an innocent endeavour. The evaluation of scale depends on multiple factors which relate to the design of the immersive environment – i.e. the 3D model – which is displayed, to the display technology itself, and, as addressed in the context of this paper, the mode of VR representation.

In this paper, we investigate a specific aspect of scale perception in VR. In particular, we compare 360° panoramic representations to a full 3D modelled VR experience of the same urban scale scene. The rationale behind the study can be formulated as this:

- VR visualization of architectural projects is becoming increasingly popular, and, as argued above, it has the potential to be a powerful tool in such contexts
- Scale perception evaluated via distance estimation has been proved to be compromised in full 3D VR. Test participants typically underestimate distances in VR by around 20%, according to the literature on the subject
- 360° panoramas are a relevant alternative to full 3D VR, as they are potentially more easy to generate and more flexible to use, i.e. with stand-alone VR headsets and smartphone based systems such

^a <https://orcid.org/0000-0003-0762-3713>

^b <https://orcid.org/0000-0001-5272-1879>

as Google Cardboard

- To our knowledge no studies exist that compare scale perception across those two viewing modes (360° panorama vs. 3D VR)



Figure 1: 3D model of section of pedestrian street used for the experiment on scale perception.

Over the past decade or so, the topic of scale perception or distance perception in VR has been extensively researched, and yet there are still many aspects that are poorly understood. Most research indicates that distance perception in VR is compromised. Specifically, the consensus is that distance perception in VR is biased towards underestimation, (Renner et al., 2013), (Interrante et al., 2006), (Loyola, 2017), (Ng et al., 2016) (Peer and Ponto, 2017), (Hornsey and Hibbard, 2021), (Jensen et al., 2020). The literature even speaks of “distance misperception” and “distance compression”, (Peer and Ponto, 2017). Most literature on the subject covers room scale experiments, i.e., personal space (0m to 2m), and the lower part of action space (2m to 30m), (Lucaci et al., 2022), (Renner et al., 2013). Vista space (+30m) is still quite unexplored, but highly relevant for urban scale architectural visualization. In this paper we address the mid to upper part of action space, by exploring scenes that extend to around 20m to 30m from the test participant. The use of 360° visual representation has also been compared to 3D VR solutions, but primarily from an exploration and immersion point of view, not from the point of view of evaluation of scale perception, (Boukhris et al., 2017), (Ritter and Chambers, 2021).

The main contribution of this paper is that we experimentally demonstrate that there is a significant difference in how accurately test participants estimate distances in the two modes. Our experiment surprisingly shows that when exploring 360° panoramas, participants can estimate distances quite accurately, whereas, given a full 3D VR experience, participants underestimate distances by about 20%.

2 VIEWING MODES

As mentioned, the aim of this paper is to investigate how viewing modes influence scale perception in virtual representations of small urban sub-environments, where the key components of the environment are well within the participant’s action space. By viewing modes, we mean different “formats” in which the user gets to visually experience the environment. We study two fundamentally different viewing modes, each of which have two “sub-modes”.

2.1 360° Photo



Figure 2: 360° photo of pedestrian street environment.

This viewing mode is included to investigate if photographic visual fidelity influences scale perception, Figure 2. The 360° photos are shown to test participants in a Head Mounted Display (HMD). In this paper all experiments are performed using Oculus Quest 2 HMDs. Both eyes are shown the same image, and hence the experience does not provide stereoscopic information. A Ricoh Theta Z1 360° camera was used to capture all the 360° photos used in the tests. Each of the 360° photos has a resolution of 6720x3360 pixels. Photos were taken from a height of 1.8 meters.

2.2 360° CityEngine



Figure 3: 360° visualization of pedestrian street environment rendered in CityEngine.

This viewing mode, rendered using the procedural modeler CityEngine 2021.1, compares directly to the

360° photo mode, although the visual aesthetics are very close to those of the VR viewing modes (see below). The 360° panoramas are presented to participants in an HMD and do not provide stereoscopic information. The 360° CityEngine panoramas are rendered from a height of 1.8 meters, Figure 3.

2.3 3 DoF Unreal Engine



Figure 4: VR visualization of pedestrian street environment rendered in UnrealEngine; user is not allowed to move around.

The 3 Degrees of Freedom (DoF) Unreal viewing mode, Figure 4, is rendered in real-time using the Unreal game engine. It is denoted 3 DoF, since the test participant is experiencing the mode wearing an HMD, but any translation done by the user does not influence viewing position in the VR experience, i.e. the participant only has 3 rotational degrees of freedom. Hence, from a mobility perspective, this mode corresponds to the 360° modes, but offers stereoscopic information.

2.4 6 DoF Unreal Engine



Figure 5: VR visualization of pedestrian street environment rendered in UnrealEngine; user is allowed to move around.

The 6 DoF Unreal viewing mode, Figure 5, is rendered in real-time using the Unreal game engine. The participant has 6 DoF, i.e. can move around freely (within an approx. 3x3m exploration area). The mode

is experienced wearing an HMD, providing stereoscopic information. Teleportation is not enabled; the participant can only move within the exploration area.

2.5 Summary of the Viewing Modes

Intuitively, we would expect test participants to struggle with scale perception in the 360° modes, since there is no binocular stereo, nor any motion parallax from moving around. The 360° photo mode might offer an advantage over the 360° CityEngine mode given the photographic realism and higher density of pictorial cues. Conversely, we would expect test participants to perform better in the 3 DoF Unreal mode than in the 3 DoF CityEngine; they have similar aesthetics and amount of pictorial cues, but the former offers binocular stereoscopic information. Finally, the 6 DoF Unreal mode offers binocular information and motion parallax arising from test participants' translation motion, and thus participants should have good conditions for scale perception.

3 SCALE PERCEPTION IN VR

Virtual Reality is most commonly experienced using a HMD, for example the Oculus Quest 2 used for this research. The performance of such ordinary consumer HMDs is now truly amazing compared to just a few years ago, and consumer VR is definitely becoming mainstream. Good as they may be, commercially available HMDs suffer a wide range of challenges from a perceptual point of view, which constitute a challenge in terms of allowing users to faithfully/realistically experience the dimensions, or sense of scale, as it were, of virtual representations of architectural spaces, be it indoor or outdoor.

The most obvious challenges are related to sensory aspects: limited field-of-view, limited image resolution (although very high end HMDs, such as the Varjo VR and XR series, are now coming close to retinal image resolution), limited dynamic range (inability to realistically represent the dynamic range of luminance levels known from the physical world), the so called vergence/accommodation conflict (HMDs' inability to provide a natural relationship between where your eyes converge and where they accommodate), etc.

Slightly less obvious are challenges that relate more to a perceptual level: lacking visual representation of the user's body inside the VR experience (you do not see yourself when looking down), lacking ability to move around across larger distances (VR tracking typically only supports a few meters of positional

movement, which is insufficient for naturally moving around in a large indoor VR scene, let alone an outdoor scene), lacking or inconsistent audio (humans' perception of space is not only visual but also relies heavily on audio perception of reverberation effects; this is often overlooked in VR representations of architectural use cases).

In essence, the challenges regarding faithful perception of space, distance, scale, etc. in VR representations are substantial. Quite specifically, current research indicates that distance perception in VR is compressed, i.e., users' estimation of distances in VR are approximately 20% lower than the same users' estimation of distances in real life, (Jensen et al., 2020). Being able to accurately estimate distances is not the only important factor in perceiving scale, but it is obviously related.

4 METHODS

In this section we describe the approaches we have taken, and the choices we have made, towards designing a way to experimentally evaluate how viewing mode influences accuracy of distance estimation. First we discuss our approach to letting test participants evaluate distances in the various modes. Subsequently we describe central aspects of the technical implementation behind the experiments.

4.1 Estimating Distances in Viewing Modes

The literature on distance estimation in VR separates distances into ego-centric distances (distances from one self to some location in the environment) and exo-centric distances (distances between two environment locations), (Renner et al., 2013). We believe both to be equally important for the purpose of evaluating urban scale architecture in VR, and hence include both types in our experiment (Fig. 3). The literature is also extensive in terms of which method to apply when getting test participants to estimate experienced distances, e.g. (Peer and Ponto, 2017). Examples of applied methods are verbal reporting, blind walking, throwing, etc. We opt for verbal reporting for two main reasons: 1) it is the method that best suits having 360° viewing modes in the experiment, and 2) blind walking, although most popular for VR research, is not realistic for scenarios of mid to upper action space dimension (10m or higher), as it can be challenging to find a suitable environment for carrying out the experiment.

In terms of test participant locomotion in the 6 DoF viewing mode, we opted to avoid teleportation. Only 1-to-1 physical movement is possible for exploring a local area of the virtual 3D scene, and participants are only allowed to move within an approximately 3m by 3m area, and there is a mark on the ground in the virtual environment, where the participants is to return to prior to verbally reporting on questions regarding estimated distances.

A final important thing regarding what the experiment should entail, concerns test participants' perception of their own height in the different viewing modes. As described, the two 360° modes are presented to participants as monocular experiences. This causes perceptual confusion, especially when looking down on the ground, as the only perceptually plausible explanation for experiencing no binocular disparity is that what is viewed must be located at infinity. Thus, participants should subjectively feel "floating" high above the ground. For the 3 DoF and 6 DoF VR modes, test participants do perceive the correct stereo disparities, but, similarly to the 360° modes, there is no visual representation of self when looking down (you do not see your own legs and feet). For these reasons we included into the experiment that test participants would be asked to verbally answer whether they felt shorter than normal, normal height, or taller than normal.

4.2 3D Model

The VR models for the user tests are made using a workflow developed for the production of architectural urban VR scenarios using parametric urban design to feed a VR model. This is done within a software framework comprising open GIS data (as a basis for the parametric generation of real-world urban environments), the CityEngine (CE) parametric urban modeler (for the parametric generation of 3D urban models), and the Unreal Engine (UE) game engine (for the preparation of the final rendered VR model).

While the parametric modeling approach potentially allows for model representation at different levels of detail (LOD) and texture sets, for the tests presented in this paper, a consistent LOD and texturing was chosen using the metaphor of two types of cardboard (plain and corrugated) architectural scale model. Geometrically, detailing was limited to adding windows, doors, pitched roofs, and cornices to buildings. Ground surfaces were textured using the same corrugated cardboard texture as for the buildings with no accentuation of curbs or other 3D features of the horizontal plane. While the 3D model represents a real-world urban space, which, in reality, has a

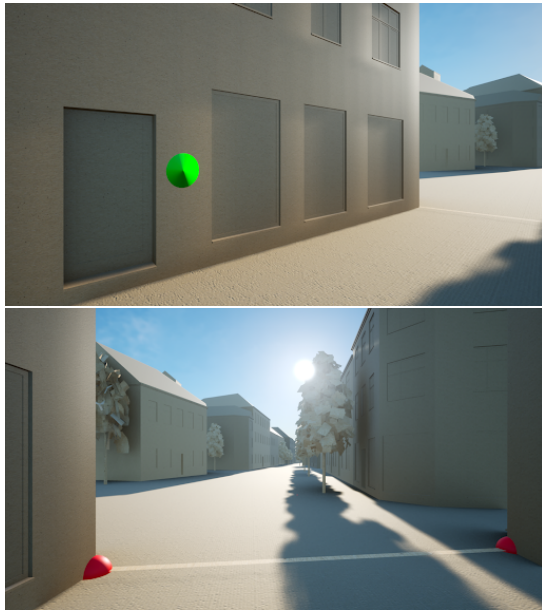


Figure 6: Top: ego-centric distance estimation is evaluated by asking participants what they believe to be the distance from themselves to a feature indicated in the environment. Bottom: exo-centric distance estimation is evaluated by asking participants what they believe to be the distance between two indicated features in the environment. These examples are from the 6 DoF Unreal viewing mode; the same approach is adopted for all viewing modes.

slightly sloping ground, the model was simplified so that buildings were projected onto a perfectly horizontal plane.

Compared to a simple 3D model where buildings are represented as simple, white boxes, the addition of geometric (3D) architectural detailing as described offers depth cues. While doors and building floors (as indicated by the vertical distribution of windows) offer scale cues which relate to the human body (human scale), niches for doors and windows generate cast shadows and ambient occlusion for spatial feel and depth cues. The choice of cardboard texturing (the architectural scale model metaphor), as opposed to more photorealistic texturing (i.e. real-world facade and pavement textures) was chosen in order to achieve a convincing feel of the model. In order to appear convincing, real-world texturing requires a high level of geometric detailing in terms of recesses and protrusions in facades, doors and other building parts. The same is true for pavements with all the geometric imperfections which they typically have in the real world.

The assumption is, that when choosing the architectural scale model metaphor instead, the VR model will be more convincing to the eye, as a scale model from simple materials like plain and corrugated card-

board is not expected to have the same level of detail as a real-world urban space. In other words, people are more tolerant with architectural scale models – they do not expect them to be “photo realistic” – and thus accept less detail.

4.3 Viewing Modes

The real-world 360 Photo testing environment is created in Unreal Engine as a map separate from the other environments. This environment uses a self-illuminating two-sided material wrapped on a simple sphere object. The VR camera is placed in the origin point of the world, with the sphere acting as the environment. The 360° model panoramas are generated in CityEngine and displayed using the ArcGIS 360 web-based VR viewing platform with the Google Chrome browser. Hence, this experimental setup is used outside of Unreal Engine. The 3DoF and 6DoF viewing modes are created in Unreal Engine and are composed of a map that contains the imported 3D model from CityEngine and a map containing the lighting conditions. In this way, the model can be re-imported or replaced under the same dynamic lighting conditions.

4.4 Lighting

The Virtual Reality Environment (VRE) was achieved using Unreal Engine 4.27. The default shading method is used due to its costly yet higher render capability. Lighting in the scene is composed of a directional light of intensity 20 lux units, with a temperature of 5900 kelvin. Furthermore, a Sky Light component is used to capture further parts of the environment and apply them as light. In this component, real-time capture is used to achieve dynamic and specular lighting. The intensity of this component is set to 3 units. This represents the total energy emitted by the capture. A SkyAtmosphere component is used to simulate the atmosphere along with the light scattering associated with it. Two reflection probes are placed in the scene to capture the reflections. These probes are placed to cover all the areas around the testing zone. Lastly, the Engine Scalability settings and Material Quality Level are set to “Epic”, which is the highest graphical setting that can yield acceptable frame rates in VR.

4.5 Textures

The model is composed from two Physically Based Rendering (PBR) textures. The ground texture resembles corrugated cardboard whereas the building texture resembles plain cardboard. Both textures use a

world aligning method to the origin point instead of the classical UV coordinate method. This method allows for a more precise calibration of textures in regard to scale for static objects.

4.6 Navigation

Navigation through the different testing conditions is achieved using a User Interface that can be called by pressing the B button on the Oculus Quest’s right controller. The menu appears attached to the left controller and can be interfaced by both controllers.

5 EXPERIMENT DESIGN

The aim is to design an experiment to evaluate the accuracy and precision of distance estimation across the four viewing modes described above. Hence, from an experiment design perspective, viewing mode represents the independent variable (with four conditions) which we assume will have an influence on the dependent variable, in the form of the accuracy of participants’ estimated ego- and exo-centric distances.

The experiment comprises four different exo-centric distance features. Each feature is made up of two opposite building corners. On either side of these features there are pre-determined viewpoints, Figure 7, which are used for creating the 360° panoramas, and well as spawning points for the 3 and 6 DoF VR modes. Each participant experiences all four viewing modes, experiencing two features in each mode, totalling eight experiences. Hence, participants get to see all features twice, but from different feature viewpoints, and in different modes.

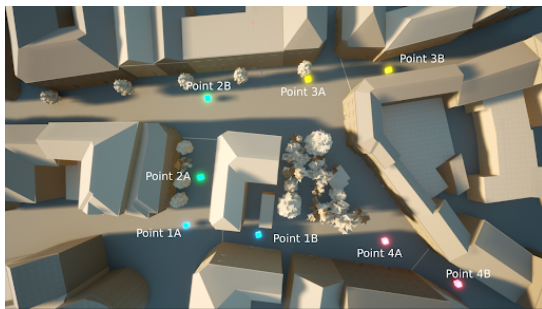


Figure 7: Overview of the urban environment modelled for the experiment, and the location’s eight viewpoints. The viewpoints form pairs (A and B) on either side of each exo-centric distance feature (1-4) for test participants to estimate.

As each participant experiences all modes, the experiment is based on a within-subject design. To alle-

viate a potential carry-over effect between the conditions, the order in which participants will be experiencing conditions is systematically arranged based on the sequence shown in Figure 8.

Order 1:	1a2b	3a4b	1b2a	3b4a
Order 2:	1b2a	3b4a	1a2b	3a4b
Order 3:	3a4b	1b2a	3b4a	1a2b
Order 4:	3b4a	1a2b	3a4b	1b2a
Order 5:	1b2a	3b4a	1a2b	3a4b
Order 6:	1a2b	3a4b	1b2a	3b4a
Order 7:	3b4a	1a2b	3a4b	1b2a
Order 8:	3a4b	1b2a	3b4a	1a2b

Figure 8: Viewing location and mode order sequence: Colors indicate viewing modes (red: 360° photo, yellow: 360° CityEngine, green: 3 DoF Unreal, blue: 6 DoF Unreal), numbers indicate locations (1 through 4), and letters indicate viewing locations (a or b). The sequence repeats after eight participants have been tested.

While the orders are not exhaustive in terms of all of the permutations possible, they are sufficiently “pseudo random” in the sense that within a full iteration of each order (1-8), each viewing location will occur exactly twice, and across the full sequence, each viewing mode will occur exactly twice for every viewing location, thus offering an even distribution of viewing locations for each participant, and an even distribution of combinations of viewing locations and modes across the experiment as a whole. With the presented scheme, each viewing mode will be the first experienced mode for one fourth of participants to rule out bias and adaptation. Each test participant will provide one “I feel shorter than normal, normal, taller than normal” opinion for each of four modes, as well as eight ego-centric distance estimates, and eight exo-centric distance estimates.

5.1 Participants

32 participants were recruited, 22 male and 10 female with ages ranging from 20y to 32y ($M=24.28y$, $SD=2.63y$) and heights ranging from 156cm to 195cm ($M=179.28cm$, $SD=9.36cm$). Four of the participants used contact lenses and five of the participants used glasses. 27 of the participants had prior experiences in using VR.

5.2 Procedure

Upon arrival, test participants were greeted and introduced to the experiment. They were asked to

sign a consent form allowing the data collected to be used for analysis, and a demographics questionnaire to enquire about relevant information such as their height, age, possible visual impairments and experience with using VR. Subsequently, they were escorted to the designated experiment area and equipped with an Oculus Quest 2 HMD. The experiment would begin and participants would go through each condition until all four conditions had been tested. For each condition, the participant would be asked the ego-centric distance to a predetermined point on a wall at the current viewing location followed by the exo-centric distance between two street corners. For the 6 DoF condition specifically, the participants were systematically encouraged to move around.

6 RESULTS AND DISCUSSION

We start the analysis of the results by looking into the test participants' evaluation of how tall they felt in the different modes. Getting a realistic sense of one's own height is very important for estimation of distances. Figure 9 clearly shows that test participants generally felt taller than normal in the 360° modes (the graph for 360° CityEngine mode is very similar but not shown here), as 22 out of 32 participants indicate feeling taller than normal. Conversely, 19 out of 32 report feeling normal height in the 6 DoF Unreal mode. The fact that 9 out of 32 participants feel shorter than normal in the 6 DoF Unreal mode might actually be an indication of distance compression, i.e. that participant distance perception is biased towards underestimation. All in all, though, the participants generally report feeling taller than normal in the 360° modes.

Turning towards the estimated ego- and exo-centric distances, Figure 10 summarizes those results. Qualitatively, merely from visual inspection of the plots, it is clear that ego- and exo-centric distance estimates follow the same overall pattern: distances are perceived quite accurately in the 360° modes as the medians of the normalized errors are very close to zero (all are 0.0 apart from 360° CityEngine ego-centric, where the median normalized error is -0.2, but we cannot statistically prove a significant difference between 360° photo and 360° CityEngine modes regarding ego-centric distances).

An ANOVA test shows that there are statistically significant differences between the distance estimates across modes for both ego- and exo-centric distances. The ANOVA test was configured as a one-way test with one independent variable (viewing mode) and 4 conditions (the 4 modes). The p-values for for

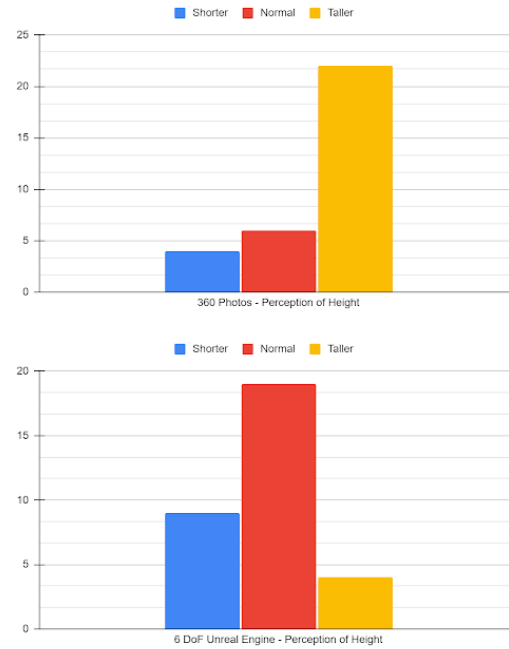


Figure 9: Top: test participants' evaluation of their own height in the 360° photo viewing mode. Bottom: similar for the 6 DoF Unreal viewing mode.

the Null hypothesis (users perform equally well in all modes) for the ego- and exo-centric cases were 0.00025 and 0.00013, respectively, indicating that equal performance in all viewing modes is highly unlikely.

A Tukey test was then performed to figure out which viewing mode performances were statistically significantly different from other viewing modes. This Tukey test showed that the 360° CityEngine mode performance is statistically significantly different from either of the Unreal modes (3 or 6 DoF). Hence participants, in this experiment, are statistically worse at estimating in 3 and 6 DoF Unreal. Or in other terms: they perform worse in full binocular VR than in the monocular 360° modes. The median errors for the 3 and 6 DoF Unreal modes, combined with ego- and exo-centric, are -0.23, -0.12, -0.29, and -0.19, respectively. So, roughly 20% underestimation.

This is a surprising result. Why would the absence of binocular information lead to better performance? Only more research and experimentation can answer this, but we may offer a few conjectures. Firstly, as mentioned previously, research has consistently demonstrated 3D VR environments to cause underestimation. So, our experiment is consistent with that. And for the 360° modes, which are experienced in the exact same HMD, participants subconsciously experience a similar HMD-induced effect, but the underestimation-effect might

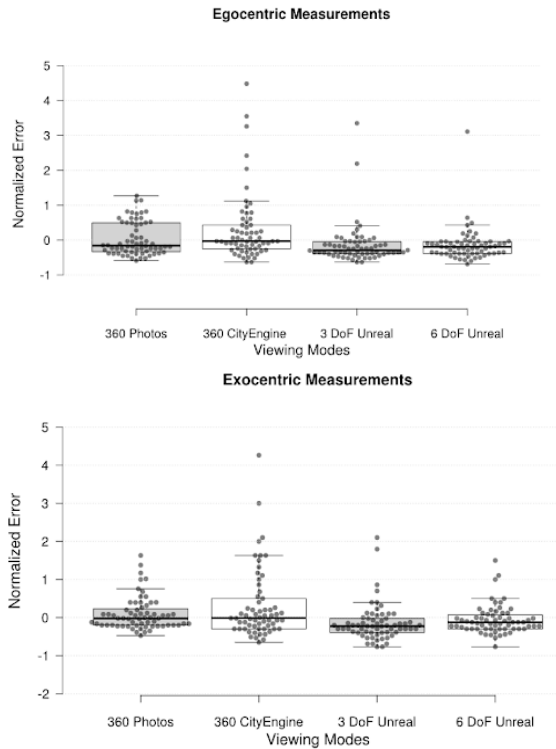


Figure 10: Top: ego-centric distance estimates across viewing modes. Bottom: exo-centric distance estimates. The y-axis indicates normalized error. This number is computed by subtracting the true value from the participant’s estimate, and dividing by the true value. Hence a value of 0 means a perfectly accurate distance estimate. A value of +1.0 corresponds to overestimating the distance by 100%, and value of -0.6 corresponds to underestimating the distance by 40%.

be offset/counter-acted by the “eerie” sense of being very tall in the 360° modes, combined with the perceptual stimuli of watching everything as being at infinity.

Our experiment did show some underestimation of ego-centric distances in 360° photo mode (median error -0.16), although not quite close enough for it to be statistically significant. So, there is a noticeable difference between ego- and exo-centric estimation. We might conjecture that distance estimation in the 360° modes is a more cognitive process than a perceptual process due to the conflicting sensory information and lower sense of immersion, leading participants to rely solely on higher level cognitive estimation based on familiarity with typical elements in the scenarios. Perhaps this functions better for exo-centric distances.

Another interesting aspect of our experiment is that the 360° photo and the 360° CityEngine results are statistically very similar, even if the visual appearances/aesthetics are extremely different. Photographic realism and detail, versus stylised com-

puter graphics. This is somewhat contradicting other research in the area, as consensus is that pictorial cues aid distance perception, (Hornsey and Hibbard, 2021). The difference in our experiment may lie in it being conducted in a mid to upper action space scale scene, as opposed to other research focusing on room scale (lower action space ranges).

7 CONCLUSIONS

The paper investigates distance perception in VR visualization of architecture at urban scale, i.e. outdoor spaces in the 15m to 30m range. More specifically, we investigate two different visualization modes: 360° panoramas and full 3D VR environments, both experienced wearing an HMD (Oculus Quest 2).

It is demonstrated experimentally that test participants systematically underestimate distances in 3D VR, which is in alignment with the distance compression consensus in the research literature. The underestimation is on the order of 20%. Conversely, the experiment shows that participants perform exocentric distance estimation very accurately in the 360° panorama mode, with a near-zero median error. Ego-centric distances appear to be compromised towards underestimation, but the experiment could not prove this with statistical significance.

The experiment does clearly show that test participants feel unusually tall when experiencing the 360° mode. We conjecture that this effect, combined with absence of binocular cues in the 360° mode, may counteract the HMD-induced distance compression.

ACKNOWLEDGEMENTS

This research was partially funded by the CityVR project, funded by the Department of Architecture, Design and Media Technology, Aalborg University. This funding is gratefully acknowledged.

REFERENCES

- Boukhris, M., Paljic, A., and Lafon-Pham, D. (2017). 360° versus 3D Environments in VR Headsets for an Exploration Task. In Lindeman, R. W., Bruder, G., and Iwai, D., editors, *ICAT-EGVE 2017 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association.
- Hornsey, R. L. and Hibbard, P. B. (2021). Contributions of pictorial and binocular cues to the perception of dis-

- tance in virtual reality. *Virtual Reality*, 25(4):1087–1103.
- Interrante, V., Ries, B., and Anderson, L. (2006). Distance perception in immersive virtual environments, revisited. In *IEEE Virtual Reality Conference (VR 2006)*, pages 3–10.
- Jensen, T. D., Kasprzak, F., Szekely, H.-G., Nikolov, I., Høngaard, J. S., and Madsen, C. (2020). Preliminary study on the influence of visual cues, transitional environments and tactile augmentation on the perception of scale in vr. *HCI International 2020 – Late Breaking Posters*, pages 156–164. Springer International Publishing.
- Loyola, M. (2017). The influence of the availability of visual cues on the accurate perception of spatial dimensions in architectural virtual environments. *Virtual Reality*, 22(3):235–243.
- Lucaci, A.-I., Jakobsen, M., Jensen, P., and Madsen, C. (2022). Influence of texture fidelity on spatial perception in virtual reality. In *Proceedings of Computer Graphics Theory and Applications (GRAPP 2022)*, pages 244–251. SCITEPRESS.
- Ng, A. K. T., Chan, L. K. Y., and Lau, H. Y. K. (2016). *Depth Perception in Virtual Environment: The Effects of Immersive System and Freedom of Movement*, pages 173–183. Virtual, Augmented and Mixed Reality. Springer International Publishing, Cham.
- Peer, A. and Ponto, K. (2017). Evaluating perceived distance measures in room-scale spaces using consumer-grade head mounted displays. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 83–86.
- Renner, R., Velichkovsky, B., and Helmert, J. (2013). The perception of egocentric distances in virtual environments - a review. *ACM computing surveys*, 46(2):1–40.
- Ritter, K. A. and Chambers, T. L. (2021). Three-dimensional modeled environments versus 360 degree panoramas for mobile virtual reality training. *Virtual Reality*, 1.

A-Eye: Driving with the Eyes of AI for Corner Case Generation

Kamil Kowol¹, Stefan Bracke² and Hanno Gottschalk¹

¹*School of Mathematics and Natural Sciences, University of Wuppertal, Gaußstraße 20, Wuppertal, Germany*

²*Chair of Reliability Engineering and Risk Analytics, University of Wuppertal, Gaußstraße 20, Wuppertal, Germany*
{kowol, bracke, hgottsch}@uni-wuppertal.de

Keywords: Corner Case, Human-centered AI, Human-in-the-Loop, Automated Driving, Test Rig.

Abstract: The overall goal of this work is to enrich training data for automated driving with so-called corner cases in a relatively short period of time. In road traffic, corner cases are critical, rare and unusual situations that challenge the perception by AI algorithms. For this purpose, we present the design of a test rig to generate synthetic corner cases using a human-in-the-loop approach. For the test rig, a real-time semantic segmentation network is trained and integrated into the driving simulation software CARLA in such a way that a human can drive on the network's prediction. In addition, a second person gets to see the same scene from the original CARLA output and is supposed to intervene with the help of a second control unit as soon as the semantic driver shows dangerous driving behavior. Interventions potentially indicate poor recognition of a critical scene by the segmentation network and then represent a corner case. In our experiments, we show that targeted and accelerated enrichment of training data with corner cases leads to improvements in pedestrian detection in safety-relevant episodes in road traffic.

1 INTRODUCTION

Despite AI systems achieve impressive performance in solving specific tasks, e.g. in automated driving, they lack understanding of the context of safety in traffic. In contrast, while humans are often described as lousy drivers, as they tend to be diverted or feel fatigue, humans have a fine understanding, when a traffic scene could lead to a situation, where humans are at risk.

It has been observed previously, that to increase robustness and performance of AI algorithms a large number of clean and diverse scenes is needed (Karpathy, 2021). However, a large amount of annotated data per se might not imply safe operation in those rare situations, where road users are exposed to a substantial risk. In this work, we aim to present an accelerated testing strategy that leverages human risk perception to capture corner cases and thereby achieve performance improvement in safety-critical scenes. In order to obtain many safety-critical corner cases in a short time, we stop training at an early stage so that the network is sufficiently well trained. Nevertheless, the scenes generated in this way are still useful to improve fully trained networks.

For this purpose, a semantic network is trained with synthetic images from the open-source driving simulation software CARLA (Dosovitskiy et al.,

2017). In addition, a test rig consisting of 2 control units is connected to CARLA in such a way that the ego vehicle can be controlled with both control units by a human. In this process, the semantic segmentation network is integrated into CARLA in such a way that first the original CARLA image is sent through the network and the prediction is displayed on the screen of one driver. The second driver, in turn, sees the real CARLA image and is supposed to intervene as a safety driver only if he or she feels that a situation is being wrongly assessed by the other driver. We aim to consider situations in which the AI algorithms lead to incorrect evaluations of the scene, which we refer to as safety-relevant corner cases, in order to improve performance through targeted data enrichment. This is done by exchanging images from the original dataset with the safety-critical corner cases, thus keeping the total amount of data fixed. We show that the semantic segmentation network that contains safety-critical corner cases in the training data performs better on similar critical situations than the network that does not contain any safety-critical situations.

Our approach somehow follows the idea of active learning, where we get feedback on the quality of the prediction by interactively querying the scene. However, unlike in standard active learning we do not leave the query strategy to the learning algorithm,

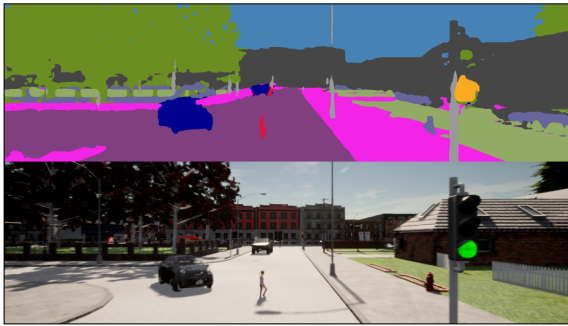


Figure 1: View of the semantic driver (top) and the safety driver (bottom).

but make use of the human’s fine tuned sense of risk to query safety-relevant scenes from a large amount of street scenes, leading to enhanced performance in safety-critical situations.

The contributions of this work can be summarized as follows:

- An experimental setup, that could also be implemented in the real world, permits testing the safety of the AI perception separately from the full system safety including the driving policy of an automated vehicle.
- A proof of concept for the retrieval of training data for automated driving with a human-in-the-loop approach that is safety-relevant.
- A proof that training on safety-relevant situations generated during poor network performance is beneficial for the recognition of street hazards.

Outline. Section 2 discusses related work on corner cases, human-centered AI and human-in-the-loop. In Section 3 we briefly describe the experimental setup used for corner case generation and which data and network were used for our experiments. In Section 4 we explain our strategy to generate corner cases. In Section 5 we demonstrate the beneficial effects of training with corner cases for safety-critical situations in automated driving. Finally, we present our conclusions and give an outlook on future directions of research in Section 6.

2 RELATED WORKS

2.1 Corner Cases

Training data contains few, if any, critical, rare or unusual scenes, so-called corner or edge cases. In the technical fields, the term corner case describes spe-



Figure 2: Test rig including steering wheels, pedals, seats and screens.

cial situations that occur outside the normal operating parameters (Chipengo et al., 2018).

According to (Bolte et al., 2019), a corner case in the field of autonomous driving describes a “non-predictable relevant object/class in relevant location”. Based on this definition, a corner case detection framework was presented to calculate a corner case score based on video sequences. The authors of (Breitenstein et al., 2020) subsequently developed a systematization of corner cases, in which they divide corner cases into different levels and according to the degree of complexity. In addition, examples were given for each corner case level. This was also the basis for a subsequent publication with additional examples (Breitenstein et al., 2021). Since the approach in these references is camera-based, a categorization of corner cases at sensor level was adapted by (Heidecker et al., 2021), where RADAR and LiDAR sensors were also considered. Furthermore, this reference presents a toolchain for data generation and processing for corner case detection.

Outside normal parameters also includes terms such as anomalies, novelties, or outliers, which, according to (Heidecker et al., 2021), correlate strongly with the term corner case. In road traffic, the detection of new and unknown objects, anomalies or obstacles, which must also be evaluated as ‘outside the operating parameters’, is essential. To measure the performance of methods for detecting such objects, the benchmark suite “SegmentMeIfYouCan” was created (Chan et al., 2021a; Chan et al., 2021b). In addition, the authors present two datasets for anomaly and obstacle segmentation to help autonomous vehicles better assess safety-critical situations.

In summary, the term “corner case” can encompass rare and unusual situations that may include anomalies, unknown objects or outliers which are outside of operating parameters. Outside the operating parameters, in the context of machine learning, means that these situations or objects were not part of the training data.

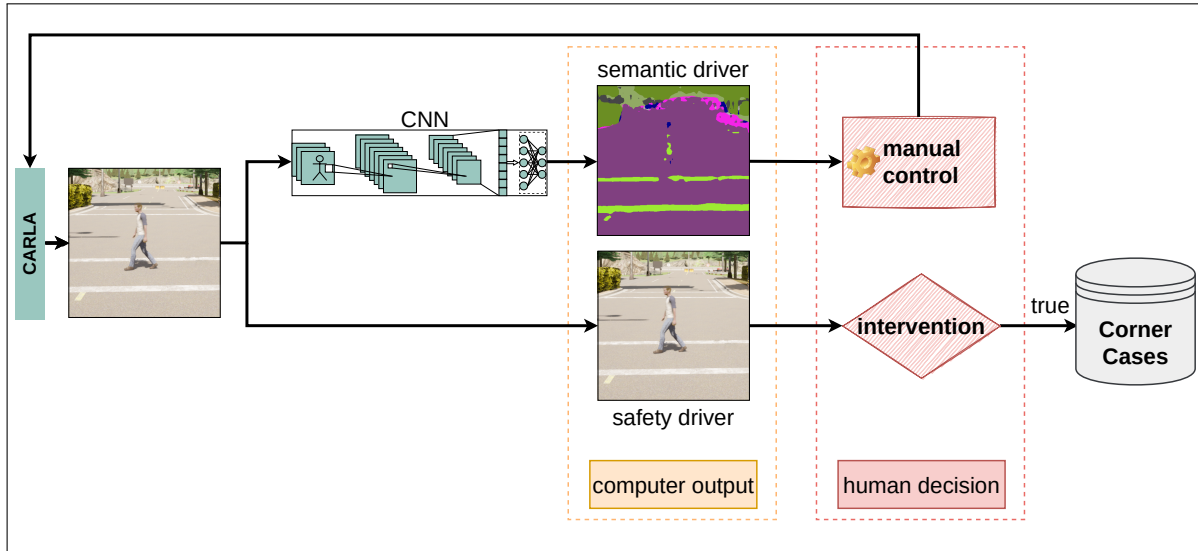


Figure 3: Two human subjects can control the ego vehicle. The semantic driver moves the vehicle in compliance with traffic rules in the virtual world and sees only the output of the semantic segmentation network. The safety driver, who sees only the original image, assumes the role of a driving instructor and intervenes in the situation by braking or changing the steering angle as soon as a hazardous situation occurs. Intervening in the current situation indicates poor situation recognition of the segmentation network and represents a corner case. Triggering a corner case ends the acquisition process and a new run can be started.

2.2 Human-centered AI

According to the Defense Advanced Research Project Agency (DARPA), the development of AI systems is divided into 3 waves (Defense Advanced Research Projects Agency, 2018; Highnam, 2020). While in the first wave patterns were recognized by humans and linked into logical relationships (crafted knowledge), statistical learning could be used in the second wave due to improved computing power and increased memory capacity.

We are currently in the third wave, which relates to the explainability and contextual understanding of AI. Here, the black-box approaches, which emerged in the second wave, are to be understood so that AI decision-making becomes comprehensible to humans. This last point is therefore also an important one in the HCAI initiative, which aims to connect different domains with human-centered AI (Li and Etchemendy, 2018).

On this basis, the authors of (Xu, 2019; Xu et al., 2022) developed an extended HCAI framework that defines the following three different design goals for the human-centered AI: *Ethically Aligned Design* that avoid biases or unfairness of AI algorithms such that these algorithms make decisions according to human criteria and rules. *Technology Design* which considers human and machine intelligence to exploit synergies. *Human Factors Design* to make AI-solutions explainable. The aim is to give humans an insight

into the decision-making process of AI algorithms, so that trust in the current technology can be increased. For this purpose, we have built a specially developed test rig to incorporate human behavior into the further development of AI in the field of autonomous driving. At the same time, we would like to use the test rig to provide a demonstration object to illustrate AI algorithms to society. The focus will be on allowing humans to visually perceive and interact with the decision-making of AI algorithms. In the context of autonomous driving this would mean: driving with the eyes of AI.

2.3 Human-in-the-Loop Approaches

As we are in the third wave of AI systems, there has been increased interest in human-in-the-loop (HITL) and machine learning approaches, where humans interact with machines to combine human and machine intelligence to solve a given problem (Wu et al., 2022). For this purpose simulators were used to improve AI systems by means of human experience or to study human behavior in field trials, which will now be briefly summarized.

The use of simulators and human drivers is applied in reinforcement learning, where an agent learns faster from human experience. For instance, the authors of (Wu et al., 2021) propose a real-time Deep Reinforcement Learning (Hug-DRL) method based on human guidance, where a person can intervene

Table 1: Performance measurement on two test datasets. The comparison shows that the addition of safety-critical scenes in training also improves performance in testing with safety-critical scenes.

traindata			safety-critical testdata		natural distributed testdata	
no.	name	mean _{pixels/scene}	IoU _{pedestrians}	mIoU	IoU _{pedestrians}	mIoU
1	natural distribution	3583.6	0.4600	0.6954	0.4937	0.761
2	pedestrian enriched	6101.1	0.5399	0.6911	0.5586	0.7554
3	corner case enriched	6215.7	0.5683	0.7173	0.5384	0.7517

in driving situations when the agent makes mistakes. These driving errors can be fed directly back into the agents' training procedure and improve the training performance significantly. Furthermore, a method for generating decision corner cases for connected and automated vehicles (CAVs) for testing and evaluation purposes is proposed in (Sun et al., 2021). For this, the behavioral policy of background vehicles (BV) is learned through reinforcement learning and Markov's decision process, which leads to a more aggressive interaction with the CAV which forces more corner cases under test conditions. The tests take place on the highway and include lane changes or rear-end collisions.

Human behaviors can also be analyzed using driving simulators. For example, in (Driggs-Campbell et al., 2015) a realistic test rig including a steering wheel and pedals for data collection was developed. Therefore, thirteen subjects were recruited to drive on different routes while being distracted by static or dynamic objects or by answering messages on their cell phones. By adding nonlinear human behaviors and using realistic driving data, the authors have been able to predict human driving behavior more accurately in testing. Another driving simulator was presented in (Gómez et al., 2018) to develop and evaluate safety and emergency systems. The control units are connected to a generic simulator for academic robotics which uses the Modular Open Robots Simulation Engine *MORSE* (Echeverria et al., 2011). They used an experiment with four road users, one human driver and three vehicles driving in pilot mode and forcing two out of 36 collision situations (*a lead vehicle stopped* and *a vehicle changing lanes*) defined by the National Traffic Safety Administration (NHTSA). The impact of a driver assistance system on the driver was one of the factors studied.

3 EXPERIMENTAL SETUP

3.1 Driving Simulator

Targeted enrichment of training data with safety-critical driving situations is essential to increase the performance of AI algorithms. Since the genera-

tion of corner cases in the real world is not an option for safety reasons, generation remains in the synthetic world, where specific critical driving situations can be simulated and recorded. For this purpose, the autonomous driving simulator CARLA (Dosovitskiy et al., 2017) is used. It is open-source software for data generation and/or testing of AI algorithms. It provides various sensors to describe the scenes such as camera, LiDAR and RADAR and delivers ground truth data. CARLA is based on the Unreal Engine game engine (Epic Games, 1998), which calculates and displays the behavior of various road users while taking physics into account, thus enabling realistic driving. Furthermore, the world of CARLA can be modified and adapted to one's own use case with the help of a Python API.

For our work, we used the API to modify the script for manual control from the CARLA repository. In doing so, we added another sensor, the inference sensor, which evaluates the CARLA RGB images in real-time and outputs the neural network semantic prediction on the screen. An example is shown in Figure 1. By connecting a control unit including a steering wheel, pedals and a screen, to CARLA, we make it possible to control a vehicle with 'the eyes of the AI' in the synthetic world of CARLA. We also connected a second control unit with the same components to the simulator, so that it is possible to control the same vehicle with 2 different control units, see Figure 2. The second control unit is thus operated on the basis of CARLA clear image and can intervene at any time. It always has priority and triggers that the past 3 seconds of driving, which are buffered, are written to the dataset on the hard disk. In order for the semantic driver to follow the traffic rules in CARLA, the script had to be modified additionally. The code has been modified to display the current traffic light phase in the upper right corner and the speed in the upper center.

3.2 Test Rig

The test rig consists of the following components: a workstation with CPU, 3x GPU Quadro RTX 8000, 2 driving seats, 2 control units (steering wheel with pedals), one monitor for each control unit and two monitors for the control center. The driving simula-



Figure 4: Two examples of a corner case where the safety driver had to intervene to avoid a collision due to the poor prediction of the semantic segmentation network.

tor software used is the open-source software CARLA version 0.9.10.

3.3 Dataset for Initial Training and Testing

For training, a custom dataset was generated using CARLA 0.9.10, consisting of 85 scenes with 60 frames each. In addition, there is a validation dataset with 20 scenes. The dataset was generated on seven maps with one fps and contains the corresponding semantic segmentation image in addition to the rendered synthetic image. The maps include the five standard maps in CARLA and two additional maps that offer a mix of city, highway and rural driving. Various parameters can be set in CARLAmi we focused on the number of Non-Player-Characters (NPCs), including cars, motorcycles, bicycles and pedestrians, and on environment parameters such as sun position, wind and clouds. Depending on the size of the map, the number of NPCs ranged from 50 to 150.

The clouds and wind parameters can be set in the range between 0 and 100, with 100 being the highest value. The wind parameter is responsible for the movement of tree limbs and passing clouds and was in the range of 0 and 50. The cloud parameter describes the cloudiness, where 0 means that there are no clouds at all and 100 that the sky is completely covered with clouds. We have chosen values between 0 and 30. The altitude describes the angle of the sun in relation to the horizon of the CARLA world, with values between -90 (midnight) and 90 (midday). Values between 20 and 90 were used for our purpose. The other environmental parameters like rain, wetness, puddles or fog are set to zero. The parameters are chosen so that the scenes reflect everyday situations with a natural scattering of NPCs and in similar good weather. During data generation, the movement of all NPCs was controlled by CARLA.

Furthermore, 21 corner case scenes were used as test data, each containing 30 frames. Another test dataset containing 21 standard scenes without corner cases serves as a comparison, each containing 30 frames.

3.4 Training

To drive on the predicted semantic mask, a real-time capable network architecture is needed. For these purposes, the Fast Segmentation Convolutional Neural Network (Fast-SCNN) model was used (Poudel et al., 2019). It uses two branches to combine spatial details at high resolution and deep feature extraction at lower resolution achieving a mean Intersection over Union (mIoU) of 0.68 at 123.5 fps on the Cityscapes dataset (Cordts et al., 2016). The network was implemented in the python package PyTorch (Paszke et al., 2019) and training was done on a NVIDIA Quadro RTX 8000 graphics card. Sixteen of the 23 classes available in CARLA were used for training. Cross entropy was used as the loss function and ADAM as the optimization algorithm. A polynomial decay scheduler was also used to gradually reduce the learning rate.

We intentionally stopped the training after 5 epochs to increase the frequency of perception errors for the network. The resulting network is sufficiently well trained to recognize the road and all road users, although objects further away are poorly recognized. An example is shown at the top of Figure 1.

3.5 Experimental Design

Two operators conducted the driving campaign and the duration of driving as a safety driver or semantic driver was set at 50:50. Both participants had time to familiarize themselves with the hardware and the CARLA world before the start of the first driving campaign so that driving errors could be minimized. Two driving campaigns are planned; the first campaign will generate targeted corner cases and the second will test whether adding the corner cases included in campaign 1 leads to an improvement in the perception of safety-critical situations.

4 RETRIEVAL OF CORNER CASES

For the generation of corner cases, we consider the following experimental setup. Two test operators

Table 2: Corner case appearances on Fast-SCNN trained with 3 different datasets.

dataset		driven distance d	driven time t	number CCes	mean _{d_{CC}}	std _{d_{CC}}	mean _{t_{CC}}	std _{t_{CC}}
no.		[km]	[min]	[-]	[km/CC]	[km/CC]	[min/CC]	[min/CC]
1	natural distribution	121.32	411	13	7.73	14.25	25.93	39.60
2	pedestrian enriched	163.09	500	21	7.52	10.47	23.25	28.72
3	corner case enriched	153.38	528	11	13.84	8.68	47.47	31.87

record scenes in our specially constructed test rig (see Figure 2), where one subject (safety driver) gets to see the original virtual image and the other (semantic driver) the output of the semantic segmentation network (see Figure 1). The test rig is equipped with controls such as steering wheels, pedals and car seats and connected to CARLA to simulate realistic traffic participation.

The corner cases were generated as shown in Figure 3. For this purpose, we used a real-time semantic segmentation network from Section 3.4 where visual perception was limited. We note that autonomous vehicles according to (Favarò et al., 2017) 67619.81 km drive until an accident happens. Using a poorly trained network as a part of our accelerated testing strategy, we were able to generate corner cases after 3.34 km in average between interventions of the safety driver. We note however that the efficiency of the corner cases was evaluated using a fully trained network. Figure 4 shows two safety-critical corner cases where the safety driver had to intervene to prevent a collision.

If the safety driver triggered the recording of a corner case the test operators label the corner case with one of four options available (overlooking a walker or a vehicle, disregarding traffic rules, intervening out of boredom) and may leave a comment. Furthermore, the kilometers driven and the duration of the ride are notated. The operators were told to obey the traffic rules and not to drive faster than 50 km/h during the test drives. After a certain familiarization period, driving errors decreased and sudden braking by the semantic driver was also reduced. The reason for this is that the network partially represents areas as vehicles or pedestrians with fewer pixels. Over time, a learning effect occurred in the drivers to hide such situations because experience showed that there was no object there due to the previous frames.

The rides are tracked and by the intervention of the safety driver the last 3 seconds of the scene are saved. Subsequently, the scenes can be loaded and images saved from the ego vehicle’s perspective using the camera and the semantic segmentation sensor. We collect 50 corner cases before retraining from scratch with a mixture of original and corner case images. For each corner case, the last 3 seconds are saved at 10 fps before the intervention by the safety driver. In total, we get 1500 new frames. When using this corner

case data for retraining, we delete the same number of frames from the original training dataset.

We selected 50 corner cases in connection with pedestrians. Therefore, the inclusion of corner case scenes into the training dataset significantly increases the average number of pixels with the pedestrian class in the training data. To establish a fair comparison of the efficiency of corner cases as compared to a simple upsampling of the pedestrian class, we created a third dataset that contains approximately the same number of pixels per scene for the pedestrian class as the dataset with the corner cases, see Table 1.

5 EVALUATION AND RESULTS

All results in this section are averaged over 5 experiments to obtain a better statistical validity. For testing purposes, we generated 21 additional corner cases for validation. With the same setup as before, we train the Fast-SCNN for 200 epochs on all three datasets and thereby obtain three networks. Table 1 shows the evaluation of all three models on the class pedestrian for the 21 safety-critical test corner cases. We see that adding corner cases to the training data leads to an improvement in pedestrian detection in safety-critical situations, which can also be shown by an example in Figure 5. There we see a situation with a pedestrian crossing the road, with a slope directly behind him that seems to end the road at the level of the horizon. Therefore, the networks that did not have corner cases in the training data seem to have problems with this situation, while the model with corner cases detects the humans much better.

While training the network using naive upsampling of pedestrians does not have any positive effect on the classe’s Intersection over Union (IoU) as compared with the original training data, we achieve a gain in the IoU by 2.19% when using the dataset containing corner cases. In addition, the 3 models were tested on a dataset with a natural distribution of pedestrians. Here it can be seen that the model trained with corner cases does not perform as well as the model with the same number of pedestrians. It follows that the model performs better in critical situations, while the models without corner cases perform less well.

Since our method for generating corner cases in safety-critical situations provides an improvement in

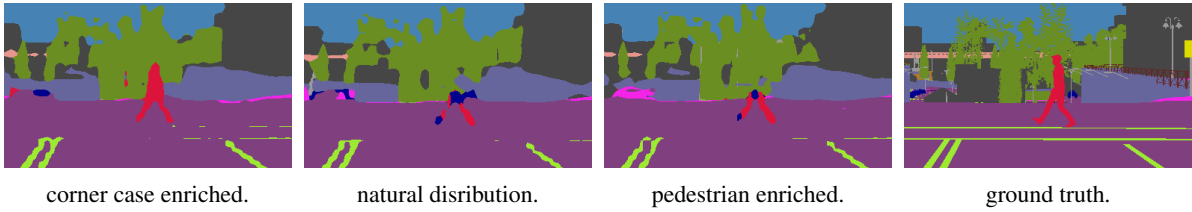


Figure 5: Evaluation on corner case test data shows that the model using corner case data in training recognizes pedestrians better than the model trained with the natural distributed dataset or the dataset which contains more pedestrians.

detecting pedestrians in safety-critical situations, we launched a second campaign to verify how long it takes driving with the 3 trained networks to generate a corner case. When conducting the second driving campaign, the same operating parameters were set as in the first campaign and the duration until a corner case occurred was recorded. This includes the same maps and weather conditions as well as the same two drivers. However, the two drivers were not aware of the network’s underlying training data. Table 2 shows the duration and kilometers driven for the different data sets, as well as the occurrence of corner cases during these rides. We see that adding corner cases during training also reduces the frequency until new corner cases reappear.

We therefore demonstrated the benefits of our method to generate corner cases, especially for safety-critical situations. We were also able to show that adding safety-critical corner cases recorded by intentional perceptual distortions improves performance, so future datasets should include such situations.

6 CONCLUSION

Due to the lack of explanation and transparency in the decision-making of today’s AI algorithms, we developed an experimental setup that allows to visualize these decisions and thus to allow a human driver to evaluate the driving situations while driving with the eyes of AI, and from this to extract data that includes safety-critical driving situations. Our self-developed test rig provides two human drivers controlling the ego vehicle in the virtual world of CARLA. The semantic driver receives the output of a semantic segmentation network in real-time, based on which she or he is supposed to navigate in the virtual world. The second driver takes the role of the driving instructor and intervenes in dangerous driving situations caused by misjudgments of the AI. We consider driver interventions by the safety driver as safety-critical corner cases which subsequently replaced part of the initial training data. We were able to show that targeted data enrichment with corner cases created with lim-

ited perception leads to improved pedestrian detection in critical situations.

In addition, we take up the idea of the HAI framework and continue the further development of AI by means of human risk perception to identify situations that are particularly important to humans, and thus train the AI precisely where it is particularly challenged by a human perspective.

Future research projects include the use of networks of different quality, changing the weather parameters and provoking accident scenarios so that the number of corner cases can be artificially increased in test operation. In addition, multi-screen driving should be expanded to increase the field of view (FOV) for a more realistic driving experience. The intervention of the safety driver in the driving situation will also be observed. To this end, criteria for measuring human-machine interaction (HMI) will be developed to track, for example, latency, attention, and intervention due to boredom of the drivers. In addition, contextual and personal factors of human drivers will be investigated to assess uncertainty or anxiety while driving.

ACKNOWLEDGEMENTS

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project KI Data Tooling under the grant number 19A20001O. We thank Matthias Rottmann for his productive support, Natalie Grabowsky and Ben Hamscher for driving the streets of CARLA and Meike Osinski for the support in the field of human-centered AI.

REFERENCES

- Bolte, J.-A. et al. (2019). Towards Corner Case Detection for Autonomous Driving. In *2019 IEEE Intelligent Vehicles Symposium, IV 2019, Paris, France, June 9-12, 2019*, pages 438–445. IEEE.
- Breitenstein, J. et al. (2020). Systematization of corner cases for visual perception in automated driving. In

- IEEE Intelligent Vehicles Symposium, IV 2020, Las Vegas, NV, USA, October 19 - November 13, 2020*, pages 1257–1264. IEEE.
- Breitenstein, J. et al. (2021). Corner cases for visual perception in automated driving: Some guidance on detection approaches. *CoRR*, abs/2102.05897.
- Chan, R. et al. (2021a). SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
- Chan, R., Rottmann, M., and Gottschalk, H. (2021b). Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5128–5137.
- Chipengo, U., Krenz, P., and Carpenter, S. (2018). From antenna design to high fidelity, full physics automotive radar sensor corner case simulation. *Modelling and Simulation in Engineering*, 2018:1–19.
- Cordts, M. et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Defense Advanced Research Projects Agency (2018). AI Next Campaign. Available: <https://www.darpa.mil/work-with-us/ai-next-campaign>. Accessed 07-Jun,-2022.
- Dosovitskiy, A. et al. (2017). CARLA: an open urban driving simulator. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR.
- Driggs-Campbell, K., Shia, V., and Bajcsy, R. (2015). Improved driver modeling for human-in-the-loop vehicular control. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1654–1661.
- Echeverria, G. et al. (2011). Modular open robots simulation engine: Morse. In *2011 IEEE International Conference on Robotics and Automation*, pages 46–51.
- Epic Games (1998). Unreal Engine. Available: <https://www.unrealengine.com>.
- Favarò, F. M. et al. (2017). Examining accident reports involving autonomous vehicles in california. *PLOS ONE*, 12:1–20.
- Gómez, A. E. et al. (2018). Driving simulator platform for development and evaluation of safety and emergency systems. *Cornell University Library*.
- Heidecker, F. et al. (2021). An Application-Driven Conceptualization of Corner Cases for Perception in Highly Automated Driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, Nagoya, Japan.
- Highnam, P. T. (2020). The defense advanced research projects agency’s artificial intelligence vision. *AI Magazine*, 41:83–85.
- Karpathy, A. (2021). Cvpr 2021 workshop on autonomous driving, keynote. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*.
- Li, F.-F. and Etchemendy, J. (2018). Introducing Stanford’s Human-Centered AI Initiative. Available: <https://hai.stanford.edu/news/introducing-stanford-human-centered-ai-initiative>. Accessed 03-Jun,-2022.
- Paszke, A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Poudel, R. P. K. et al. (2019). Fast-scnn: Fast semantic segmentation network. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 289. BMVA Press.
- Sun, H. et al. (2021). Corner case generation and analysis for safety assessment of autonomous vehicles. *Transportation Research Record*, 2675(11):587–600.
- Wu, J. et al. (2021). Human-in-the-loop deep reinforcement learning with application to autonomous driving. *CoRR*, abs/2104.07246.
- Wu, X. et al. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.
- Xu, W. (2019). User-centered design (iv): Human-centered artificial intelligence. *Journal of Applied Psychology*, 25:291–305.
- Xu, W. et al. (2022). Transitioning to human interaction with ai systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human-Computer Interaction*, pages 1–25.

An Integrated Neural Network and Structural Equation Modeling Approach for Modeling Activity Trackers Use

Ricardo Sol^a and Karolina Baras^b

Exact Sciences and Engineering, University of Madeira, Madeira, Portugal
ricardosolj@gmail.com, karolina.baras@staff.uma.pt

Keywords: Artificial Neural Networks (ANN), Ubiquitous Systems, Personal Informatics, Personal Data Tracking, Sports/Exercise, Technology Acceptance Model (TAM), Embodied Interaction.

Abstract: The objective of this study is to enhance a Technology Acceptance Model (TAM) with an Artificial Neural Network (ANN) approach in order to obtain substantially accurate results when compared to Structural Equation Modeling (SEM). This study looked at another paper that created a TAM dedicated to activity trackers (AT) obtained via SEM from a questionnaire to 247 participants. This study uses the constructs of that paper in an ANN as the input units and the Root Mean Square of Errors to indicate that the ANN method achieves high prediction accuracy. The results provide conclusive evidence that Perceived Usefulness is the most significant factor affecting AT acceptance. Perceived Ease of Use and Image affect acceptance, however their impact is much lower. Hedonic Motivation and Habit were found to have a significant relationship with TAM while Self-Efficacy showed mixed results. This confirmation can be useful for future designs of activity trackers.

1 INTRODUCTION

Wearable devices such as activity trackers have become important in monitoring health behavior, for recreation, and socialization, and thus are a viable and significant research topic in Human Computer Interaction. Confirming this trend is the International Data Corporation in a press release, stating that worldwide shipments of wearables grew 9.9% throughout the third quarter of 2021 reaching 138.4 million units (IDC 2021). The improvement and the commercialization of activity trackers have helped many users to reach the recommended goal of ten thousand steps per day in order to maintain or improve their health (Akers 2012). However, a study on the acceptance of a particular activity tracker device discovered that half of the users stop using the device after two weeks (Shih 2015).

One possibility, to help solve design issues that lead to loss of interest or decrease of device usage, is the use of models. Even though some researchers think of them as excessively theoretical. In fact, researchers working with interfaces who had often been skeptical, started to acknowledge that models

could be helpful in the design of interfaces (Myers 2000). Since Li et al.'s seminal work, researchers have been trying to describe the use of trackers through a model (Li 2010). Li et al. presented a model with five iterative stages: preparation, collection, integration, reflection, and action; later the model was refined by these authors. Also, Epstein et al. looked at that model and expanded on it by including the lapses and interruptions of tracking, and highlighting the intricacy of integration, collection and reflection (Epstein 2015). Narrowing the scrutiny, Sol and Baras obtained a model dedicated to activity trackers use (Sol 2016) that is used in this paper. The most important advantage of this model is that it is quantitative. It was obtained by expanding the Technology Acceptance Model (TAM), with health oriented, data control, and other constructs. TAM assumes that user acceptance can be described by two ideas: Perceived Usefulness (PU), and Perceived Ease of Use (PEoU) which determine Intention to Use through Attitude (Davis 1989).

A TAM based model for activity trackers, as most of the research on technology acceptance models is done simply with Structural Equation Modeling

^a  <https://orcid.org/0000-0003-4333-7140>

^b  <https://orcid.org/0000-0002-2050-6565>

(SEM) methods, or other models, for example, the Ubiquitous Computing Acceptance Model (Spiekermann 2007), and the Health Information Technology Acceptance Model (Kim 2013). SEM is a sophisticated multivariate technique that can be used to scrutinize multiple dependence associations between variables simultaneously. It is useful for hypothesis specification and testing, can suggest novel hypotheses that were not considered initially. Nevertheless, SEM may lead frequently to an oversimplification of the complexities involved as it is simply detecting linear relationships (Ringle 2012). To address this issue, undertaking a second step using an Artificial Neural Network (ANN) allows for further scrutiny and examination.

An ANN is “a massively parallel distributed processor made up of simple processing units, which have a neural propensity for storing experimental knowledge and making it available for use” (Haykin 2004). Contrary to SEM an ANN is not suitable for hypotheses testing, but further to linear relationships can also deal with non-linear relationships. Moreover, an ANN has the capability to assess non-compensatory processes (Svozil 1997). Additionally, an ANN is more robust and can offer greater prediction accuracy than linear models (Tan 2014). This study uses the constructs of the TAM based model for activity trackers as the input units of an ANN in order to obtain a more accurate view of the acceptance and use of these devices.

In the following sections, we contextualize the study with the literature review, and present the model. Next, we describe the methodology, and discuss the results of ANN analysis. Finally, we conclude, and envision the possibility of future research.

2 LITERATURE REVIEW

In this section we primarily review the literature related to activity trackers. We also look into Artificial Neural Networks research to understand its relations, importance and classifications. The acceptance and use of activity trackers is due to many reasons and motives, several of which seem to clash. Users may begin tracking their activity since they have a certain goal in mind (Epstein 2015). Still, there are users who start to use activity trackers with no goal in mind and use the device to help them set an objective. This objective becomes clearer as the usage changes after transitioning from the discovery phase to the maintenance phase of pondering (Li 2010). Other users start tracking merely moved by concern and curiosity of the quantitative data (Lindqvist

2011). Nonetheless, goal setting is just one notion to help and persuade health-related behavior change.

For example, when the user wants to implement a habit in daily life, one of the best ways is for the activity tracker to help implementing routines (Lazar 2015).

An egocentric perspective for these devices (Elsden 2015) can be looked at as a form of hedonic motivation, as is individual encouragement (Patel 2015), the acknowledgement of effort (Kim 2016), and giving credit (Consolvo 2006).

When looking at image one can look into the lifestyle of the user (Consolvo 2006) or to the aesthetics and form of the devices (Harrison 2015). Other notions embrace social comparison (Harrison 2015), social competition and collaboration (Patel 2015).

Users manage to change their goals, habits, and devices; however, the applications or dashboards are ill equipped to allocate this change. For tracking, users tend to change devices often or even use several devices in parallel, which leads to complications in measuring and associating data (Rooksby 2014). This issue has many impacts as it increases the difficulty to provide a tailored efficacy evaluation (Klasnja 2011) that is important for the users’ self-efficacy.

When one approaches to data control the information that activity trackers are gathering can be extremely sensitive (Lupton 2017) and the risk of third-party recording is real (Elsden 2015).

When looking at how to improve the usefulness of activity trackers different researchers produced several design ideas, one of which was the facilitating of micro-plans (Gouveia 2018), another was to give meaningfulness in context (Rooksby 2014), yet another was to provide a wide variety of adjustable goals (Clawson 2015), or to appeal to identity (Lazar 2015), and another was the idea of adherence (Tang 2018).

The idea that the devices have to “speak” the language of the users (Lazar 2015) because users are not data scientists (Rooksby 2014), and the need for devices to remind them (Shih 2015) are ways to improve the usability of activity trackers. Nevertheless, these devices are still being used in a rather limited manner (Didziokaite 2017).

Numerous statistical techniques are parametric, such as SEM and Multiple Regression Analysis (MRA), requiring a great statistic background, while artificial neural networks are non-parametric models, which can provide higher prediction accuracy (Tan 2014). An ANN uses a considerable interconnection of simple computing units called neurons or nodes as inputs, hidden, and outputs layers with connection values called synaptic weights that are adaptable via

an iterative process. A classic ANN consists of several layers: one input layer, one or more hidden layers and one output layer (Negnevitsky 2011). In ANNs for technology acceptance typically only one hidden layer is used (Tan 2014). There are several types of ANNs, but the most common is feed forward back propagation multilayer perceptron (BPFF). In this kind of network, belonging to supervised learning ANNs, the knowledge stored in the network by iteratively subjecting it to patterns of known inputs and outputs (Negnevitsky 2011). The difference between desired and actual output, is calculated and propagated back, in order to change the synaptic weights and by doing so minimize the estimation error (Haykin 2004).

A node uses a function f defined as a weighted sum of its inputs based on equation 1 where the w are the weights, and the x are the inputs, the bias is b and the output is Z (Haykin 2004).

$$Z = f(w_1x_1 + w_2x_2 + b) \quad (1)$$

There are many activation functions for the output layer, however Sigmoid, shown in equation 2, is generally used in a technology acceptance context (Tan 2014).

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

The root mean square of errors (RMSE) is used to predict accuracy and is calculated using equation 3 and 4 (Tan 2014), where SSE is the sum of squared error, and MSE is the mean squared prediction error.

$$MSE = \frac{SSE}{n} \quad (3)$$

$$RMSE = \sqrt{MSE} \quad (4)$$

3 MODEL OBTAINED VIA SEM

In order to obtain the model for activity tracking use, via Structural Equation Modeling, the paper targets a population of actual activity trackers users. These users were recruited from social media and on a micro work site. A survey with 80 questions adapted from prior research was deployed. The items of the survey were considered using a seven-point Likert scale, amid between “Strongly Disagree” and “Strongly Agree.” Specifically, the strength and significance of direct effect of nineteen independent variables on behavioral intention were determined. From a total of 17 tested relationships, 11 were statistically significant.

There were a total of 247 users, mostly from Western Europe and North America, that completed

Table 1: Definitions of the constructs present in the model.

	Definition
Perceived Susceptibility of Disease	“The perception of the likelihood of experiencing a condition that would adversely affect one's health” (Jayanti 1998)
Perceived Severity of Disease	“The beliefs a person holds concerning the effects a given disease or condition would have on one's state of affairs” (Hochbaum 1952)
Habit	“The extent to which people tend to perform behaviors automatically because of learning” (Limayem, 2007)
Health Consciousness	“The degree to which health concerns are integrated into a person's daily activities” (Jayanti, 1998)
Hedonic Motivation	“The fun or pleasure derived from using a technology, and it has been shown to play an important role in determining technology acceptance and use” (Brown, 2005)
Image	“The degree to which use of an innovation is perceived to enhance one's image or status in one's social system” (Moore, 1991)
Self-Efficacy	“The judgment of one's ability to use a technology (e.g., computer) to accomplish a particular job or task” (Compeau 1995)
Perceived Data Control	“The degree to which a person feels they have control over the use of, and access to, the data collected” (Lindqvist 2011)
Perceived Usefulness	“The degree to which a person believes that using a particular system would enhance his or her job performance” (Davis 1989)
Perceived Ease of Use	“The degree to which a person believes that using a particular system would be free of effort” (Davis 1989)

the survey, from these 144 were male (58.3 percent) and 103 were female (41.7 percent). The average age was 33 with a standard deviation of 10.6.

The sample size of 247 has exceeded the recommended minimum sample size of 111 obtained from G*Power with an effect size of 0.3, an alpha level of 0.05 and a power of 0.95 (Faul 2009). In Table 1 we display the definitions of the constructs that make up the model.

The previously found model was obtained using maximum likelihood parameter estimation. Descriptive statistics, and Exploratory Factor Analysis were conducted using IBM SPSS version 23. The structural equation model was built-in with maximum likelihood estimation routines in IBM SPSS Amos 23.

The Kurtosis analysis found normality issues, with values higher than 2, in item 2 of the construct Perceived Usefulness, item 2 of the construct Perceived Ease of Use, item 1 of the construct Intention to Use, and in all items of the construct Behavioral Intention. However, these constructs passed in the Exploratory Factor Analysis. In Figure 1 we show the path diagram of the activity tracker acceptance model with the respective path coefficients.

4 METHODOLOGY

In this work, an Artificial Neural Network is applied to analyze, complement and verify the SEM approach and measure the effectiveness of the constructs that prevailed for the acceptance of activity trackers.

We used a Multilayer Perceptron (MLP) back propagation feed-forward (BPFF) method. The MLP is the most used and widespread ANN method (Liébana-Cabanillas 2017). The ANN contains three layers: the input layer, the hidden layer, and the output layer. In this work, the ANN is created using SPSS 24. The model obtained from SEM is divided into four ANN models with one output variable. Model A has the output as the construct Health Consciousness and has three inputs Perceived Susceptibility to Disease, Perceived Severity to Disease, and Habit. Model B has the output as the construct Perceived Usefulness and has five inputs Health Consciousness, Hedonic Motivation, Image, Self-Efficacy, Perceived Ease of Use. Model C has the output as the construct Perceived Ease of Use and has three inputs Image, Self-Efficacy, and Perceived Data Control. Model D has the output as the construct Behavioral Intention to Use (BIU)/ Acceptance and has three inputs Perceived Usefulness, Image, Perceived Ease of Use. The four models ANN are shown in Figure 2. The nodes (hidden neurons) are

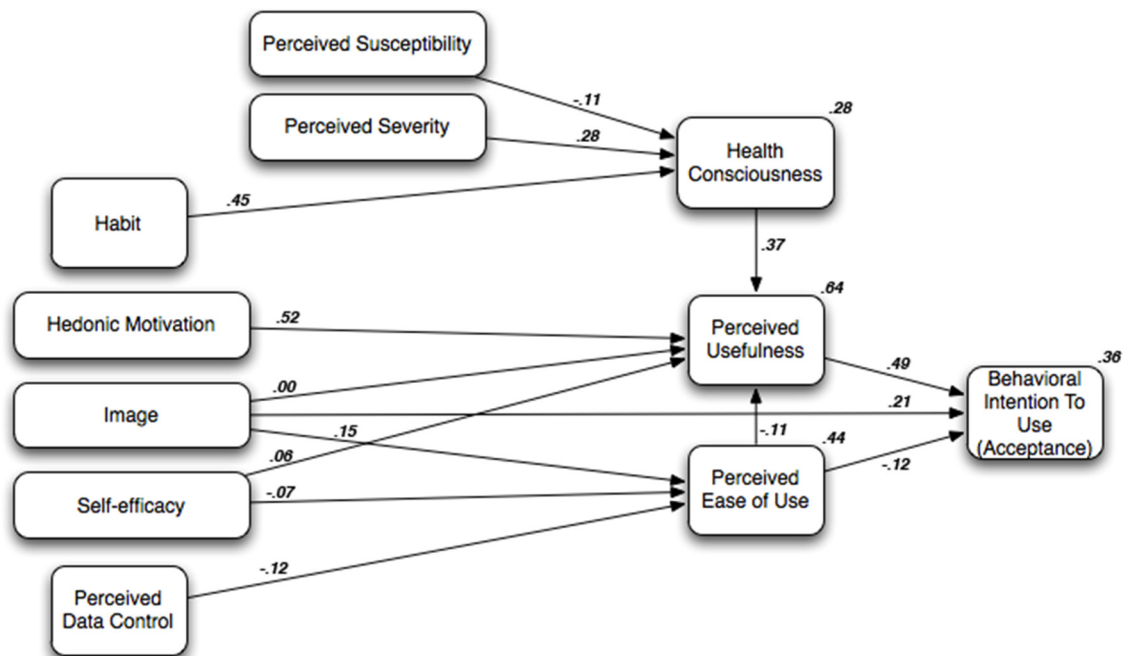


Figure 1: Path Diagram of the Activity Trackers Acceptance Model obtained via SEM with respective path coefficients (Sol 2016).

automatically generated by SPSS and the activation function used for both hidden and output layers was Sigmoid Function. We assigned 90 % of the samples to the training procedure and the remaining 10% were used for the testing procedure. To avoid the risk of over-fitting, we employed a ten-fold cross-validating process. The root mean square of errors (RMSE) was used to assure the predictive accuracy of the four ANNs. The next section analyzes the results of the ANN.

5 ARTIFICIAL NEURAL NETWORK RESULTS

An ANN is helpful in discovering both linear and non-linear relationships without requiring any distribution assumptions like linearity, normality, or homoscedasticity as in Structural Equation Modeling (Leong 2013). By doing so, an ANN can provide higher prediction accuracy (Tan 2014).

Table 2: RMSE values of ten artificial neural networks.

	Model A		Model B		Model C		Model D	
Input Neuron	Perceived Susceptibility to Disease (PSusD), Perceived Severity to Disease (PSevD), Habit		HC, Hedonic Motivation (HM), Image, Self-Efficacy, Perceived Ease of Use		Image (I), Self-Efficacy (SE), Perceived Data Control (PDC)		Perceived Usefulness, Image, Perceived Ease of Use	
Output Neuron	Health Consciousness (HC)		Perceived Usefulness (PU)		Perceived Ease of Use (PEoU)		Behavioral Intention to Use (BIU)/ Acceptance	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
ANN 1	0.114	0.134	0.072	0.073	0.104	0.096	0.075	0.057
ANN 2	0.117	0.107	0.077	0.077	0.101	0.099	0.072	0.058
ANN 3	0.118	0.116	0.083	0.055	0.100	0.120	0.072	0.067
ANN 4	0.119	0.106	0.076	0.056	0.104	0.094	0.074	0.076
ANN 5	0.119	0.104	0.078	0.077	0.105	0.102	0.087	0.054
ANN 6	0.115	0.129	0.087	0.072	0.102	0.104	0.073	0.071
ANN 7	0.117	0.080	0.080	0.063	0.105	0.159	0.071	0.062
ANN 8	0.116	0.107	0.079	0.049	0.111	0.130	0.079	0.053
ANN 9	0.130	0.108	0.073	0.066	0.100	0.114	0.069	0.083
ANN 10	0.117	0.072	0.069	0.067	0.103	0.079	0.078	0.057
Mean RMSE	0.118	0.106	0.077	0.066	0.103	0.110	0.075	0.064
Standard Deviation	0.005	0.019	0.005	0.010	0.003	0.022	0.005	0.010

Table 3: Neural network sensitivity analysis.

	Model A			Model B					Model C			Model D		
Output	Health Consciousness (HC)			Perceived Usefulness (PU)					Perceived Ease of Use (PEoU)			Behavioral Intention to Use (BIU)/ Acceptance		
	Relative Importance			Relative Importance					Relative Importance			Relative Importance		
ANN	PSusD	PSevD	H	HC	HM	I	SE	PEoU	I	SE	PDC	PU	I	PEoU
1	0.324	0.240	0.437	0.040	0.671	0.065	0.113	0.112	0.301	0.429	0.270	0.730	0.218	0.052
2	0.326	0.256	0.417	0.070	0.603	0.050	0.123	0.154	0.352	0.356	0.291	0.731	0.166	0.103
3	0.298	0.238	0.464	0.237	0.514	0.025	0.097	0.127	0.358	0.306	0.336	0.690	0.174	0.136
4	0.350	0.161	0.488	0.062	0.603	0.035	0.071	0.229	0.411	0.331	0.258	0.651	0.213	0.136
5	0.298	0.248	0.454	0.091	0.584	0.023	0.052	0.250	0.311	0.423	0.266	0.360	0.188	0.453
6	0.379	0.251	0.370	0.168	0.367	0.033	0.170	0.264	0.324	0.419	0.257	0.694	0.149	0.157
7	0.315	0.217	0.467	0.060	0.567	0.096	0.117	0.161	0.335	0.376	0.289	0.754	0.193	0.053
8	0.311	0.252	0.437	0.039	0.601	0.041	0.029	0.289	0.237	0.503	0.260	0.579	0.174	0.247
9	0.223	0.106	0.671	0.023	0.651	0.018	0.102	0.205	0.304	0.299	0.397	0.723	0.195	0.082
10	0.255	0.275	0.470	0.049	0.604	0.054	0.146	0.146	0.380	0.353	0.267	0.525	0.405	0.070
Average	0.308	0.224	0.468	0.084	0.577	0.044	0.102	0.194	0.331	0.380	0.289	0.644	0.208	0.149
Average	68%	50%	99.8%	16.4%	100%	7.7%	18.9%	35.2%	83.3%	93.4%	72.5%	97.9%	33.3%	25.9%

As shown in Table 2, the RMSE values for the training data and the testing data are low, representing a higher predictive accuracy and better data fit.

In Table 3, we show the results of the sensitivity analysis that assessed the strength of the predictive power of each of the input neurons. In order to have the normalized importance of these neurons in percentage we divided the relative importance by the maximum importance.

Habit (H) was found to be the key determinant in predicting Health Consciousness (HC) followed by Perceived Susceptibility to Disease (PSusD) and lastly Perceived Severity to Disease (PSevD) in model A. In model B, the order of importance towards Perceived Usefulness (PU) in descending order is Hedonic Motivation (HM), followed by Perceived Ease of Use (PEoU) and Self-Efficacy (SE) and the least important were Health Consciousness (HC) and Image (I). For model C, Self-Efficacy (SE) is the most prominent predictor for Perceived Ease of Use

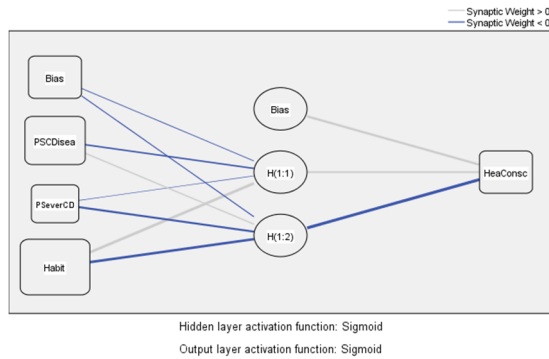


Figure 2: Neural Network between Perceived Susceptibility to Disease, Perceived Severity to Disease, and Habit with Health Consciousness.

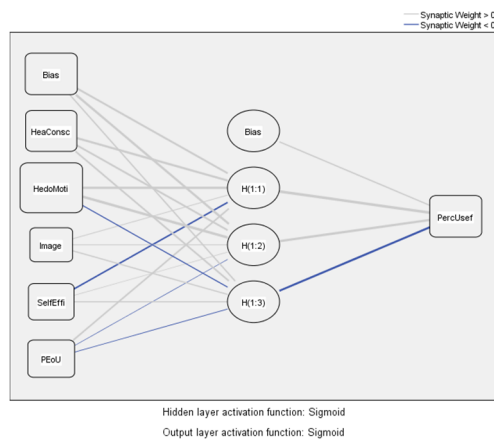


Figure 3: Artificial Neural Network between Health Consciousness, Hedonic Motivation, Image, Self-Efficacy, and Perceived Ease of Use with Perceived Usefulness.

(PEoU), followed by Image (I) and lastly Perceived Data Control (PDC). Finally, Perceived Usefulness (PU) constituted the most effective in term of predicting Behavioral Intention to Use (BIU)/ Acceptance, followed by Image (I) and lastly Perceived Ease of Use (PEoU).

All constructs in all ten ANNs for each model had at least one non-zero synaptic weight connected to the hidden neurons which validates the relevance of the constructs as variables as Figures 2 to 5 show.

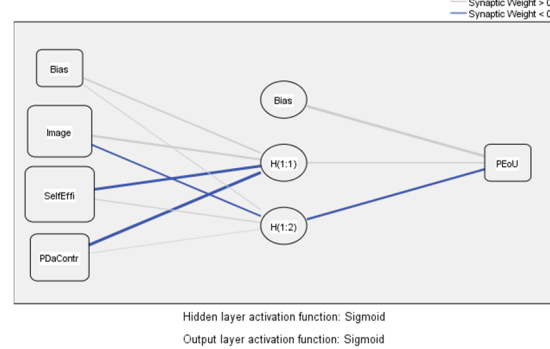


Figure 4: Artificial Neural Network between Image, Self-Efficacy, and Perceived Data Control with Perceived Ease of Use.

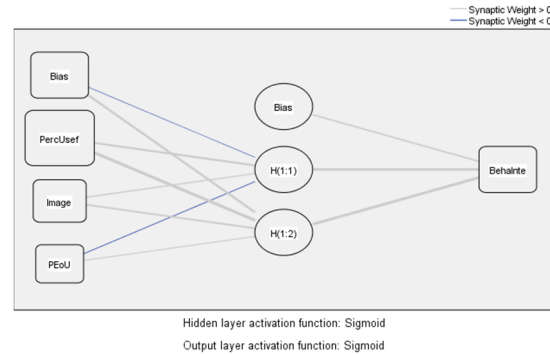


Figure 5: Artificial Neural Network between Perceived Usefulness, Image, and Perceived Ease of Use with Behavioral Intention to Use / Acceptance.

6 DISCUSSION

The Technology Acceptance Model (TAM) is based on many theories and grounded in many studies. In this work, the determinants of activity trackers use include TAM constructs and other constructs such as Image, Hedonic Motivation, Habit and Self-Efficacy. The results show that the research model studied in this work is acceptable. Next, we discuss the findings in more detail.

6.1 Relationships between Perceived Susceptibility to Disease, Perceived Severity to Disease, and Habit with Health Consciousness

As shown in Table 4, the construct Habit showed a significant relationship with Health Consciousness with a path coefficient of 0.451 obtained during the Structural Equation Modeling and has the highest normalized importance according to the Model A of the Artificial Neural Network analysis. While to our knowledge this relation is not found in the literature, the fact that both approaches ranked Habit in first place makes one ponder that a Health Conscious person has health habits.

The findings in this study also show that the construct Perceived Severity to Disease with 50% normalized importance is positively related to Health Consciousness. Looking at this result one might consider that if a person is Health Conscious, then that person should have a high degree of awareness of disease and related issues. Nevertheless, the construct Perceived Susceptibility of Disease showed mixed results within the two approaches.

6.2 Relationships between Health Consciousness, Hedonic Motivation, Image, Self-Efficacy, and Perceived Ease of Use with Perceived Usefulness

During to the SEM the path coefficient between Hedonic Motivation and Perceived Usefulness is 0.515, which is a significant positive correlation with the highest normalized importance given by model B of the ANN. For many studies, the perception of Hedonic Motivation has been viewed as egocentric (Elsden 2015) and an important issue for individual encouragement (Patel 2015), provide motivation (Lazar 2015), acknowledgement of effort (Kim 2016) or giving credit (Consolvo 2016).

The construct Perceived Ease of Use was promoted by the ANN when compared to SEM approach, however, the normalized importance was only 35.2%. This result is in line with previous work that demanded for reminders to be added to the devices (Shih 2015).

Table 4: Comparison between SEM and ANN analysis (output: Health Consciousness).

	SEM Path	SEM Ranking	ANN normalized relative importance	ANN Ranking	Rank Matched?
Perceived Susceptibility to Disease	-0.114	3	68%	2	No
Perceived Severity of Disease	0.283	2	50%	3	No
Habit	0.451	1	99.8%	1	Yes

Table 5: Comparison between SEM and ANN analysis (output: Perceived Usefulness).

	SEM Path	SEM Ranking	ANN normalized relative importance	ANN Ranking	Rank Matched?
Health Consciousness	0.372	2	16.4%	4	No
Hedonic Motivation	0.515	1	100%	1	Yes
Image	-0.013	3	7.7%	5	No
Self-Efficacy	-0.060	4	18.9%	3	No
Perceived Ease of Use	-0.110	5	35.2%	2	No

Table 6: Comparison between SEM and ANN analysis (output: Perceived Ease of Use).

	SEM Path	SEM Ranking	ANN normalized relative importance	ANN Ranking	Rank Matched?
Image	0.146	1	83.3%	2	No
Self-Efficacy	-0.067	2	93.4%	1	No
Perceived Data Control	-0.115	3	72.5%	3	Yes

Table 7: Comparison between SEM and ANN analysis (output: Behavioral Intention to Use / Acceptance).

	SEM Path	SEM Ranking	ANN normalized relative importance	ANN Ranking	Rank Matched?
Perceived Usefulness	0.492	1	97.9%	1	Yes
Image	0.207	2	33.3%	2	Yes
Perceived Ease of Use	-0.115	3	25.9%	3	Yes

Regarding the construct Self-Efficacy, the weak influence is corroborated by model B of the ANN. To some extent, this result is partially contradicted by earlier studies that ask for a tailored efficacy evaluation (Klasnja 2011).

The ANN demoted Health Consciousness giving a low normalized importance that shows a weak influence in Perceived Usefulness.

Since the normalized importance for Image is less than 10%, we may conclude that the effect of Image in Perceived Usefulness is very small in comparison to Hedonic Motivation. This result seems to contradict past research (Harrison 2015) however one should keep in mind that here we are relating Image to Perceived Usefulness.

6.3 Relationships between Image, Self-Efficacy, and Perceived Data Control with Perceived Ease of Use

The construct Image showed a significant relationship with Perceived Ease of Use obtained during the SEM and even though it does not have the highest normalized importance it is a high value according to Model A of the Artificial Neural Networks analysis. This finding of this research is compatible with the findings of existing studies that state that Image is present as a component of social tracking (Rooksby 2014) and that it exists as both social competition and social comparison (Patel 2015).

The ANN came to empower Self-Efficacy as a relevant construct in its relationship with Perceived Ease of Use opposing a path coefficient of -0.067 that SEM found. The ANN result corroborates with previous research that demanded for good inter-device reliability (Dontje 2015). Nevertheless, one has to take into consideration that an ANN measure with high predictive accuracy has both a linear and non-linear relationship among variables.

During the SEM the path coefficient between Perceived Data Control and Perceived Ease of Use is -0.115, which is in accordance with the lowest normalized importance ranking but a high value of 72.5% given by model B of the ANN. To some extent, this result is partially supported by earlier studies

which emphasize that the personal information collected by self-tracking can be highly sensitive (Lupton 2017).

One should note that these results are influenced by the fact that in Model C the Average RMSE value of the testing is higher than the Average RMSE for training.

6.4 Relationships between Perceived Usefulness, Image, and Perceived Ease of Use with Behavioral Intention to Use / Acceptance

Perceived Usefulness with the highest normalized importance (97.9%) given by Model D of the Artificial Neural Networks approach was found in the Structural Equation Modeling to have a significant relationship in predicting Behavioral Intention to Use / Acceptance. This finding supports prior research, as the suggestions for the designers of activity trackers to facilitate micro-plans (Gouveia 2018), add a wide variety of adjustable goals (Clawson 2015), and have adjustable tracking goals (Epstein 2015).

The construct Image shows a significant influence in predicting Acceptance in both approaches. The finding of this research is compatible with the findings of existing studies as the influence of activity trackers on lifestyle (Consolvo 2006) and the importance of aesthetics and form (Harrison 2015).

Concerning the SEM, the path coefficient from Perceived Ease of use to Acceptance is -0.115, nevertheless the normalized importance given by model D of the ANN was 25.9%. This result is in line with previous work which points out that people are using activity trackers in a rather limited manner (Didziokaite 2017).

7 CONCLUSIONS

This research aimed to study beliefs and behavioral variables that impact the acceptance and use of activity trackers. It looked to an established technology acceptance model dedicated to activity trackers that was obtained via Structural Equation

Modeling. These constructs of the model are used in an Artificial Neural Network as the input units of four ANN. The Root Mean Square of Errors with the highest value of 0.118 indicates that the ANN method achieves high prediction accuracy.

The constructs of the model were divided in four ANNs. Model A had as inputs the constructs: Perceived Susceptibility to Disease, Perceived Severity to Disease, and Habit, while the output was the construct Health Consciousness. Model B had as inputs Hedonic Motivation, Image, Self-Efficacy, Health Consciousness, and Perceived Ease of Use, while the output was Perceived Usefulness. Model C had as inputs Image, Self-Efficacy, and Perceived Data Control, while the output was Perceived Ease of Use. Model D had as inputs Perceived Usefulness, Image, and Perceived Ease of Use, while the output was Behavioral Intention to Use (BIU) / Acceptance.

When comparing the results of SEM and ANN analysis, the main disparity lies in the strength of the effect of the construct Self-Efficacy with regards to Perceived Ease of Use. The ANN analysis increases the importance of Self-Efficacy in the Perceived Ease of Use of activity trackers. Even though with a lower impact, the ANN also increases the importance of Perceived Susceptibility of Disease when related to Health Consciousness. On the other hand, it also decreases, with a not so high impact the importance of Health Consciousness in Perceived Usefulness.

The ANN were able to emphasize the strengths and weaknesses of the model obtained via SEM. Furthermore, this research shows the relevance of the two-stage approach integrating SEM and ANN techniques to fine-tune this technology acceptance models and to present valuable information that can be utilized to increase the acceptance and usability of activity trackers as well as to enhance device designs. This research is restricted in the sense that it would be interesting to include control variables such as age and gender and compare the results. Also, it used a cross-sectional approach to obtain the responses of the activity trackers users at one point in time. Hence, in a future study one may repeat the questionnaire to the same users in a longitudinal approach to examine the temporal effects.

REFERENCES

- Akers, J. D., Cornett, R. A., Savla, J. S., Davy, K. P., & Davy, B. M. (2012). Daily self-monitoring of body weight, step count, fruit/vegetable intake, and water consumption: a feasible and effective long-term weight loss maintenance approach. *Journal of the Academy of Nutrition and Dietetics*, 112(5), 685-692.
- Brown, S. A., & Venkatesh, V. (2005). A model of adoption of technology in the household: A baseline model test and extension incorporating household life cycle. *Management Information Systems Quarterly*, 29(3), 11.
- Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., & Mamykina, L. (2015, September). No longer wearing investigating the abandonment of personal health-tracking technologies on craigslist. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 647-658).
- Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS quarterly*, 189-211.
- Consolvo, S., Everitt, K., Smith, I., & Landay, J. A. (2006, April). Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 457-466).
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
- Didziokaite, G. (2017). Mundane self-tracking: calorie counting practices with MyFitnessPal (Doctoral dissertation, Loughborough University).
- Dontje, M. L., De Groot, M., Lengton, R. R., Van Der Schans, C. P., & Krijnen, W. P. (2015). Measuring steps with the Fitbit activity tracker: an inter-device reliability study. *Journal of medical engineering & technology*, 39(5), 286-290.
- Elsden, C., Kirk, D., Selby, M., & Speed, C. (2015, April). Beyond personal informatics: Designing for experiences with data. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2341-2344).
- Epstein, D. A., Ping, A., Fogarty, J., & Munson, S. A. (2015, September). A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 731-742).
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4), 1149-1160.
- Gouveia, R., Karapanos, E., & Hassenzahl, M. (2018, April). Activity tracking in vivo. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).
- Harrison, D., Marshall, P., Bianchi-Berthouze, N., & Bird, J. (2015, September). Activity tracking: barriers, workarounds and customisation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 617-621).
- Haykin, S. (2010). *Neural networks: a comprehensive foundation*. 1999. Mc Millan, New Jersey, 1-24.
- Hochbaum, G., Rosenstock, I., & Kegels, S. (1952). Health belief model. *United states public health service*, 1.

- IDC. Press release from IDC on December 06, 2021. Retrieved March 04, 2022, from <https://www.idc.com/getdoc.jsp?containerId=prUS48460121>.
- Jayanti, R. K., & Burns, A. C. (1998). The antecedents of preventive health care behavior: An empirical study. *Journal of the academy of marketing science*, 26(1), 6-15.
- Kim, D. J., Lee, Y., Rho, S., & Lim, Y. K. (2016, May). Design opportunities in three stages of relationship development between users and self-tracking devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 699-703).
- Kim, J., & Park, H. A. (2012). Development of a health information technology acceptance model using consumers' health behavior intention. *Journal of medical Internet research*, 14(5), e133.
- Klasnja, P., Consolvo, S., & Pratt, W. (2011, May). How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3063-3072).
- Lazar, A., Koehler, C., Tanenbaum, T. J., & Nguyen, D. H. (2015, September). Why we use and abandon smart devices. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 635-646).
- Leong, L. Y., Ooi, K. B., Chong, A. Y. L., & Lin, B. (2013). Modeling the stimulators of the behavioral intention to use mobile entertainment: does gender really matter?. *Computers in Human Behavior*, 29(5), 2109-2121.
- Li, I., Dey, A., & Forlizzi, J. (2010, April). A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 557-566).
- Liébana-Cabanillas, F., Marinković, V., & Kalinić, Z. (2017). A SEM-neural network approach for predicting antecedents of m-commerce acceptance. *International Journal of Information Management*, 37(2), 14-24.
- Limayem, M., Hirt, S. G., & Cheung, C. M. (2007). How habit limits the predictive power of intention: The case of information systems continuance. *MIS quarterly*, 705-737.
- Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011, May). I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2409-2418).
- Lupton, D. (2017). Self-tracking, health and medicine. *Health Sociology Review*, 26(1), 1-5.
- Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information systems research*, 2(3), 192-222.
- Myers, B., Hudson, S. E., & Pausch, R. (2000). Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1), 3-28.
- Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Pearson education.
- Patel, M. S., Asch, D. A., & Volpp, K. G. (2015). Wearable devices as facilitators, not drivers, of health behavior change. *Jama*, 313(5), 459-460.
- Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-SEM in "MIS Quarterly". *MIS quarterly*, iii-xiv.
- Rooksby, J., Rost, M., Morrison, A., & Chalmers, M. (2014, April). Personal tracking as lived informatics. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1163-1172).
- Shih, P. C., Han, K., Poole, E. S., Rosson, M. B., & Carroll, J. M. (2015). Use and adoption challenges of wearable activity trackers. *ICConference 2015 proceedings*.
- Sol, R., & Baras, K. (2016, September). Assessment of activity trackers: toward an acceptance model. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (pp. 570-575).
- Spiekermann, S. (2007). *User control in ubiquitous computing: design alternatives and user acceptance*. Habilitation Humboldt Universität Berlin.
- Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1), 43-62.
- Tan, G. W. H., Ooi, K. B., Leong, L. Y., & Lin, B. (2014). Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-Neural Networks approach. *Computers in Human Behavior*, 36, 198-213.
- Tang, L. M., Meyer, J., Epstein, D. A., Bragg, K., Engelen, L., Bauman, A., & Kay, J. (2018). Defining adherence: Making sense of physical activity tracker data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(1), 1-22.
- Temir, E., O'Kane, A. A., Marshall, P., & Blandford, A. (2016, May). Running: A flexible situated study. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2906-2914).

Between the Buttons: Stress Assessment in Video Games using Players' Behavioural Data

Susanna Brambilla^a, Giuseppe Boccignone^b, N. Alberto Borghese^c
and Laura A. Ripamonti^d

Department of Computer Science, University of Milan, Via Celoria 18, Milan, Italy
susanna.brambilla@studenti.unimi.it, {giuseppe.boccignone, alberto.borghese}@unimi.it, ripamonti@di.unimi.it

Keywords: Affective Computing, Video Game Design, Human-Computer Interaction, Stress, Physiological Data, Behavioural Data, Virtual Reality, Machine Learning.

Abstract: Flow lies at the heart of the interaction between players and video games. It is usually regarded as the optimal experience blooming in the fragile equilibrium that floats between boredom and anxiety. Under such circumstances, stress assessment can be a crucial experiential marker. In this preliminary study, we propose a computational approach to characterise the stress level of video game players, suitable to be exploited in the development of adaptive video games while enhancing players' experience. To such purpose, a Virtual Reality (VR)-based video game has been created to gather data from participants. The information collected includes both physiological data and motion behavioural data (from game controllers), as well as the subjects' self-reports of perceived stress. Behavioural data are specifically considered in the work presented here. We characterize the stress level evolution in terms of state-space dynamics, which is suitable for either discrete (classification) and continuous stress level assessment. Different experiments have been performed and results so far obtained are encouraging. In particular, along the stress vs. no-stress classification test, an accuracy of up to 84.4% is achieved by using VR-based data.

1 INTRODUCTION

The creation of player-centered technologies, putting players and their objectives at the center of the design and development process, is heavily emphasized in the current video game research realm. Due to their interactive nature, games have a significant potential to elicit in players a variety of cognitive, affective, and behavioral reactions. In such perspective, entertainment and engagement play an essential role and are often related to the Csikszentmihalyi's 'Flow theory' (Csikszentmihalyi and Csikszentmihalyi, 1990). Flow is a state of elevated concentration and enjoyment, in which a person is neither anxious nor bored, but completely absorbed in the game.

As such, flow is related to the stress level experienced by the player. Indeed, according to flow theory, a lack of stress can lead to boredom and loss of motivation, while too much stress can cause the player

to suffer from anxiety. Both of them can kill the enjoyment of the gaming experience. It is worth noting that video games do not generate stress *per se*; rather, it is the player's own interaction with the game that results in a more or less stressful personal experience, depending on the challenge level of the game, the skill level of the player and game genre preferences (Fullerton, 2014; Schell, 2008).

On the other hand, the expectancy of the players and its fulfillment or failure not only are ingredients triggering a stress response, but also contribute to the players' affective state (Lebois et al., 2016). Players are constantly demanding more immersive experiences, and players' emotional involvement can be seen as a sign of a high level of immersion in the game. Affective-based interaction can increase player engagement, and the emotion-driven game adaptation helps with the personalization of the playing experience, allowing for the fulfillment of each player's unique demands (Yannakakis and Togelius, 2018).

Eventually, stress response and emotional involvement dynamics together contribute in modulating behavioural and physiological response along the gaming session. Yet, there are few examples of how emo-

^a  <https://orcid.org/0000-0002-7975-8724>

^b  <https://orcid.org/0000-0002-5572-0924>

^c  <https://orcid.org/0000-0002-0925-3448>

^d  <https://orcid.org/0000-0001-8167-7870>

tion recognition systems can be used in video games.

The hypothesis that the players' arousal is correlated with the pressure applied to the buttons of the game pad has been investigated by Sykes and Brown (2003). With this aim, they created a video game similar to the classic arcade game *Space Invaders*, with three difficulty levels (easy, medium, hard). The pressure used by players during a game session was recorded and compared across the three levels. According to the study's findings, players push the game pad buttons far harder as the difficulty level rises.

Hiramon (Frommel et al., 2018) is another good example of technique for players' emotional state detection involving their input on game controllers. *Hiramon* is a serious game which aims at teaching players to write Japanese hiragana characters. Players are first shown how characters are written and are required to replicate them (learning period). Then, players are challenged by enemies to write specific characters randomly chosen from those that they have already learned (fight period). During the game, input parameters on a graphic tablet and in-game performance are collected. Self-reports of emotions were also gathered via a questionnaire after each fight of the game, in order to train an ML model to predict levels of valence, arousal, and dominance for a classification problem, with an accuracy of up to 74%.

We aim at overcoming the limitations of the presented works by exploiting new technologies, which provide more opportunities for interaction, and by considering the evolution of the stress level in terms of state-space dynamics.

The main goal of the work presented here is to develop a method for improving the game-play experience by exploiting a computational approach to seamlessly assess players' level of stress while playing, the latter being inferred from motion behavioural data collected from game controllers.

To this purpose, the game used for the experiment has been developed for a Virtual Reality (VR) environment, due to its high level of immersion and realism. Immersive technologies like VR can help to arouse a sense of presence in the game, enhancing players' fun and involvement and eliciting stronger emotional reactions (Pallavicini et al., 2018b; Marín-Morales et al., 2020).

The paper is organized as follows. Section 2 describes the methods with a focus on the design of the game and its elements. Section 3 details the experimental setting used to collect the users' data. Section 4 illustrates the analyses carried out using the collected data, Section 5 focuses on the learning models used to build the stress assessment system. Section 6, presents and discusses the results achieved so

far; eventually, in Section 7, some preliminary conclusions are drawn.

2 METHODS

Input data from controllers, coupled with players' self-assessment, are used to infer the stress level. Self-assessment annotations correspond to their perceived stress, used as labels of the generated dataset. For some subjects, self-reported valence and arousal levels were also collected. The survey of Klein-smith and Bianchi-Berthouze (2012) and of Karg et al. (2013) are significant reviews concerning the affective body expression perception and recognition. They showed that the body is a valid modality for recognizing affect. Body movements were analyzed and used to identify emotion-specific features in order to recognize basic emotions (Ahmed et al., 2019)

The motion behavioural data were taken from the *Oculus Quest 2*¹ devices; jointly, physiological signals were recorded via the *Empatica E4*² wristband, to be exploited in future analyses.

2.1 Game Design

The video game³, developed in *Unity Engine 2020.3.19f1*, is a first-person horror survival game for a VR environment (Figure 1). The game is set on an abandoned space station and the players' goal is to escape from the station, facing enemies in a hostile environment, surviving with few resources. The horror survival sub-genre was chosen because it has been proven excellent at evoking players' intense emotions, particularly stress (Vachiratamporn et al., 2013).

Some stressors, such as the presence of monstrous enemies, scary sounds and disturbing music, the lack of visibility due to the poor lighting, and the scarcity of resources, were accurately created to enhance a stressful affective state in the players. Lebois et al. (2016) argue that when typical features of stressful situations are present, people categorize events and situations as stressful. Under such circumstances, the most typical features associated with a situation perceived as stressful are: (a) the situation violates personal expectations, there is a discrepancy between the

¹<https://store.facebook.com/it/quest/products/quest-2/>

²<http://www.empatica.com/research/e4/>

³A demonstration of the recorded gameplay can be found at https://drive.google.com/file/d/1K-DNeW14NNvwgorEw9AEHh_0EnwJIS50/view?usp=sharing

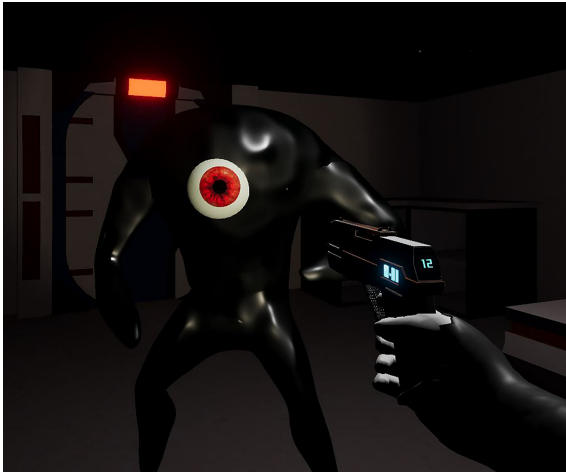


Figure 1: Screen of the video game developed for the experiment.

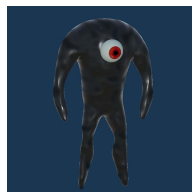
individual's expectation and the actual situation (Higgins, 1989); (b) the situation or event threatens the well-being of self (Lazarus, 1993); (c) the subject's personal resources available for coping with the situation are not sufficient (Lazarus, 1993). When the core stress conditions are met, they produce different responses of negative emotions and physiological reactions. An individual may experience anxiety, fear, sadness, danger, or even a combination of these categorical emotions during a stressful situation. Subsequently, physiological changes occur in the cardiovascular, endocrine, respiratory, and autonomic systems. Stress perception is thus the result of assigning a situation to the category of a stressful experience.

The two types of enemies designed for the game are described in Table 1. Both of them have a body composed by black slime and a big red eye.

All the designed elements were placed on the game map to create different phases, each one with the aim of inducing a different stressful experience. The structure of the map is linear: six rooms were designed and divided into four stress phases (baseline, low stress, medium stress, and high stress), arranged so that the amount of generated stress gradually increases from the first phase to the last one. The stress phases are organized as follows.

- **Baseline Phase:** a little room with a terminal introducing the players to the game world. The players are asked to put their right hand on a panel and remain still for 15 seconds. This expedient is used to collect baseline data from users.
- **Low Stress Phase:** the astronaut's room where the players are safe. The oxygen does not drop and no enemy can be sound. The light allows the players to see quite everything, and some far noises can be heard. All the items necessary to advance in

Table 1: The two types of enemies with their description.

	DESCRIPTION
	Creatures similar to a frog with a big eye at the center of the head. They move quickly and attack jumping towards the player, but they are not really strong.
	Humanoid-shaped creatures with an eye protruding from their torso. They are not very agile and hit the player by waving their arms. If stabbed in the eye, they remain stunned for a few seconds and stop moving.

the game (a gun, a knife, and ammunition) can be found here.

- **Medium Stress Phase:** composed of two rooms and two corridors, here the enemies are introduced. The light starts to lower, and many scary sounds come from everywhere.
- **High Stress Phase:** the last two rooms with a total of five enemies scattered around the areas, in the rooms and corridors that connect them. Approaching the first room, an alarm starts to bother the players. The oxygen runs out, and the character begins to gasp quickly. Here, it is practically impossible to see without using a flashlight. The final room is the one in which there is the greatest concentration of enemies, and it is where the players find the escape pod.

The game session of the test is brief, lasting approximately 5-10 minutes, during which all the game elements and events are presented in the same way to all players, making the experiment more controllable. This also allows a comparison between the data acquired from each different subject.

3 EXPERIMENTAL SETTING

This section covers the experimental process, providing details on the subjects, the data acquisition during the game session, and the self-reports.

Before starting the experiment, a pilot test phase was conducted. This was necessary to understand if there was something to modify in order to accomplish

the research goal, to check the validity of the collected data in relation to both the temporal duration of the game session and designed stress phases in the game. It was especially useful to adjust the baseline phase.

The experiment lasted, at most, 30 minutes for the participants who just annotated the stress values: 5-10 minutes for the early questionnaires, 5-10 minutes for the game session, and 5-10 minutes for the labeling phase (see Section 3.3, for details). For the participants who also self-assessed valence and arousal values, the experiment had a total duration of about 50 minutes.

3.1 Participants

A total of 16 volunteer students participated in the experiment. Volunteers did not receive any payment or credit for their collaboration. All of them reported having no anxiety disorders and no neurological alterations. Prior to the study, participants read and signed a consent form because biometric and demographic, personal data were handled. Before starting the game session, participants were required to complete two questionnaires: the Perceived Stress Scale (PSS) questionnaire (Cohen et al., 1983) and a demographic questionnaire.

The first one is a standard psychological instrument for measuring perceived stress and refers to feelings and thoughts during the last month. From the PSS, it resulted that all the participants felt moderately stressed in their life before starting the experiment.

The second questionnaire, which was designed specifically for this work, refers to demographic and static data about the participants. The answers indicate that 15 males and 1 female in the range of 22 and 29 years, with a mean age of 24.68, took part in the experiment. The majority of the subjects claimed to spend, on average, around 11-20 hours per week playing video games; two participants only declared they never play video games. The most popular game genres were action and adventure. Only two participants reported they like horror games. Finally, half of the participants reported that they had never experienced VR environments before the test, whilst, three participants declared they often play VR games.

3.2 Data Acquisition

After completing the two questionnaires, each participant was equipped with the sensors. First, they wore the *E4* on their non-dominant hand, to minimize the movement artifacts. Then, they put on the headset and its controllers, and instructions about the buttons

needed to play were given. During the game session, participants were standing and could rotate freely on themselves.

Once this preparation phase was completed, the game, during which all the data were collected, was started. When the participants reached the end of the game, all the devices were removed. A problem that can arise when using VR devices is the motion sickness, which is a feeling of disorientation or nausea. All of them, even the novices, declared that they did not suffer from motion sickness.

3.3 Self-assessment

In order to validate the experimental protocol and to label the data, self-reports of each participant were collected. Indeed, asking the participants directly about their perceived experience is the most direct way to record their inner state.

Self-reports were collected by exploiting *DANTE*⁴ (Dimensional ANnotation Tool for Emotions). *DANTE* is a web-based annotation tool helpful for studying affective responses in presence of a stimulus (Boccignone et al., 2017). It provides an interactive version of the Self-Assessment Manikin (SAM) questionnaire (Bradley and Lang, 1994), and allows for continuous annotation of valence and arousal by moving a slider on a scale with values ranging from -1 to 1 and a step of 0.001.

Specifically for this work, *DANTE* was extended by adding a stress slider in order to allow participants for the annotation of their own perceived stress level. Such slider represents the continuous extension of a perceived stress Likert scale (visualised in the form of a color bar, cfr. Figure. 2), spanning a range from low (-1) to high (1) stress level.

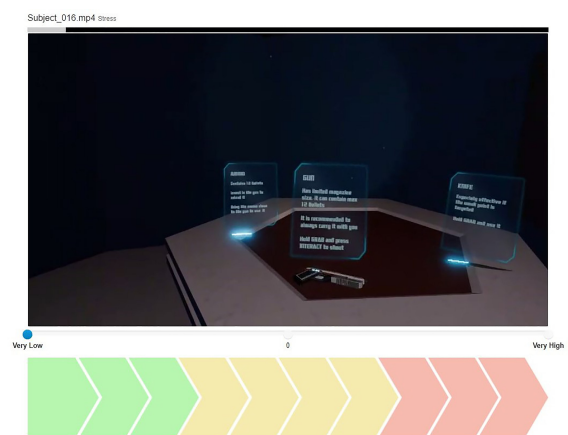


Figure 2: Self-assessment of stress using DANTE.

⁴<https://github.com/phuselab/DANTE>

The gameplay of each participant was recorded during the game session and the stress label sequence or targets over time $\{l_t\}_{t=1}^T$ were collected from all participants, while seeing their own gameplay video.

Seven participants were also asked to evaluate in real-time their subjective emotional experiences in the 2-dimensional affective space, defined by valence and arousal dimensions. In our context, the valence dimension indicates how pleasurable the gaming experience was and (negative to positive), while the arousal dimension denotes how arousing the gaming experience was (low to high).

4 DATA ANALYSIS

The acquired data had to be carefully scrutinized before confronting with the stress level assessment process. Pre-processing, feature extraction, data analysis, and feature selection are the traditional steps considered in the following.

In this paper, we focus on the data related to the *Quest 2* headset and controllers, since, in the pilot test phase, they proved to be particularly significant for the achievement of the initial purpose. The *Quest 2* data concern movement and behavioral information. They represent both the players' action in the video game (which keys are pressed and how hard) as well as the players' movements (basically, their rotation with velocities and accelerations) in the real world. The data were sampled at 64 Hz.

4.1 Preprocessing

The raw data were pre-processed in order to check missing values, noisy data, and other inconsistencies.

The *Quest 2* data and the self-assessed stress labels were first synchronized using the start and end timestamps of the experiment. The timestamps were also saved after each stress phase to make it possible to separate the data related to different stress levels.

Then, the data were standardized using the baseline signal, with the Z-score (or standard score) method, first participant-wise and then using the whole subjects' data. Using the baseline to standardize data provides a common reference for further processing of each feature.

4.2 Feature Extraction

All signals were segmented using a sliding window with a size of 6 seconds and a shift of 1 second. Since the system was designed to be used in real-time dur-

ing the game session and require a fast response to identify stress, a small window size was chosen.

There is a limited literature on the feature extraction process for the data gathered from *Quest 2*. Thus, we selected the most significant statistical features for all the data obtained from head and left/right hands: mean, standard deviation, minimum and maximum values are used, as well as the average number of times the buttons were pressed within the designated time period. The features extracted for each category of data are shown in Table 2.

Table 2: List of the features extracted from the data collected from *Quest 2*. Abbreviations: h = head, lh = left hand, rh = right hand, vel = velocity, acc = acceleration, ang = angular, press = pressure, pos = position, std = standard deviation, # = number.

Data	Feature	Description
h vel	μ	mean
h ang vel	σ	std
h acc	min	min value
h ang acc	max	max value
lh/rh vel		
lh/rh ang vel		
lh/rh acc		
lh/rh ang acc		
lh/rh grip press		
lh/rh trigger press		
lh/rh thumbstick pos x		
lh/rh thumbstick pos y		
lh/rh grip pressed	$\mu\#press$	mean of the
lh/rh trigger pressed		# of times the button is pressed

After the extraction step, the features were normalized by scaling and translating them in the range between 0 and 1.

4.3 Statistical Analysis

A preliminar statistical analysis was conducted on the features that were extracted. This task was completed using *Autorank*⁵ (Herbold, 2020). *Autorank* is a Python package used to automatically compare paired populations. The package analyzes the distribution of the data and automatically decides which tests to perform. The first populations compared coincide with the stress phases. The means of all the extracted features from the *Quest 2* data were used. Each subject's data was first analyzed individually, and all the subjects' data were then combined to perform the analysis. This analysis produced no useful

⁵<https://github.com/sherbold/autorank>

findings. We surmise that this outcome might be due to the fact that the stress experienced during a certain game phase is not continuous throughout it. So, when comparing the complete temporal sequence of one phase with another, no relevant differences are likely to emerge. Then, samples were drawn from subjects with different VR-experience level (experienced and non-experienced). This value was extracted from the demographic questionnaire. To carry out the comparison, an analysis for each stress level was carried out. Even in this case, no statistically significant differences emerged, most likely because of the gap between more experienced and less experienced subjects is not very great.

Subsequently, data were analyzed by comparing their correlations. Pearson correlation was used. Correlation matrices were computed first for every single subject, and then combining the subjects' data together, first dividing the stress phases and then using them together. According to the analyses carried out by differentiating the four designed levels of stress, it appears that, in general, all the features correlate better with the perceived stress level in the baseline phase and the high stress phase. Even using the stress phases together, a good level of correlation emerged with some of the features. In the last study, participants who assessed their emotional state using valence and arousal in addition to stress were used. Their data were separated from those of the other participants to examine the correlation between valence and arousal levels and self-reported stress level. In all cases, the arousal level correlates positively with the stress level. By contrast, the correlation between valence and stress level can be either positive or negative, depending on whether the subject enjoys or not the horror games. A good degree of correlation results between the features and the valence level and, even more markedly, between the features and arousal level, which raises the possibility of further analyses.

4.4 Feature Selection

The final step in processing the extracted features involved selecting the most relevant ones. In this work, a filter approach was used. Since the relationship (i.e. the correlation) between each numerical feature as input and the target label (the stress level) had been calculated, it could be used to select the most informative subset of the original features. Univariate feature selection was performed. The p-values and the F-scores were computed. The features with a p-value less than 0.05 and the higher score were selected. In the Tab.3, all the features selected to be used as the input of the learning models are shown.

5 MODEL-BASED STRESS ASSESSMENT

We assume a State Space Model (SSM) for stress level dynamics. Namely, we use a partially observed Markov model, in which the hidden state s_t is a random variable that evolves over time according to a Markov process and each hidden state generates some behavioural observations, the random vector y_t at each time step (in what follows to avoid cumbersome notation we do not distinguish, unless needed, between a RV X and its realization $X = x$). In brief, a general SSM defines the joint distribution

$$P(y_{1:T}, s_{1:T}) = P(s_1) \prod_{t=1}^T P(y_t | s_t) P(s_t | s_{t-1}), \quad (1)$$

where $P(s_t | s_{t-1})$ is the state transition model and $P(y_t | s_t)$ is the behavioural observation model.

Here we are interested in exploiting the SSM to perform posterior inference about the hidden states or stress level state estimation. In particular, it can be used either for online inference, by inferring the probability of the hidden state s_t at current time $t < T$ via the filtering posterior distribution $P(s_t | y_{1:t})$, or to estimate s_t , at any time $t \in [1, T]$, given the full sequence of observations, via the smoothing distribution $P(s_t | y_{1:T})$.

Further, the SSM is a suitable approach since it allows characterising stress level s_t either as a discrete RV or a continuous RV. The former can be exploited to segment/classify the behavioural observations into a finite set of stress states; the latter can be used in the service of continuously tracking stress level dynamics over time.

For the simulations reported in the present work we adopted the Hidden Markov Model (HMM, see Bishop, 2006 for a review) and a variant of the Kalman Filter (KF), the Discriminative Kalman Filter (DKF, Burkhart et al., 2020) as implementation models of the discrete and the continuous SSM, respectively.

It is worth noting, that, in principle, more complex implementation models (e.g., resorting to deep neural nets-based models, Girin et al., 2021) could be used, provided that, different from here, a sufficiently large dataset is available.

5.1 Simulation Details

5.1.1 Discrete Characterization of Stress Dynamics

HMM is an extensively used technique to model temporal information (Bishop, 2006), in particular for

Table 3: List of the features. The ones selected to be given as input to the learning models are checked with a $\times$. Abbreviations: vel = velocity, acc = acceleration, ang = angular, press = pressure, pos = position, # = number.

TYPE	DATA	FEATURE				
		μ	σ	min	max	$\mu\#press$
head	vel	$\times$	$\times$	$\times$	$\times$	
	ang vel	$\times$	$\times$	$\times$	$\times$	
	acc	$\times$	$\times$	$\times$	$\times$	
	ang acc	$\times$	$\times$	$\times$	$\times$	
left hand	vel	$\times$	$\times$	$\times$	$\times$	
	ang vel	$\times$		$\times$		
	acc	$\times$	$\times$	$\times$	$\times$	
	ang acc	$\times$	$\times$	$\times$	$\times$	
	grip press	$\times$	$\times$	$\times$	$\times$	
	trigger press	$\times$	$\times$		$\times$	
	thumbstick pos x	$\times$	$\times$	$\times$	$\times$	
	thumbstick pos y	$\times$	$\times$	$\times$	$\times$	
	grip pressed					$\times$
	trigger pressed					$\times$
right hand	velocity	$\times$	$\times$	$\times$	$\times$	
	ang vel	$\times$	$\times$	$\times$	$\times$	
	acc	$\times$	$\times$	$\times$	$\times$	
	ang acc	$\times$	$\times$	$\times$	$\times$	
	grip press	$\times$		$\times$	$\times$	
	trigger press	$\times$	$\times$	$\times$	$\times$	
	thumbstick pos x			$\times$		
	thumbstick pos y					
	grip pressed					$\times$
	trigger pressed					$\times$

speech recognition or facial expressions recognition applications, and it has been used before in video games by Mishra and Ratnaparkhi (2018) for real-time recognition of players' emotions.

Here, we exploited the HMM implementation provided by the *hmmlearn*⁶ library; the behavioural observation distribution was assumed to be continuous and Gaussian, $P(y_t | s_t = j) = \mathcal{N}(y_t | \mu_j, \Sigma_j)$, μ_j, Σ_j being the mean and covariance, respectively, of the observation y_t conditioned on the j -th state.

In the discrete setting, the stress assessment via HMM can be conceived in broad terms as that of solving a classification problem.

To this end, experiments were conducted with different number of discrete states (classes). First, a binary stress vs no-stress classification was performed. Next, three classes, i.e. no-stress vs low stress vs high stress, were used. Finally, a multiclass problem was solved, with the classes identifying the stress phases of the game partition.

In order to employ the HMM for classification purposes, label values were discretized into different levels and a separate model was built for each class and trained on the class samples, separating each sub-

ject's sequence of observation to preserve the temporal information.

Leave-One-Out Cross-Validation (LOOCV) was applied. The learning was performed once for each subject, using the selected subject as a test set and all the others as a training set. Then, the estimate of the performance of each model was computed by averaging the scores over all the trials.

The training process was repeated as many times as the number of subjects. Each model was trained using the forward-backward algorithm. The parameters were initialized randomly.

After training, each test sequence was scored in relation to each model. The log-likelihood of the sequence given as input was returned with respect to the model in use. The test set was then classified in the class of the model returning the highest log-likelihood result. The overall accuracy was then calculated by comparing the predicted stress labels to the actual stress labels. This accuracy was given as the average of the accuracy of all the models trained across all LOOCV iterations.

⁶<https://github.com/hmmlearn/hmmlearn>

5.1.2 Continuous Characterization of Stress Dynamics

The DKF is a novel filtering method that provides a fast, analytic approximation for models with linear, gaussian dynamics but nonlinear, nongaussian observations. For continuous SSMs, the conditional distribution $P(s_t | y_{1:t})$ can be expressed recursively using Bayes' rule, $P(s_t | y_{1:t}) = \frac{P(y_t | s_t)P(s_t | y_{1:t-1})}{P(y_t | y_{1:t-1})}$, together with the Chapman-Kolmogorov equation, $P(s_t | y_{1:t-1}) = \int P(s_t | s_{t-1})P(s_{t-1} | y_{1:t-1})ds_{t-1}$. The standard KF models the densities $P(s_t | s_{t-1})$ and $P(y_t | s_t)$ as linear and gaussian so that the posterior $P(s_t | y_{1:t})$ is also Gaussian and efficiently computable. The DKF is based on the key approximation for the likelihood

$$P(y_t | s_t) = P(y_t) \frac{P(s_t | y_t)}{P(s_t)} \approx k(y_t) \frac{\mathcal{N}(s_t | f(y_t), Q(y_t))}{\mathcal{N}(s_t | 0, S)} \quad (2)$$

where $k(\cdot)$ is a normalizing constant, $f(\cdot), Q(\cdot)$ the conditional mean and covariance of s_t , respectively. The advantage of such approximation is that $f(\cdot), Q(\cdot)$, and S are easier to learn from training data than the full conditional density (see Burkhart et al., 2020 for details). In particular, the conditional mean f can be learned using a number of regression methods; here we adopted a neural network regression (Burkhart et al., 2020).

Before DKF training, the signals were smoothed by applying the Savitzky-Golay filter (Savitzky and Golay, 1964). To evaluate the performance of the model, the normalized Root Mean Squared Error (nRMSE) was calculated. DKF was trained as many times as the number of LOOCV iterations, and the average of all the computed nRMSE values for each iteration was used to determine the overall nRMSE value.

6 RESULTS AND DISCUSSION

Results of the evaluation of the models are presented in this section, along with a discussion on the importance of each individual device. Only the stress labels were taken into account for the evaluation, and different tests were run for each model.

6.1 Discrete SSM Performance

In order to evaluate the results of in terms of classification, confusion matrices and accuracy scores were reported. Based on the affective states designed for the different game phases, different classification tasks were distinguished:

1. using baseline, low stress, medium stress, and high stress as classes, a four-class problem was defined;
2. using baseline, low stress, and high stress as classes, a three-class problem was defined;
3. using the baseline condition acting as the no-stress class and the other states acting as the stress class, a binary problem was defined.

Each setup defined by the classification problem was run sixteen times (the number of folds in LOOCV), and the accuracy measures are averaged over all runs of the repeated LOOCV.

As it can be seen in Figure 3a, the first classification task revealed that the collected VR-based data perform better for the high stress condition. The features have difficulties, in fact, in distinguishing between different stress levels, particularly the low stress state.

Figure 3b shows the results of the second classification problem. The model proved to be valid at recognizing the data in the high stress class and quite good at recognizing the baseline data. In this case, a great deal of the low stress data is incorrectly classified as high stress.

The third classification task is the one that led to the best results, which are shown in the confusion matrix illustrated in Figure 3c. The algorithm proves to be able to discriminate between the two classes, the best score is achieved with the baseline class.

In Table 4, the accuracy scores for all the described classification tasks are reported.

The binary problem was the one in which the model performed better, reaching the accuracy of 84.4%. Overall, results provided evidence that discrete SSMs were good at distinguishing between a state of stress and a state of baseline, but they could hardly classify different stress levels.

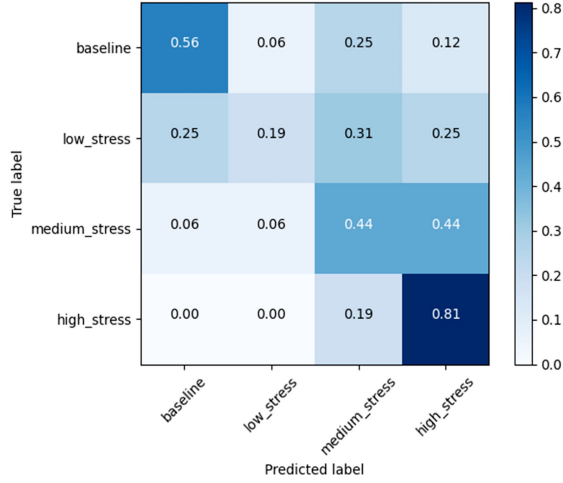
Table 4: Accuracy scores of the model for all the classification tasks.

Accuracy score		
Four-class	Three-class	Two-class
50.0	56.2	84.4

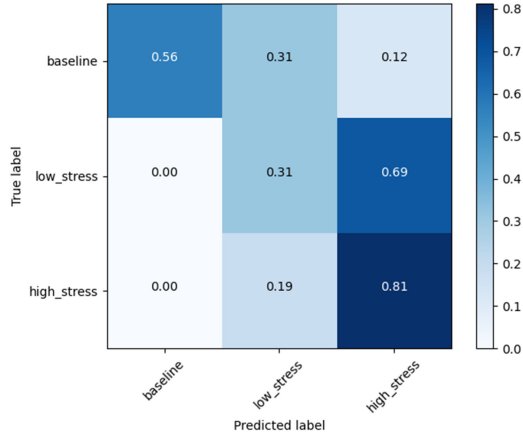
6.2 Continuous SSM Performance

As a preliminary qualitative evaluation, the actual label curve $\{l_t\}_{t=1}^T$, defining the self-assessed stress level, and the predicted stress curve were also plotted and compared with the purpose of visualizing and scrutinizing the results.

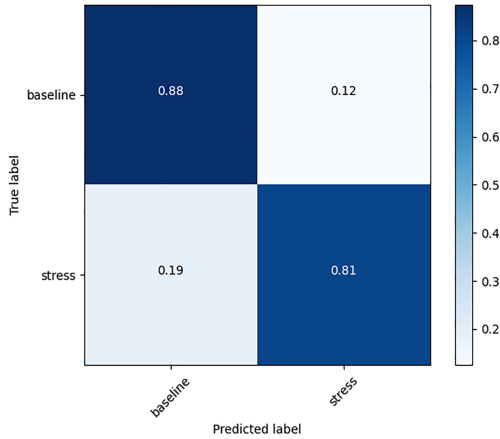
The two curves were plotted for each fold of LOOCV procedure, hence for each test subject. Fig-



(a) Confusion matrix of the four-class problem.



(b) Confusion matrix of the three-class problem.



(c) Confusion matrix of the two-class problem.

Figure 3: Confusion matrices of all the classification tasks using VR-based features.

ure 4 shows one typical example of the inferred latent stress level dynamics, which refers to the trial involv-

ing subject 13 as the test and all the other subjects as training set. The red line indicates the inferred stress curve, whilst the blue line represents the actual label curve $\{l_t\}_{t=1}^T$. Even though the model was able to identify the general trend of the stress level dynamics, these results suggest that the continuous SSM is not accurate enough, at least in the current experimental setting, to make the model fully reliable for game control purposes.

A quantitative evaluation was thus performed in terms of the nRMSE values obtained by comparing the target (self-assessment) and estimated perceived stress curves. The computed nRMSE values are reported in Table 5. The three different input types are compared, nRMSE was calculated for each iteration of LOOCV process using each subject as test set in turn, and then the nRMSE average with respect to all the iterations was computed.

Table 5: The nRMSE values computed for each trial in which the DKF model was run. Subject column indicates the subject who was left out during LOOCV iteration, and AVG is the average value across all trials.

subject	nRMSE
01	0.44
02	0.39
03	0.59
04	0.42
05	0.61
06	0.49
07	0.89
08	0.57
09	0.60
10	0.73
11	0.50
12	0.45
13	0.43
14	0.44
15	1.0
16	0.80
AVG	0.57

6.3 Discussion

The VR-based data provide quite generalizable measures: behavioral and movement data, such as controllers and head velocities, which keys are pressed and at what pressure level, are indicative of the users' emotional states. Information gathered from *Quest 2* devices proved to be useful to identify the perceived stress level of the players. This suggests that, in the context of video games, the VR-based features can offer a rich understanding of the affective states. The

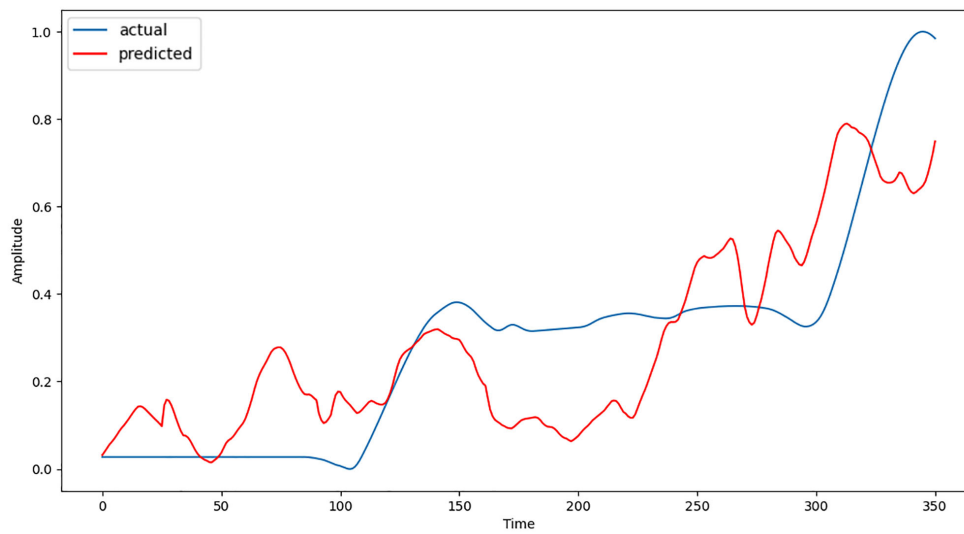


Figure 4: Curves of the actual stress labels and the stress labels predicted by DKF for the VR-based features, using subject 13 as test set.

analyzed context, in fact, differs from any other real context, and players' mental states are strongly related to game events and actions they can do.

7 CONCLUSION

The present work aimed at a preliminary investigation concerning the assessment of players' perceived stress level in VR-based video games. The developed systems can in principle be applied to video games of the same genre, with the same structure and elements as the one developed in this project: a survival horror video game with a baseline phase for calibration and stress elements established on the features of perceived stress defined by Lebois et al. (2016).

When comparing how the utilized algorithms performed on the various tasks, it becomes apparent that the binary state discrete model produced the highest accuracy scores. The achieved classification accuracies are up to 84.4% using the VR-based data.

In conclusion, the results obtained are encouraging and this preliminary study can be useful as a starting point for future research. Further work is required to gather more data from a major number of subjects, and to take the questionnaires into account. Questionnaire answers were not included in the development of the presented model. Training the models using different classes of players who have the same profile (Yannakakis and Togelius, 2018) would result in a more personalized experience.

The potential benefits of the explored approach are twofold: a system incorporating a stress-

dependent control can adapt the content of the game in response to the players' experiential state; stress level modelling/inference can help in developing new game mechanics relying on such index to provide a customized experience for players.

Video games are increasingly being used in the field of mental health (Colder Carras et al., 2018; Pallavicini et al., 2018a). The enjoyment and intrinsic motivation often associated with video games make them a powerful and attractive tool to provide psychological support to people. According to studies, video games can help individuals cope with their stressful life experiences (Pallavicini et al., 2021; Maarsingh et al., 2019). A stress-based serious game that incorporates biofeedback techniques into the game could assist people with stress to handle their emotional and physiological responses to stressors, improving their abilities in everyday life and their mental health.

Ongoing research is focused on two aspects. The first concerns the integration of physiological information (electrodermal activity) with behavioural data for stress level assessment. This aspect paves the way to consider the relationship between the stress response and the affective state of the user.

The second, more generally, relates to the use of stress/affective assessment for the development of affective feedback-based video games. In this perspective, feedback from players' affective states (Bersak et al., 2001) is incorporated into the development of new game content and mechanics, manipulating gameplay so to keep the players in the flow state.

REFERENCES

- Ahmed, F., Bari, A. H., and Gavrilova, M. L. (2019). Emotion recognition from body movement. *IEEE Access*, 8:11761–11781.
- Bersak, D., McDarby, G., Augenblick, N., McDarby, P., McDonnell, D., McDonald, B., and Karkun, R. (2001). Intelligent biofeedback using an immersive competitive environment. In *Paper at the designing ubiquitous computing games workshop at UbiComp*, pages 1–6. Citeseer.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Boccignone, G., Conte, D., Cuculo, V., and Lanzarotti, R. (2017). Amhuse: a multimodal dataset for humour sensing. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 438–445.
- Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.
- Burkhardt, M. C., Brandman, D. M., Franco, B., Hochberg, L. R., and Harrison, M. T. (2020). The discriminative kalman filter for bayesian filtering with nonlinear and nongaussian observation models. *Neural computation*, 32(5):969–1017.
- Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396.
- Colder Carras, M., Van Rooij, A. J., Spruijt-Metz, D., Kvedar, J., Griffiths, M. D., Carabas, Y., and Labrique, A. (2018). Commercial video games as therapy: A new research agenda to unlock the potential of a global pastime. *Frontiers in psychiatry*, 8:300.
- Csikszentmihalyi, M. and Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*, volume 1990. Harper & Row New York.
- Frommel, J., Schrader, C., and Weber, M. (2018). Towards emotion-based adaptive games: Emotion recognition via input and performance features. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 173–185.
- Fullerton, T. (2014). *Game design workshop: a playcentric approach to creating innovative games*. CRC press.
- Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., and Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, 15(1-2):1–175.
- Herbold, S. (2020). Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173.
- Higgins, E. T. (1989). Self-discrepancy theory: What patterns of self-beliefs cause people to suffer? In *Advances in experimental social psychology*, volume 22, pages 93–136. Elsevier.
- Karg, M., Samadani, A.-A., Gorbet, R., Kühnlenz, K., Hoey, J., and Kulić, D. (2013). Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359.
- Kleinsmith, A. and Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33.
- Lazarus, R. S. (1993). From psychological stress to the emotions: A history of changing outlooks. *Annual review of psychology*, 44(1):1–22.
- Lebois, L. A., Hertzog, C., Slavich, G. M., Barrett, L. F., and Barsalou, L. W. (2016). Establishing the situated features associated with perceived stress. *Acta Psychologica*, 169:119–132.
- Maarsingh, B. M., Bos, J., Van Tuijn, C. F., and Renard, S. B. (2019). Changing stress mindset through stress-jam: a virtual reality game using biofeedback. *Games for health journal*, 8(5):326–331.
- Marín-Morales, J., Llinares, C., Guixeres, J., and Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors*, 20(18):5163.
- Mishra, P. and Ratnaparkhi, S. (2018). Hmm based emotion detection in games. In *2018 3rd International Conference for Convergence in Technology (I2CT)*, pages 1–4. IEEE.
- Pallavicini, F., Ferrari, A., and Mantovani, F. (2018a). Video games for well-being: A systematic review on the application of computer games for cognitive and emotional training in the adult population. *Frontiers in psychology*, 9:2127.
- Pallavicini, F., Ferrari, A., Pepe, A., Garcea, G., Zancacchi, A., and Mantovani, F. (2018b). Effectiveness of virtual reality survival horror games for the emotional elicitation: Preliminary insights using resident evil 7: Biohazard. In *International Conference on Universal Access in Human-Computer Interaction*, pages 87–101. Springer.
- Pallavicini, F., Pepe, A., Mantovani, F., et al. (2021). Commercial off-the-shelf video games for reducing stress and anxiety: systematic review. *JMIR mental health*, 8(8):e28150.
- Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.
- Schell, J. (2008). *The Art of Game Design: A book of lenses*. CRC press.
- Sykes, J. and Brown, S. (2003). Affective gaming: measuring emotion through the gamepad. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 732–733.
- Vachiratamporn, V., Legaspi, R., Moriyama, K., and Numa, M. (2013). Towards the design of affective survival horror games: An investigation on player affect. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 576–581. IEEE.
- Yannakakis, G. N. and Togelius, J. (2018). *Artificial intelligence and games*, volume 2. Springer.

Augmented Reality-based Visualization of the Reach of a Collaborative Robot

Fabio Schwarz¹, Maxim Feinleb¹ and Gerrit Meixner² ^a

¹*Robotic Testing and System Engineering, Festo SE & Co. KG, Ruiter Straße 82, Esslingen, Germany*

²*UniTyLab, Heilbronn University, Max-Planck-Str. 39, Heilbronn, Germany*
fasc@hotmail.de, maxim.feinleb@festo.com, gerrit.meixner@hs-heilbronn.de

Keywords: Augmented Reality, Collaborative Robot, Unity3D, Robot, Cobot.

Abstract: With increasing globalization and higher competition on the one hand, and the same high level of personnel costs on the other, the pressure on companies to rely more on automation is growing. At the same time, with increasing individualization and complexity of products, batch sizes are shrinking, making versatile production machines with quick production changes and easy set-up processes necessary. In order to still cope with the complexity and at the same time use the human worker only for responsible tasks, robots have to take over the tedious tasks and work together with the human worker at the same time. To enable this working condition, the collaborative robots need to be set up and installed at the human's working place first. This increases the safety requirements and leads to a large installation effort. To tackle these problems, improve the set-up and to simplify the collaborative work of the robot and the human worker, an augmented reality application is presented that helps by visualizing the reach of the robot. The proposed solution uses Unity3D to create an augmented reality application which recognizes the robot with the help of ArUCO markers. Once recognized, the reach of the robot is shown in the application in such way that the human worker is able to see the reach of the robot.

1 INTRODUCTION

With the change of industry to an "Industry 4.0" and increasing globalization the ever-growing need for collaborative robots increases (Galin and Meshcheryakov, 2019). These robots help to tackle the higher competition in the market by automating production steps. This will help workers in manufacturing products on mass or more individualized products with smaller batch sizes. They can either be installed as a third arm, to hand the worker parts for assembly, or work autonomously in supplying chains. They are small and outfitted with sensors to perceive their surroundings and not to collide with humans, but also 'see' what they can grip and where to move. Because of this versatility, they need to be set up for each specific task. They need to be placed in such way that they can access all the instruments and parts they need to complete their task. They also must not interfere with the already existing machines and must not interrupt a worker in his workflow (Kopp et al., 2021).

To improve the collaboration between robot and

worker, the worker must be able to have a way to communicate with the robot or at least understand what the robot is working on at any moment in time. To come closer to a solution, this paper describes how to create an augmented reality (AR) application which will show the worker the reach of the robot.

2 STATE OF THE ART

AR gained a lot of traction in the last few years not only in pop culture but also in the industry. This validates a forecast given by Shao-Ning Chang and Wei-Lun Chen (Chang and Chen, 2017) in 2015 where they predicted an 450% increase in AR and virtual reality (VR) users. They contributed this to the growth of the ecosystem of AR and VR that enabled more uses with these devices. Furthermore, the new devices come with improved sensors, which also improve the AR and VR experience. They also claim that this increase will mainly happen because the industry is interested in these technologies. For the consumer market, the authors claim that there is a chicken-egg-problem between manufacturers building devices and

^a  <https://orcid.org/0000-0001-9550-7418>

developers building applications. Either of them waits for the other to saturate the market which in turn stops the general progress in the consumer market. According to the authors the industry must step in to create a need for AR applications and devices such that progress can still be made.

Many AR approaches in industry fall into the teaching /supporting area, where they support the worker in his task or simplify teaching of new tasks. Mourtzis et al. (Mourtzis et al., 2018) simplifies the teaching of new designers by simplifying the designing process with the help of AR. Furthermore, Mourtzis et al. (Mourtzis et al., 2017) shows a remote support system where they improve the maintenance procedure. A similar approach is also described by Peng Wang et al. (Wang et al., 2019) where they describe how a remote expert can teach a local worker through spatial augmented reality (SAR) which projects information directly on the surfaces around the worker. A further solution is described by Volker Paelke (Paelke, 2014) where the author supports workers with visual aid in a simplified production environment that synchronizes the work of the machines and the workers.

Using AR as shown by Dimitris Mourtzis et al. (Mourtzis et al., 2018) improves the teaching of new designers and also helps professionals. The authors create an application which allows designers to view a product through a head mounted AR device. Designers can manipulate the product design with their gestures while it is shown in real time, with the product designs' size adjusted to its surroundings. The application helps new designers to rule out mistakes that would otherwise emerge later in the designing process. It also enables the designers to discuss their designs in group sessions, so it simplifies pointing out flaws to their colleagues. The proposed approach of the authors is working as follows:

1. The requirements of the new product are defined
2. A prototype is designed
3. The prototype is viewed in AR by the designing group
4. If a flaw is found the designers go back to 2)
5. If product is production ready the designers can proceed to the next production steps

Through the help of AR, the design process is made more accessible to newcomers, but also helps professionals by simplifying the design process in general. It shortens design time and designers become more aware which challenges they will face in the final assembly.

In a second paper the same authors show a maintenance-as-a-service approach (Mourtzis et al.,

2017). The authors create a system that lets onsite technicians get remote support for broken machines. After a malfunction report is filed to the cloud service by an onsite technician, the manufacturer generates an AR scene which is sent back to the onsite technician. The AR scene is mostly auto-generated and only little help is required from the manufacturer. Already existing scenes are reused, and a scene shows a step-by-step walk through of the repair. This enables the onsite technician to handle the repair without the need of a manufacturer to send experts to the site. Since the systems depends on a cloud infrastructure, the maintenance service is always accessible. The authors claim to reduce the maintenance cost by this approach and better integrate maintenance solutions into the product-as-a-service market.

Similar to the aforementioned solution, Peng Wang et al. (Wang et al., 2019) describe a mixed reality (MR) approach to teaching. Here, a remote expert shows in VR how to assemble a product to a local worker. The worker will apply this knowledge which is shown to him through spatial augmented reality, and he reproduces these steps to build the physical product. The authors also implemented a two-way feedback by sharing the video stream of the local worker to the remote expert, so the remote expert is able to modify or repeat steps if necessary. The goal of this research is to determine whether the remote expert is more comfortable using a controller or his hands / gestures to manipulate objects in AR. The authors conclude that a gesture-based system is generally more accepted than a controller-based system.

While these papers focus on knowledge transfer with the help of AR, there are solutions for helping the workers actively in their daily working routine as well. Volker Paelke shows an application which supports the worker in a simulated production environment (Paelke, 2014). His solution is integrated into a modular production system and offers the worker visual hints through a head-mounted-display. The tasks performed by the worker vary depending on the production steps of the machines ahead of the worker's station. These steps are considered when providing visual support to the worker. The worker then performs the required build task. The author compares different visualization methods, from text-only to visual cues, like concentric circles. While the author cannot establish which method of user guidance is best, he states the results to be very promising in general. The feedback received from the test subjects is very positive.

In the following paragraphs, this paper will describe an approach (similar like (Paelke, 2014)) for helping the worker through visual guidance in his

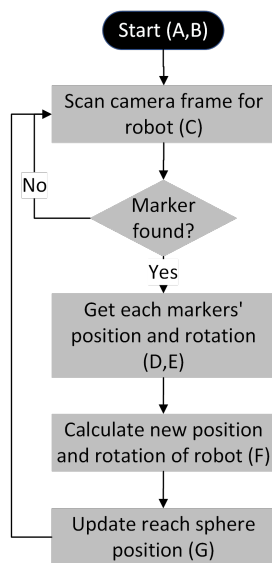


Figure 1: Flow chart of application.

work with, and around, a collaborative robot. Our approach shows the worker information for his task. This will help to optimize the worker's safety.

The next section explains how the robot is recognized from the AR device and how the AR application is created. The subsequent section explains the generation of the sphere representing the reach of the robot. Finally, this paper gives a conclusion and further considerations on future extension to the presented approach.

3 APPROACH

In this section, the specifics of the approach are presented and discussed in detail. Figure 1 shows the simplified sequence diagram of the AR application. As soon as the application is started, it is scanning the received camera image for markers. Once one or more markers are found, it calculates the rotation of these markers and infers the position of the robot base from them. Afterwards, it updates the position of the reach sphere accordingly.

3.1 Foundation

Unity3D is a game engine mainly used to create video games. Since Unity3D supports development of AR and VR applications, it is selected for the purpose of this paper. Unity3D supports building AR applications for IOS, Android and Microsoft HoloLens. An Android tablet (Samsung Galaxy Tab S6 Lite) and an Android smartphone (Samsung Galaxy S20) are used exemplarily throughout this paper.

3.2 Recognition Approaches

With the foundation set, the first step is to recognize the robot in the camera image of the tablet. Three recognition methods are evaluated for the robot recognition:

1. Recognition with artificial intelligence (AI)
2. Recognition through image recognition algorithms
3. Recognition with markers

The first method, an AI based approach, is ruled out because the necessary information for visualizing the reach of the robot cannot be extracted from the image. An artificial neural network must be trained thousands of epochs with millions of images, to be fitted to the purpose of recognizing the robot. The artificial neural network is then able to detect objects in the image and is providing either a bounding box around the object in the image or a segmentation pattern which only gives the pixels of the found object. Thus, information of the pose of the robot and the distance to the camera must be inferred by subsequent steps in the image processing pipeline. A single unified approach is subject to further research.

The second approach, recognizing the robot through image recognition algorithms, is also ruled out. Using recognition algorithms to detect salient features on the robot cover would recognize the robot, but there is also no further information about the robot pose and distance to the camera.

The third approach is the recognition with the help of markers (e.g., ArUCO markers). These are the work of Romero-Ramirez et al. (Romero-Ramirez et al., 2018) and Garrido-Jurado et al. (Garrido-Jurado et al., 2015). The authors show in their work how to generate and detect the markers with low effort and high precision. They claim to recognize the markers reliably in a 4k video stream with 1000 FPS even when the markers are printed on normal paper. For the recognition of the markers the OpenCV library (Bradski, 2000) is used with its ArUCO marker detection algorithms. This implementation is selected because of the high precision and speed that is also achieved on mobile devices.

3.3 Marker Recognition

OpenCV provides methods to recognize ArUCO markers in images. These methods are integrated into the Unity3D application as a dynamic-link library. The library is responsible for detecting markers in camera images and returning its encoded ID, its normal vector and the coordinates of the marker corners.

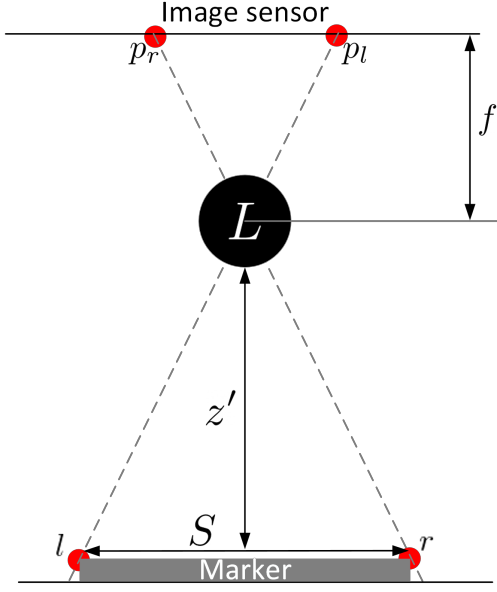


Figure 2: Conversion of the binocular disparity.

C#'s invoking and marshalling mechanisms are used to call the methods of the library.

The process of recognition of these markers works as follows: First, the image is turned to gray scale. Color is not necessary for this process. Then square structures are searched in the image. These are possible candidates for the markers, since the markers are squares as well. Once the candidates are found, they are further examined to figure out, if they contain an encoded ID and thus are valid markers. The valid markers are then returned, with their corresponding recognized data, to the Unity3D application.

3.4 Distance Estimation

OpenCV returns the (x, y) coordinates of the markers corners in the image. To add a 3D component to this information, the distance of the marker to the camera must be calculated. Most of the so-called 3D cameras use a similar approach to the binocular disparity, as described by Asra Aslam and Mohammad Samar Ansari (Aslam and Ansari, 2019). The authors describe how to calculate the distance to an object by taking two images from two cameras. These two cameras are placed with a fixed and known distance T from each other. Both cameras use the same focal length f and photograph the same scene. Now the pixels on each image are compared and the pixel-wise distance between pixels showing the same object is computed. Using equation (1) the distance is then calculated.

$$z = \frac{f \cdot T}{|x_l - x_r|} \quad (1)$$

A typical Android tablet only possesses one camera. To calculate the distance with one camera, the binocular disparity approach must be altered. Using one camera, the size of the object must be known. This is given, since the markers have a fixed and known size S . Also, the focal length f must be known. This value is provided by Unity3D. Calculating the pixel-offset between the provided corners of a marker delivers the third value necessary to calculate the distance with only one camera. This is also shown in figure 2. Using equation (2), it delivers the distance z' .

$$z' = \frac{f \cdot S}{|p_l - p_r|} \quad (2)$$

With this equation, the distance to a marker is calculated once a marker is found by OpenCV. The application now has the location of a marker in the image and also the distance to the marker from the camera. With these, the marker is located in world coordinates and an object is placed at its location in Unity3D's coordinate system. Since the markers placements on the robot is known, the position of the robot is extrapolated from the marker position.

As soon as a position of the robot is located, Unity3D keeps its relative position to the robots' surroundings. This is done by tracking the tablet movements with the help of gyro sensors and inertial measurement units. This allows the application to only update the robot position on a successful measurement, while still keeping the objects at the correct place when a worker is moving around.

3.4.1 Recognition Accuracy

The analysis of the accuracy of the distance measurement approach described above is shown in figure 3. It shows the deviation of the measured distance from the actual distance in percent. While the color strength indicates the horizontal angle to the marker, the different colors and line markers indicate the vertical angle to the marker. The stronger the color, the more perpendicular the angle to the marker. Measurements are taken from 25° , 50° , 75° and 90° for vertical and horizontal angles respectively. 90° corresponds to the camera being orthogonal to the marker, while 25° is the sharpest angle. The vertical angle corresponds to moving the AR device up and down, and the horizontal angle corresponds to left and right movement of the AR device. During the measurements the camera is pointed directly at the marker. The aim for this measurement is to cover one eighth of a sphere, as the results can be mirrored to the other seven eighths of the sphere. Every measurement is also carried out for six different distances. The distances tested range from 50 cm to 300 cm.

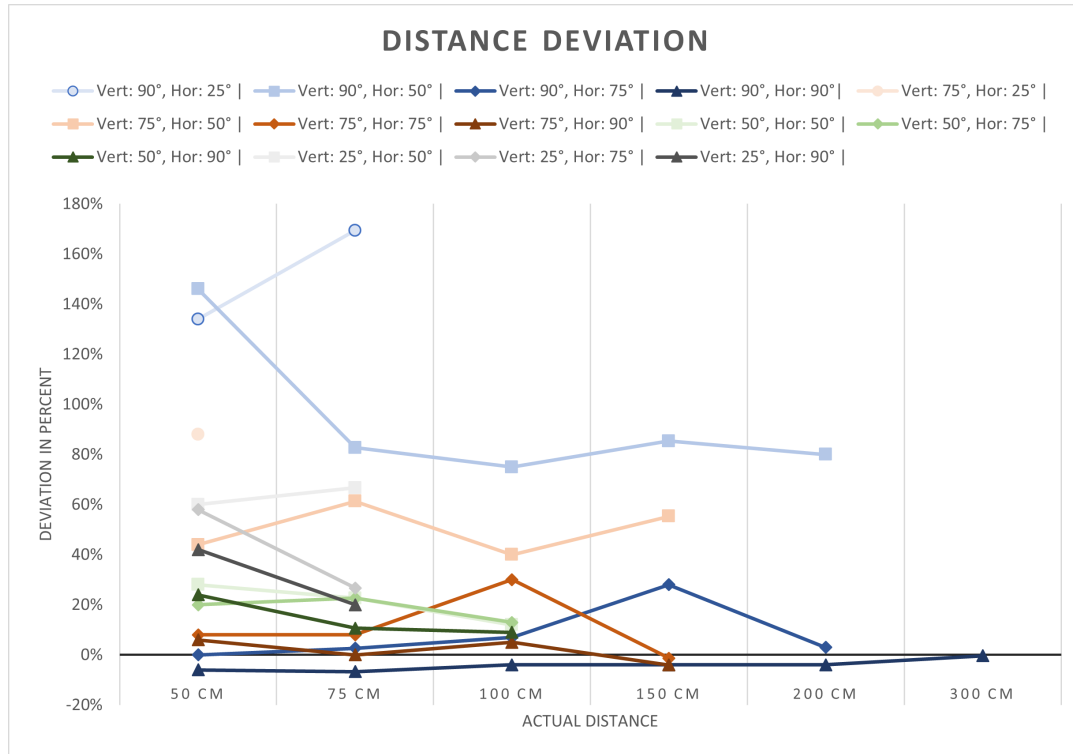


Figure 3: Deviation of the distance measured by the app in percent.

To estimate the distance, each position is tested three times and the average is taken. A test for one position is carried out by placing the AR device at the given angles to a marker fixed at a wall and with a given distance. The distance to the marker is measured directly between camera and marker, while the angles are measured from the wall to the AR device camera.

As expected, the figure shows that a perpendicular viewing angle of the marker works best for distance estimation, since it results in the smallest deviation for every tested distance. Decreasing the horizontal viewing angle increases the deviation. The accuracy loss is even bigger for decreasing the vertical angle.

This deviation happens due to a difference in the actual marker size and the projected marker size in the image. To correct this difference the viewing angle needs to be accounted for.

3.5 Viewing Angle Corrections

The projected marker is smaller in the image when the camera's normal vector $\vec{n}_c$ is not perpendicular to the marker. This happens due to rotation of the AR device. Under those circumstances the measurements become inaccurate when using the fixed marker size S .

The provided normal vector $\vec{n}_m$ of the marker

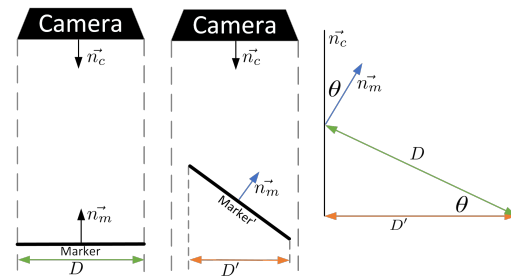


Figure 4: Marker changing size in image.

is used to counteract bad measurements on non-perpendicular angles of $\vec{n}_c$ to the marker. Knowing the relative rotation θ between $\vec{n}_c$ and $\vec{n}_m$ allows to calculate the projected size of the marker. Figure 4 shows the difference in size of the projected markers when the camera is rotated. Furthermore, it shows the correlation between the length of the projected markers and a right-angled triangle. This correlation allows to use the Pythagorean theorem in equation (3) to calculate the smaller projected marker length S' .

$$S' = \cos(\theta) \cdot S \quad (3)$$

Using S' in equation (2) will result in calculating the correct distance of a marker to the camera. Even if the marker is rotated and the camera forward vector $\vec{n}_c$ is not orthogonal to the marker.

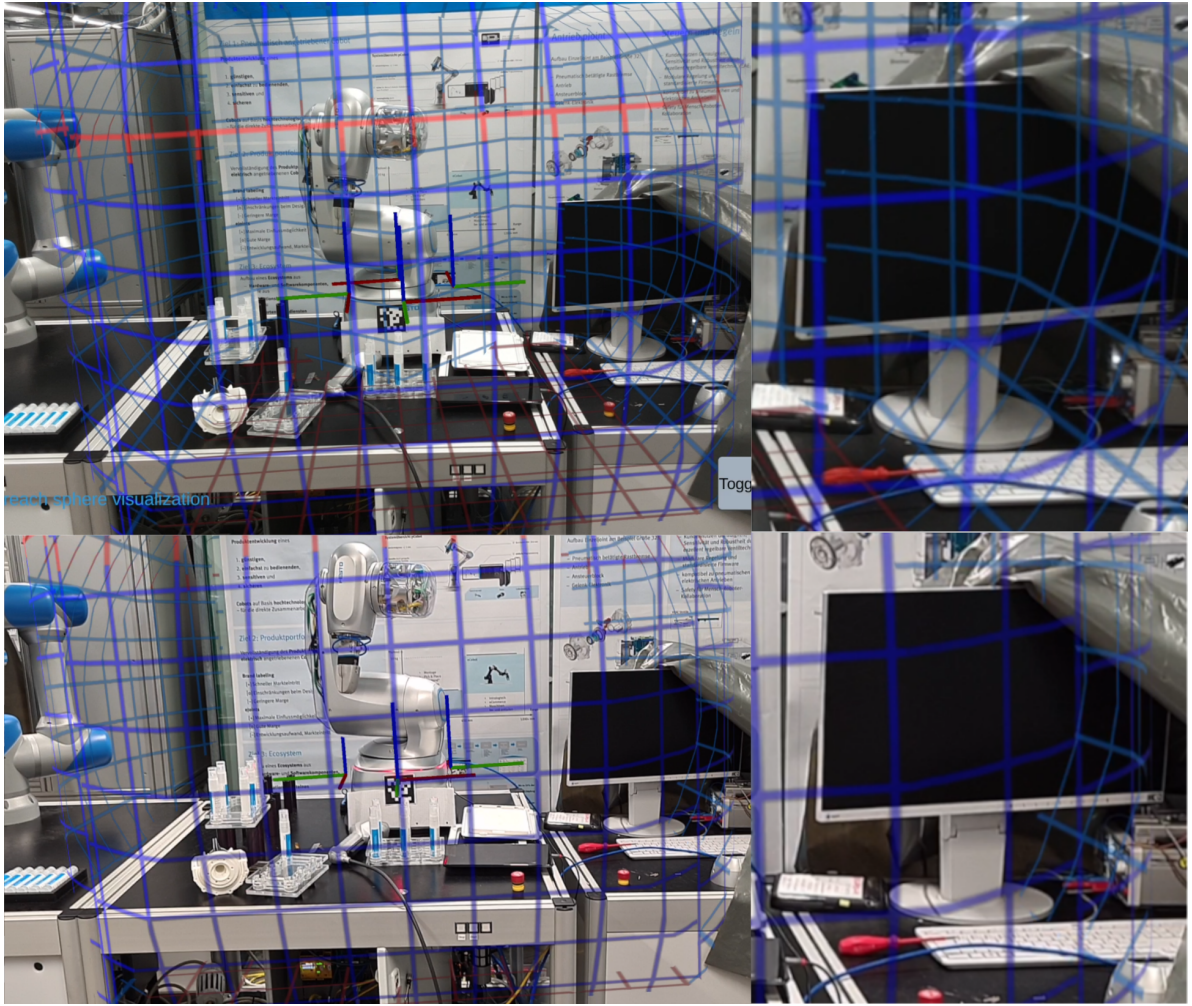


Figure 5: Comparison of the result without (top part) and with (bottom part) Depth API.

3.6 Exponentially Weighted Moving Average

To give the worker a smooth experience, larger differences in between frames need to be evened out. OpenCV delivers the rotation of the marker, which is not always the same, for consecutive frames without moving the camera. This results in a shaky movement of the virtual object representing the reach of the robot. To get improved results, an exponentially weighted moving average (EWMA) is used. By weighing the newly calculated position and rotation values against the already provided ones, the shaky behaviour gets smoothed out. How to calculate the EWMA is shown in equation (4).

$$EWMA_t = \alpha \cdot r_t + (1 - \alpha) \cdot EWMA_{t-1} \quad (4)$$

The EWMA is a recursive function which is usually used in finance. The α is the weight with which each old value is reduced before adding it to the average.

By using this equation for location and rotation values, the application filters out single bad measurements, but is still able to adapt to changes that persist over a longer period of time by overriding the old values with each new measurement and thus changing the average to be at the changed value.

3.7 Visualize the Reach

To visualize the reach of the robot, the reach must be known. Therefore, the robot is simulated to reach every possible position and the outer most position of the tool head is marked. The created point cloud gives an outer limit as to what the robot can reach. With this point cloud, a 3D object can be modeled which is representative of the robot reach and shown in figure 6. This model is loaded into Unity3D. Whenever the robot is recognized successfully, this model is positioned around the robot base and shows the user the

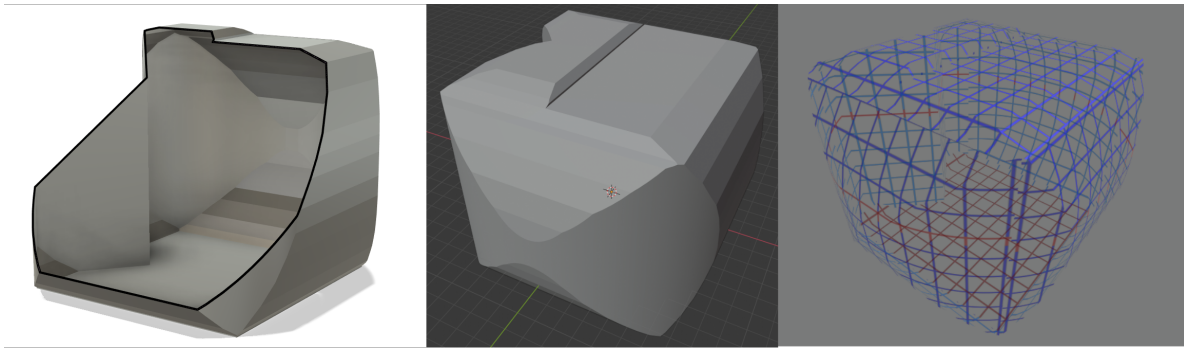


Figure 6: On the left a section analysis of the sphere object. In the middle single color approach to displaying the sphere. On the right a mesh approach to displaying the sphere.

maximum reach.

The first approach to show the reach sphere in a single color (figure 6 mid part, though the approach was done with the sphere being transparent) was discarded as it was not possible to understand where the border of the reach of the robot is. Thus, in the chosen approach, a mesh texture is displayed on the sphere (figure 6 right part). This requires the sphere object to be hollow, to also render a texture on the inside of the sphere (figure 6 left part). The coloring shows the worker from which side the sphere is viewed. Darker blue for the outside and lighter blue for the inside of the sphere. This enables the worker to distinguish the reach borders and improves understanding of the reach of the robot. Furthermore, a red scan line is added. This line moves up and down on the sphere. It is used as a moving element that helps to better differentiate between the outside and inside of the sphere.

4 OCCLUSION DETECTION

Occlusion detection enables the application to detect if virtual objects are hidden behind real world objects. In doing so it allows the application to disable rendering for hidden virtual objects. Unity3D supports occlusion detection by using device dependent APIs for depth estimation.

Depth estimation is enabled when a device has appropriate sensors (iPad: Lidar sensor) or powerful enough hardware to create a depth map through image analysis in conjunction with motion sensor input (modern Android devices (Google, 2022)). Once, Unity3D detects that depth mapping is available, it will automatically toggle the occlusion detection feature. This causes the application to pay attention to the generated depth map for each frame and render the virtual objects accordingly. If a virtual object is hidden behind a real world object then only its visible part is rendered.

Figure 5 is a comparison between what the worker sees with occlusion detection disabled vs. enabled. The top row shows the application with occlusion detection disabled, while the bottom row shows the application with occlusion detection enabled. The zoomed in view on the right enlarges the monitor of the image on the left. In the top, it is visible that the sphere is rendered even though it is overlapping with the monitor. Since occlusion detection is disabled, the application does not know that the monitor is colliding with the sphere and thus is not hiding the sphere object. With occlusion detection enabled in the bottom row, the monitor hides the sphere where it is colliding with the monitor. The effect of partially rendering the sphere improves the immersion of the worker, since this is the anticipated effect. It also improves understanding of the spheres' position in relation to the real world objects.

Occlusion detection improves immersion as it enables the application to integrate the virtual objects better into the scene with respect to the real world environment.

5 CONCLUSIONS

Visualizing the reach of a robot as proposed in this paper not only improves the set-up procedure of a new robot, but also helps the worker in his daily tasks with, and around, the robot.

Visualizing the robot reach during set-up simplifies the positioning of the robot. The robot needs to reach all of its tools and work pieces, and by enabling the setup worker to see the reach, it is possible to configure the robot position without powering the robot up. This also speeds up the set-up process by reducing the number of reach test that need to be done to confirm if the robot is able to reach everything he needs to reach.

Furthermore, the application also supports the

worker while working around the robot. It enables the worker to understand what the robot can reach. This understanding also improves the safety of the worker. The worker is able to see where the robot can reach and thus reduce accidents when working with the robot.

The quantity of these results are subject of further research.

Further improvements to the application to increase the support for the worker could be imagined. The depth map could be used to also show areas where the robot is possibly colliding with objects inside its reach sphere. The visualization of these areas could improve the programming process of the robot by showing the worker which areas to avoid during program execution. The application could also show the worker information for the ports on the backside of the robot. This will improve understanding of the ports and their corresponding functions.

REFERENCES

- Aslam, A. and Ansari, M. S. (2019). Depth-map generation using pixel matching in stereoscopic pair of images. *CoRR*, abs/1902.03471.
- Bradski, G. (2000). The OpenCV Library. *Dr. Dobbs's Journal of Software Tools*.
- Chang, S.-N. and Chen, W.-L. (2017). Does visualize industries matter? a technology foresight of global virtual reality and augmented reality industry. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 382–385.
- Galin, R. and Meshcheryakov, R. (2019). Automation and robotics in the context of industry 4.0: the shift to collaborative robots. *IOP Conference Series: Materials Science and Engineering*, 537(3):032073.
- Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F., and Medina-Carnicer, R. (2015). Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern Recognition*, 51.
- Google (2022). Overview of arcore and supported development environments. <https://developers.google.com/ar/develop?hl=en>. Online; accessed 28 July 2022.
- Kopp, T., Baumgartner, M., and Kinkel, S. (2021). Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework. *The International Journal of Advanced Manufacturing Technology*, 112.
- Mourtzis, D., Zogopoulos, V., and Vlachou, E. (2017). Augmented reality application to support remote maintenance as a service in the robotics industry. *Procedia CIRP*, 63:46–51. Manufacturing Systems 4.0 – Proceedings of the 50th CIRP Conference on Manufacturing Systems.
- Mourtzis, D., Zogopoulos, V., and Vlachou, E. (2018). Augmented reality supported product design towards industry 4.0: a teaching factory paradigm. *Procedia Manufacturing*, 23:207–212. “Advanced Engineering Education & Training for Manufacturing Innovation”8th CIRP Sponsored Conference on Learning Factories (CLF 2018).
- Paelke, V. (2014). Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment. In *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pages 1–4.
- Romero-Ramirez, F., Muñoz-Salinas, R., and Medina-Carnicer, R. (2018). Speeded up detection of squared fiducial markers. *Image and Vision Computing*, 76.
- Wang, P., Bai, X., Billinghamurst, M., Zhang, S., Han, D., Lv, H., He, W., Yan, Y., Zhang, X., and Min, H. (2019). An mr remote collaborative platform based on 3d cad models for training in industry. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 91–92.

SHORT PAPERS

Usability Evaluation of a Community-led Innovation Mobile App

André Castello Branco¹^a, Eveline Sacramento²^b, Eliza Oliveira¹^c, Oksana Tymoshchuk¹^d,
Maria Antunes¹^e, Margarida Almeida¹^f, Luis Pedro¹^g, Fernando Ramos¹^h
and Daniel Carvalho¹ⁱ

¹Department of Communication and Arts, University of Aveiro, DigiMedia, Aveiro, Portugal

²Department of Communication and Arts, University of Aveiro, INET-md, Institute of Ethnomusicology, Aveiro, Portugal
{acastello, eveline.sacramento, elizaoliveira, oksana, mariajoao, marga, lpedro, fernando.amos, daniel.carvalho}@ua.pt

Keywords: Mobile Application, Usability, Heuristic Evaluation, Community-led Innovation.

Abstract: Digital media can facilitate collaborative processes among local agents, value endogenous resources, and promote assets associated with territory. This article presents the results of a study concerning the development and validation of a mobile app for promoting the relationship among agents of the Portuguese Centro region's communities/entities. This paper focuses on the results of a heuristic evaluation of the mobile app carried out with two groups of experts in Digital Technologies, Tourism, Health, and Well-Being, besides providing an overview of the mobile app that was developed and a theoretical background regarding community-led innovation, usability, and heuristics. For the CeNTER app prototype evaluation itself, the use of Nielsen's heuristics, a MATCH-MED scale, together with a Think-Aloud Protocol allowed us to improve its usability. This article contributes to a reflection about the evaluation of mobile apps in the scope of territorial-based innovation initiatives, engaging its stakeholders in the process.

1 INTRODUCTION

Community-led initiatives have great potential in the development of cultural, tourism, and environmental projects, enabling valuing endogenous resources and promoting innovation in the territory. Digital technologies can be particularly useful in community-led initiatives, allowing to recreate a "virtual proximity" between the different actors involved in the territory's development process (Martínez-Rolán et al., 2019). They can offer several advantages to community-led initiatives, as, for example, facilitating innovation, because of access to information and experiences; sharing practices, leading to knowledge co-creation and the emergence

of new ideas; and enabling their members to improve their practices through a continuous engagement in a meaningful participatory environment (Saint-Onge & Wallace, 2012; Snyder & Wenger, 2010).

However, several studies indicate that community-led initiatives face several challenges, especially in rural areas (Brown & Nylander, 2009; Marré & Weber, 2010). These challenges are related to the fact that local participants, namely elderly people, have difficulties accessing the Internet, present limited digital skills, and lack access to technological equipment. To overcome these challenges, new political measures are required to empower local communities to the social, cultural, and economic valorisation of territories.

^a <https://orcid.org/0000-0002-6493-6938>

^b <https://orcid.org/0000-0003-0839-4537>

^c <https://orcid.org/0000-0002-3518-3447>

^d <https://orcid.org/0000-0001-8054-8014>

^e <https://orcid.org/0000-0002-7819-4103>

^f <https://orcid.org/0000-0002-7349-457X>

^g <https://orcid.org/0000-0003-1763-8433>

^h <https://orcid.org/0000-0003-3405-6953>

ⁱ <https://orcid.org/0000-0003-0108-8887>

This study was developed under the CeENTER Program, which has as one of its objectives to understand the active role those digital technologies can play in the process of territorial innovation (Silva et al., 2021). In this context, a digital platform (mobile app) is being designed, whose primary focus is to promote collaboration between the various agents involved in territorial-based innovation processes. The CeENTER mobile app intends to be a tool that encourages internal and external interactions and facilitates communication and collaboration processes, to strengthen existing mediation strategies and stimulate the creation of new ones, as well as new ideas/activities (Tymoshchuk et al., 2019). However, the design of a mobile app for community mediation is still a great challenge due to the variety of initiatives and the diversity of stakeholders (public institutions, companies, and communities) involved in territorial-based innovation processes.

This paper presents the results of a heuristic evaluation of the CeENTER app prototype, which will provide the development cycle with crucial usability data, required to converge to a final solution that will offer a coherent and effective experience about the distinguishing features of this app to its users. The study is supported by a User-Centered Design approach, which focuses on satisfying the user desires and needs (Hartson & Pyla, 2012). Also, it is framed by the User Experience (UX) theoretical basis, guiding the prototype evaluation in the CeENTER Program scope, which will crucial the development of the CeENTER mobile app.

2 BACKGROUND

2.1 Territorial Innovation

Innovation can be translated as the systematization of knowledge shared between actors in the spaces in new products and services. Systematized knowledge, when interacting with the social, historical, and cultural constructs for which they are intended, change the behaviour of social groups in a systemic way, enabling the development by promoting territorial innovation (Ferreira, 2020). In geographically close groups, these historical-socio-cultural constructs are perceived in a systemic way, which characterizes the territories (Keating, 2020). Territorial innovation is a complex process that results from an intentional collective action promoted by a set of actors and organizations (companies, universities, local government bodies, communities) that interact to develop innovation (Morgan, 1997).

So, the initiatives led by groups coordinated from the bottom up respond to the development needs of territories, as they address ecological, social, economic, and political problems of global resonance (Seyfang & Smith, 2007, p. 585).

Furthermore, the sharing of information to solve local problems has become more effective as a result of the integration provided by the dissemination of information and telecommunication technologies, especially mobile apps (Diniz et al., 2019), further enhancing the importance of mobile apps in the territorial-based innovation domain. These digital solutions must meet the community's needs to foster the local economy. To ensure that a product meets users' needs, such mobile digital solutions must be developed under the User Experience (UX) theoretical basis, which is crucial to gather relevant information regarding the interaction with the product.

2.2 User Experience

An experience is a complex event, which is created in the mind of the user and is influenced by many factors, being a completely personal issue (Knight, 2019). Therefore, User experience is subjective and is related to how the user feels regarding a created product. Further, the current concept of UX is strictly related to emotional results that emerged from the user when interacting with a product, such as pride, joy, and fun (Bernhaupt & Pirker, 2013). Therefore, creating an experience is not just about how the product is designed, which structures were implemented or whether state-of-the-art technology is used (Knight, 2019). It is also about how the specific product can help the user to accomplish their tasks successfully and how the user feels when getting involved with the product.

UX is composed of three factors: usability, usefulness, and emotional impact, and generates a memory related to the product interaction. Emotional impact is the affective component of the user experience, focusing on the system as a means that affects the user's feelings. Usefulness focuses on the use of a system to achieve goals and accomplish specific tasks. Usability is constituted by effectiveness, efficiency, ease-of-use, learnability, and the user satisfaction, being the practical component of user experience (Hartson & Pyla, 2012).

The usability factor contributes to the elaboration of a product that is easy to use and is pivotal in creating digital products since it provides products with a low level of difficulty, reducing the platform's

disuse (Partridge, 2017). Since ensuring high usability for a mobile app is of great importance, as a predictor of its acceptance and of its success, is crucial in all phases of app testing (Muchagata & Ferreira, 2019), including early testing with experts.

2.3 Usability

Usability is an essential concept in developing digital interfaces, focusing on users and contexts of use, and ensuring they can achieve their goals with efficiency, effectiveness, and satisfaction. The usability concept was defined by Nielsen (2012) as an attribute with qualitative nature that assesses the ease of use of user interfaces. Usability directly relates to methods for enhancing this attribute during the design process phase that includes the following five main components: (1) the ability to learn - the ease with which users complete basic tasks during the primary interaction with the interface; (2) efficiency - the speed with which users reach their goals after they have learned to interact with the platform; (3) memory - the easiness with which users interact with the interface after a period without using it; (4) errors - the number of errors that are made by the user and the ease with which it is possible to correct them; (5) satisfaction - a component related to the desire and pleasure expressed by the user during and after the interaction with a specific product (Nielsen, 2012).

Due to several specificities related to the small screen, storage capacity, and energy consumption of mobile phones, it is essential that these apps meet a set of basic requirements, such as: being easy to use, being flexible, having a simple and intuitive interface, allowing the user to easily adapt to different contexts of usage, etc. (Feijó et al., 2013; Kumar & Mohite, 2018).

In this sense, usability evaluation of mobile apps is a mandatory process to ensure that such apps are practical, effective, and easy to use (Kumar & Mohite, 2018). Usability evaluation is the generic designation for a set of methods aiming at interface inspection. The final objective is to identify usability issues, through the indication of the severities' level of each issue (Nielsen, 1994). These methods include heuristic evaluation, cognitive pathway, consistency, pattern inspection, among others. One of the commonly available and employed methods for assessing and improving interfaces is heuristic evaluation (Nielsen, 1994).

2.4 Heuristic Evaluation

Nielsen's heuristics (1994) is a systematic evaluation method allowing the identification of problems in user interface design that involves interface analysis. As the author states, heuristic evaluation requires fewer resources than other methods to detect usability issues. As Nielsen (1994) states, "basically, a set of evaluators inspects the interface with respect to a small set of fairly broad principles, which are referred to as the 'heuristics'" (p. 152).

This specific method is based on a 10 "heuristic" items checklist, which can be used in interface specifications, prototypes, or complete systems. Nielsen's 10 heuristics include the following guidelines (Nielsen, 1994): (1) Visibility of the system; (2) Match between system and real-world; (3) User control and freedom; (4) Consistency and standards; (5) Error prevention; (6) Minimise the user's memory load; (7) Efficiency of use and performance; (8) Aesthetic and minimalist design; (9) Help and documentation; (10) Help users identify, analyse, and improve from errors.

According to Nielsen and Landauer (1993), between three to five evaluators are recommended to perform a heuristic evaluation, which can identify around 75% to 95% of the problems, considering that an individual evaluator usually finds around 35% of them.

2.4.1 MATCH-MED Scale

Testing the usability of mobile apps can be done using several sets of checklists and protocols available to carry out heuristic evaluations. However, the vast majority of these scales are geared towards generic systems, demanding adaptations to fulfil specific requirements of mobile apps (Hashim & Isse, 2019; Inostroza et al., 2016). Considering that the CeNER application is aimed at Tourism, Health, and Well-being clusters, a set of MATCH-MED usability heuristics was used.

The MATCH-MED scale aims to evaluate the usability of mHealth systems on smartphones and was developed based on Nielsen's generic heuristics, with the addition of three heuristics specific for mobile devices (Lacerda et al., 2015; Salazar et al., 2012), which consist of: (1) Minimization of human-computer interaction: Considering that typing on mobile touch screen keyboards is more error-prone than on conventional keyboards, it is essential to minimise user interaction with the mobile app; (2) Physical interaction and ergonomics: Given the limited screen size of a mobile device, the action

controls must present suitable sizes and minimal distance from each other, ensuring the user does not press a button by mistake; (3) Readability and Quick View: Considering that mobile apps are generally used in dynamic contexts, it is essential to ensure quick access to system information by the user (e.g., at a glance).

According to Zhang and Adipat (2005), a comprehensive usability study for a mobile application should assess a variety of issues such as interface design, ease of use, and perceived attitude by the user and measures related to the application's performance.

3 THE CENTER PROTOTYPE

The team researchers developed a CeNTER mobile application prototype based on an in-depth study under this project. That research included a systematic review of the literature, a benchmarking of the communities' digital platforms, a series of interviews with local stakeholders, and two focus groups with representatives of community-led initiatives (Oliveira et al., 2021). The results of these methodological procedures made it possible to identify the main difficulties found on community-led initiatives and led to the proposed digital solution that aimed to fill these gaps.

The researchers chose a card-based user interface design that is both simple and innovative in appearance. This type of interface design tries to provide information in a readable format, easy to navigate, allowing an overview of the application's content and quick access to all categories (Seifi, 2015). Considering the requirements and functionalities defined as important in the proposed digital solution, the medium-fidelity prototype was developed with the support of "Principle" software, an interactive user interface, and animation design software.

The prototype's content evaluated by experts was organised according to the categories defined by the team as being essential in the mobile application: Initiatives; Events; Entities; Volunteers; Resources and Highlights. These categories were defined during the Focus Group sessions with the local initiatives, through a User-Centered Design approach. In the CeNTER app context, these categories are called "Tabs" and represent the main relevant elements in the community dynamics. Such "Tabs" are organised in horizontal lines on the screen and have secondary content that is modified according to the preferences

defined by the user in the app settings (Branco et al., 2021).

By touching each one of the Tabs, a vertical opening occurs on the screen, where it is possible to view suggestions (of local Initiatives, for example) through cards displayed on the screen. Each card presents a single content and appears ordered in a carousel layout.

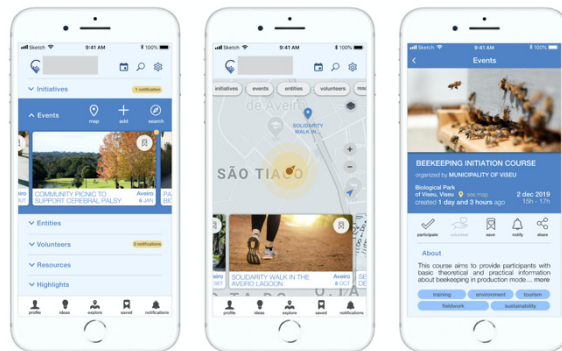


Figure 1: Home screen of the version of the medium-fidelity prototype "CeNTER" evaluated by experts.

4 METHODOLOGY

The evaluation process for the CeNTER mobile application prototype is outlined according to an iterative design methodology (Lorenz et al., 2010). It provides for the initial testing of app mockups with team members, evaluating medium-fidelity prototypes by experts and laboratory tests with end-users. This article presents the second stage of evaluation of the prototype, which consisted of heuristic evaluation by experts to detect and correct issues related to the prototype's usability.

Through dissemination seminars, organized by the CeNTER team, it was possible to recruit local stakeholders to actively participate in the project as experts. To accomplish the assessment, two panels of experts were organized according to the evaluators' expertise. In this sense, the first group consisted of five experts in the field of digital technologies, having deep knowledge and experience in developing interfaces. The second group consisted of five experts in the fields of Tourism, Health, and Well-being, who have knowledge of their specific domain and involvement in local community projects.

To determine the severity of usability issues for this study, a heuristic checklist was developed to evaluate the application prototype, based mainly on Nielsen's 10 heuristics (1994). Besides that, three heuristics of MATCH-MED scale (Salgado et al.,

2019) were added to allow the evaluation of mobile apps in health and well-being areas. It is important to note that the researchers chose to add these items on the MATCH-MED scale since Nielsen's heuristics do not consider specific characteristics and limitations of mobile devices (Hashim & Isse, 2019), as it was pointed out before.

A total of 44 items were identified in this checklist, which was employed to evaluate the prototype usability. The heuristic evaluation process took place in three stages, according to the proposal of Nielsen (1994): (i) preparation phase, in which the prototype screens for evaluation and the list of heuristics were defined and organised; (ii) evaluation phase, which consisted of collecting data from each expert, individually. The evaluators tested the prototype by identifying the guidelines that were violated and the degree of severity of the problem.

In addition, the Think-Aloud Protocol (Jaspers, 2009) was also employed to obtain immediate feedback from experts about their experience of interacting with the prototype. The application of this method allowed the qualitative evaluation of the prototype based on the experts' verbal comments.

The tests took place at the facilities of the University of Aveiro, with the first group on October 28-31, 2019, and the second group on November 11-29, 2019. In total, 78 screens of the developed application prototype were evaluated. This evaluation consisted of two phases: in the first phase, the experts freely explored the prototype and commented on their doubts about the CeENTER program, and in the second, the experts filled a table of design-oriented heuristics for mobile phone interfaces and apps. A User-Centered Design approach was used as an interactive design process to get users feedbacks and their needs in each phase of design evaluation process.

5 RESULTS AND DISCUSSION

This section presents the results obtained from the usability test, through a heuristic evaluation, of the CeENTER application prototype with the two groups of experts (five users in each group). To consolidate the heuristic validation results performed by both groups, a grid was created for each group to gather all of the obtained results.

A total of 174 usability problems were identified. After eliminating the duplicate issues, 155 unique usability problems remained. Table 1 shows the number of problems identified by each group of

evaluators, the severity of such problems and their average severity.

Regarding the severity of problems, it is important to refer that Group 1 (G1) identified 46 (42,2%) cosmetic problems, 39 (35,8%) small usability problems, 21 (19,3%) main usability problems and 2 (1,8%) usability "catastrophes". The group of digital technologies experts identified two problems with score level four. These problems referred to the second heuristic: "Correspondence between the system and the real world" and were related to the following items: "The proposed interactions in the application are similar to real actions"; and "Information appears in a logical and natural order". Those are relevant feedback because the prototype is applied to promoting community-led initiatives and those participants may present different levels of digital literacies. In this sense, a simple and intuitive interface, easily adaptable to different contexts of usage (Feijó et al., 2013; Kumar & Mohite, 2018) is important to make sure individuals use the application properly, which in this case is represented by different territorial agents (communities, entities, networks and citizens).

Table 1: Number of Problems and Average Severities identified by each group of evaluators.

Group	Experts	Total problems	Average severity
G1	Digital Technologies Experts	109	1,78
G2	Tourism, Health and Well-being Experts	46	1,53

Regarding the severity of the found problems, it is important to note that Group 2 identified 14 (30,4%) cosmetic problems, 26 (56,5%) small usability problems, and 6 (13%) main usability problems. It should be noted that no problems with a score of level 4 (usability catastrophes) were reported by the second group.

Tables 2 and 3 show the type of usability problems and average severities identified by each group of evaluators. With the highest Average Severity, Group 1 identified the heuristics: Help and Documentation (2.75), Recognition rather than Reminder (2.5), and Interaction between person and application (2.33). Group 2 reported the following Heuristic Severities Averages: Flexibility and Efficiency (2.25), Help and Documentation (2.2) Recognition rather than Reminder (2), and Interaction between person and application (2). Therefore, both

groups consistently assessed the severity of the heuristic violation. The Mean Average Severity in these two groups was low, with a score close to 1.78 (G1) and 1.53 (G2).

The difference in results among both groups also occurred in the type of the identified problems. The problems identified by Group 2 were more related to the suitability of the prototype for each application domain, as well as the efficiency and type of functionality available to users (Table 3). For example, experts of Group 2 did not observe inconsistencies rectified after the first round of tests, such as different icons representing the same function, or the absence of the return icon on some screens.

Table 2: Type of usability problems and average severities identified by Group 1 of evaluators.

Heuristic	G 1 - Digital Technologies Experts	
	Frequency N=109	Average of Severities
Visibility of system status	11	2.18
Match between system and the real world	11	2.27
User control and freedom	13	1.54
Consistency and standards	7	1.57
Error prevention	5	1.6
Recognition rather than recall	4	2.5
Flexibility and efficiency of use	7	1.85
Aesthetic and minimalist design	12	1.66
Help users recognize, diagnose, and recover from errors	10	1.1
Help and documentation	4	2.75
Interaction between person and application	3	2.33
Physical interaction and ergonomics	8	1.75
Readability and layout	14	1.5

As it is possible to see in Table 2, the maximum number of usability problems, identified by Group 1, included: Readability and layout (14, mean severity 1.5), followed by User controls and free will (13, mean severity 1.54) and Aesthetics and minimalist design (12, medium severity 1.66). In Table 3, Group 2 identified the following number of heuristic violations: User controls and exercises free will (11, mean severity 1.45), Avoid errors (7, mean severity 1.86), and Matching the system to the real world (6, mean severity 1.5). Identifying these heuristic violations will make it easier to identify and prioritise

issues that need urgent attention before the final deployment of the application.

The application of the three heuristics proposed by the MATCH-MED scale allowed the identification of 28 usability problems related to the different specificities of mobile devices, namely "Readability and layout" (16) and "Physical interaction and ergonomics" (8).

Table 3: Type of usability problems and average severities identified by Group 2 of evaluators.

Heuristic	G 2 - Tourism, Health, and Well-being Experts	
	Frequency N=46	Average of Severities
Visibility of system status	1	1
Match between system and the real world	6	1.5
User control and freedom	11	1.45
Consistency and standards	3	1.66
Error prevention	7	1.86
Recognition rather than recall	1	2
Flexibility and efficiency of use	4	2.25
Aesthetic and minimalist design	1	1
Help users recognize, diagnose, and recover from errors	4	1.9
Help and documentation	5	2.2
Interaction between person and application	1	2
Physical interaction and ergonomics	0	0
Readability and layout	2	1

The Think-Aloud Protocol application allowed the collection of 124 suggestions for improving the prototype under evaluation, which will be considered in the following stages of the mobile application development. The first group of experts made 83 suggestions for improvement, which were very focused on technical aspects and improving interactions. For example, "standardise the 'save' icons on the right side of the ideas screen with the rest of the application" (G1E1) and "Drag and drop visual feedback is required" (G1E2), respectively, for the evaluators E1 and E2 of Group 1.

In contrast, the second group of experts mainly commented on the concept of the CeNER application, functionalities, and effectiveness and did not mention technical aspects. This group made several suggestions regarding the lexical inaccuracies

used in the application. It is important to note that both experts' groups highlighted the "Resources" and "Volunteer" components as being the most innovative features of this application: "Resources and volunteers are the differentiating elements of the platform" (G2E5). Overall, the prototype was described as having a simple interface, with an appropriate layout, easy to use and not requiring much effort from the user.

The study's results demonstrate that usability testing is an effective way to significantly improve the interface of a future mobile app, favouring the user experience, as mentioned by other authors (Lacerda et al., 2016; Salazar et al., 2012). The methods and techniques of application production were supported by methodologies oriented to the user's requirements and were anchored in a spiral of evaluation of the prototype, resulting in improved versions of the pilot app, attending to users' statements, and developing an application with a more attractive look & feel.

6 CONCLUSIONS

As the development of the CeNTER prototype requires an iterative design process, performing early tests is a crucial part of the system design to detect usability problems and make essential improvements. Through the techniques used for the presented evaluation, it was possible to achieve value data, as the experts are from different areas of knowledge.

Therefore, according to the current study findings, the identification of usability problems facilitated the identification and prioritisation of problems that need urgent attention before the final implementation of the application. In this case, the heuristic "control and freedom" deserves special attention, since it received the major severity score by both groups. Furthermore, a decrease in the heuristic values between the two groups was verified, which leads to the conclusion that the changes made after the first round of tests helped for the improvement of the prototype.

This study also showed the importance of developing heuristic usability evaluation scales. As already mentioned, the researchers applied three items proposed by the MATCH-MED scale, which allowed us to identify usability problems related to the specificities of mobile devices. The evaluation of the prototype, not only by professionals in the field of digital technologies but also by professionals in areas related to the application, such as tourism, health, and well-being is also important. It is essential to mention that the professionals in the areas of tourism, health,

and well-being, who participated in this study, have a strong connection with different community initiatives and deeply know the local associative dynamics, which was very advantageous for evaluating the prototype.

Future work includes performing laboratory tests with end-users, which are of most importance in this project, requiring accommodation of the experts' suggestions for improvements. Some limitations of this work include the fact that Principle software does not allow to test some features and interactions of the prototype, such as writing on the interface and changing the letters' size for accessibility. Besides, the Principle software only works with iPhone, which limits the evaluation using Android devices.

Specifically, it is hoped that this app can make a difference in everyday life, promoting community-led initiatives, towards adding value to local resources and, therefore, the increment of the region's social, cultural, and economic levels.

ACKNOWLEDGEMENTS

This work is financially supported by national funds through FCT – Foundation for Science and Technology, I.P., under the project UIDB/05460/2020

REFERENCES

- Bernhaupt, M. & Pirker, R. (2013). Evaluating User Experience for Interactive Television: Towards the Development of a Domain-Specific User Experience Questionnaire, *Human-computer Interact. - Lect. Notes Comput. Sci.* 8118, 642–659.
- Brown, R. & Nylander, A. (2009). Community Leadership Structure: Differences Between Rural Community Leaders' and Residents' Informational Networks. *Journal of the Community Development Society*, 29, p. 71-89. <https://doi.org/10.1080/15575339809489774>
- Branco, A.C., Carvalho, D., Sacramento, E., Tymoshchuk, O., Oliveira, E., Antunes, M.J., Pedro, L., Almeida, M., Ramos, F. (2021). Prototyping and evaluating a mobile app to promote territorial innovation. In *Proceedings of CeNTER Conference*, (pp. 73-81).
- Diniz, E., Cernev, A., Bastos, T. P., Pinheiro, A., La Rovere, R., Querette, E., Morel, L., & Arroio, A. (2019). *Inovação em Serviços na Economia do Compartilhamento*. Publisher: Saraiva.
- Feijó, V. C., Gonçalves, B. S. & Gomez, L. S. (2013). Heurística para avaliação de usabilidade em interfaces de aplicativos smartphones: Utilidade, produtividade e imersão. *Design e Tecnologia*, 3(06), 33-42.

- Ferreira, L.M. (2020). Sustainable generation of innovative enterprises in territories: a proposal of an ecosystem management model. In *Proceedings of CeNTER*, 55-62.
- Hartson, R. & Pyla, P. S. (2012). *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.
- Hashim, N. & Isse, A. (2019). Usability Evaluation Metrics of Tourism Mobile Ap-plications. *Journal of Software Engineering and Applications*, 12, 267-277. <https://doi.org/10.4236/jsea.2019.127016>
- Inostroza, R., Rusu, C., Roncagliolo, S., Rusu, V., & Collazos, C. A. (2016). Developing SMASH: A set of SMArtphone's uSability Heuristics. *Computer Standards & Interfaces*, 43, 40-52.
- Jaspers, M. (2009). A comparison of usability methods for testing interactive health technologies: methodological aspects and empirical evidence. *International journal of medical informatics*, 78(5), 340-353.
- Keating, M. (2020) Beyond the nation-state: territory, solidarity, and welfare in a multiscalar Europe. *Territory, Politics, Governance*, 9, 2021 - Issue 3: Inequalities, territorial politics, nationalism.
- Knight, W., 2019. UX for Developers. Northampton. APRESS.
- Kumar, B. A. & Mohite, P. (2018). Usability of mobile learning applications: a systematic literature review. *Journal of Computers in Education*, 5(1), 1-17.
- Lacerda, T. C., Nunes, J. V. & von Wangenheim, C. G. (2015). Usability heuristics for mobile phone applications: a literature review. In *Emerging Perspectives on the De-sign, Use, and Evaluation of Mobile and Handheld Devices* (pp. 143-157). IGI Global.
- Lorenz, T., Richter, A., Köster, F. & Rataj, J. (2010, April). Iterative design process for the development and testing of cooperative applications. In *Proceedings of the 2010 Spring Simulation Multiconference*, (pp. 1-6).
- Marré, A. & Weber, B. (2010). Assessing community capacity and social capital in rural America: lessons from two rural observatories. *Community Development*, 41(1), p. 92-107, <https://doi.org/10.1080/15575331003661099>
- Martínez-Rolán, X., Tymoshchuk, O., Piñero-Otero, T. & Renó, D. (2019). Instagram como red de promoción e hipermediación del turismo rural: el caso de Aldeias Históricas. *Revista Latina de Comunicacion Social*, 74, 1610 - 1632
- Morgan, K. (1997). The learning region: institutions, innovation and regional renewal. *Regional Studies*, 31, 491-503.
- Muchagata, J. & Ferreira, A. (2019). Visual Schedule: A Mobile Application for Autistic Children-Preliminary Study. In *Proceedings of ICEIS 2*, (pp. 452-459).
- Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *Conference on Human Factors in Computing Systems - Proceedings*, 152-158. <https://doi.org/10.1145/259963.260333>
- Nielsen, J. (2012). *Usability 101: Introduction to usability (2012)*. Retrieved from URL: <http://www.nngroup.com/articles/usability-101-introduction-to-usability/>.
- Nielsen, J. & Landauer, T. K. (1993). A mathematical model of the finding of us-ability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213). ACM.
- Nielsen, J. (15 de Novembro de 2020). *10 Usability Heuristics for User Interface Design*. Retrieved from <https://www.nngroup.com/articles/ten-usability-heuristics/>
- Partridge, D. (2017). *Launch Your Dream: A 30-day Plan for Turning Your Passion In-to Your Profession*. Harper Collins Leadership.
- Oliveira, E., Branco, A. C., Carvalho, D., Sacramento, E. R., Tymoshchuk, O., Pedro, L., & Ramos, F. (2021). End-user Evaluation of a Mobile Application Prototype for Territorial Innovation. In *ICEIS (2)*, (pp. 495-504).
- Saint-Onge, H., & Wallace, D. (2012). *Leveraging communities of practice for strategic advantage*. Routledge.
- Salazar, L. H., Lacerda, T., von Wangenheim, C. G. & Barbalho, R. A. (2012). Customizando heurísticas de usabilidade para celulares. In *Companion Proceedings of the 11th BSHFCS*, (pp. 37-38).
- Salgado, A. D., Fortes, R. P. M., Hung, P. C. & Moreira, D. D. A. (2019). A Method for Classifying Usability Findings to Enhance Validation of New Heuristics. *Revista de Sistemas e Computação-RSC*, 9(1), 96-118.
- Seifi, S. (2015). *Developing an Engaging Local Community Application as a Tool for Promoting Social Integration*. (Doctoral dissertation, Malmö University). Retrieved from <http://ls00012.mah.se/handle/2043/19614>
- Seyfang, G. & Smith, A. (2007). Grassroots innovations for sustainable development: Towards a new research and policy agenda. *Environmental politics*, 16(4), 584-603.
- Silva, A., Antunes, M. J., Tymoshchuk, O., Pedro, L., Almeida, M., & Ramos, F. (2021). Barriers and incentives to territory-based innovation processes: from technology to interaction among actors. In *Handbook of Research on Cultural Heritage and Its Impact on Territory Innovation and Development*, (pp. 44-61). IGI global.
- Snyder, W. M. & Wenger, E. (2010). Our world as a learning system: A communities-of-practice approach. *Social learning systems and communities of practice* (pp. 107-124). Springer, London.
- Tymoshchuk, O., Almeida, A.M., Ramos, F., Pedro, L., Silva, P., & Renó, D. (2019). Mediação digital para a inovação territorial: um estudo de caso múltiplo em redes sociais digitais. *Observatorio (OBS*)*, 13(4), 70-88.
- Zhang, D. & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International journal of human-computer interaction*, 18(3), 293-308. https://doi.org/10.1207/s15327590ijhc1803_3

Relationship between Demographic Factors and Metacognition in Digital Library Interaction

Heesop Kim^a, Aluko Ademola^b and Yumi Kim^c

Department of Library and Information Science, Kyungpook National University, Daehak-ro 80, Daegu, South Korea
heesop@knu.ac.kr; alukojo Joshua@knu.ac.kr; yumikim@knu.ac.kr

Keywords: Metacognition Awareness, Digital Library Interaction, Information Searching, University Students, Demographic Factors.

Abstract: This paper aims to investigate the relationships between the university students' demographic factors and their metacognition in digital library interaction. To achieve the objectives of this study, the demographic factors were divided into six variables (i.e., gender, age, qualification, academic backgrounds, searching skills, and experience with the digital library) and the metacognition was classified into nine variables (i.e., schema-training, planning, monitoring, evaluation, transfer, memory, comprehension, task, technology). A total of 112 students participated in the online questionnaire. The collected data were analyzed using SPSS version 26, and a significant correlation between demographic elements and metacognition was found in digital library interactions. Interestingly, the experience with digital libraries has shown the strongest factor to be considered as the most important variable in the design of digital library interactions.

1 INTRODUCTION

Over the years, the storage and preservation of physical materials and notable books have been the hallmark of a conventional library. Users would have to go to the library to learn and search for detailed information. This trend gradually changed with the introduction of computers in the processing and searching of information some decades ago. The rapid development of Information Communication and Technology (ICT) has brought about unprecedented changes. Present developments in ICT have brought positive change to how we store, generate, access, and use information (Chowdhury and Chowdhury, 2003). Due to the role of ICT, libraries are undergoing fundamental changes from the traditional library to the digital library. Digital library searches become faster than traditional libraries and provide easy access, fast-tracking, and many electronic resources, although one can't neglect the feel of holding books which means we can't denigrate the importance of traditional libraries. (Case, 2002) describes information searching as a deliberate endeavor to obtain information to eliminate a deficiency of knowledge or meet a knowledge

demand. However, due to the significant changes in the libraries, studies show that during information searching in the digital library, users experience less optimal search results or suffer cognitive overhead. This is due to a lack of information tailored to the users' interests, preferences, needs, and context. For instance, users in traditional libraries can learn how to find information from librarians if assistance is needed.

However, that may not be the case in the modern digital library, where information professionals' help is not always available, and users are left alone to access, search, and use information. Generally, when students looking for information, they prefer to information, students prefer to look at online sources rather than practice in a methodical and organized manner, making learning information sources in the digital library difficult. In this regard, one could argue that self-knowledge may drive successful information searching in the digital library environment with the help of "metacognition." The term metacognition is most commonly connected with (Flavell, 1976), who is credited as being the first scholar to use the notion of metacognition, a term he coined from the word "metamemory" and first used in his early 1970s publications. According to (Flavell, 1979), metacognition is the act of a thinker reflecting on their mental process. It involves people comprehending both them-

^a  <https://orcid.org/0000-0002-4642-4445>

^b  <https://orcid.org/0000-0002-4314-2143>

^c  <https://orcid.org/0000-0002-1765-1999>

selves and the nature of the activity and understanding the specification of the task.

Metacognition is defined by (Swanson, 1990) as an individual awareness of their ability to monitor, regulate and manage their learning activities. Metacognition refers to higher-level planning for handling a learning activity and mechanisms for observing and evaluating comprehension. At times, individuals utilize the expression "going meta" when discussing metacognition, alluding to the interaction of reviewing a sort of flashback to see what you are doing, as though you were another person noticing it. Sternberg refers to these executive processes as meta-components in his triarchic theory of intelligence (Sternberg, 1984). According to (Sternberg, 1986) meta-components are in charge of "determining how to perform a certain task or set of activities, and then ensuring that the job or set of tasks is completed correctly and the ability to organize cognitive resources effectively, such as selecting how and when a task should be completed is centered to intelligence. Metacognition is a fundamental human trait (Muhali, 2018) that is defined as thinking about thinking by going through a process of review (McCormick, 2003).

Having metacognition awareness assists individuals in taking a more active role in planning, analyzing, evaluating, and monitoring their learning processes (Akyol and Garrison, 2011). Individuals' metacognition awareness is assumed to play an essential part in the online information searching process by structuring the searching process, monitoring, and assessing much as it is in the learning process. Resources on the relationship between metacognition and the search process and its applications are needed (Bowler, 2009). So much research has been done in the academic field, but much research has not been done in the digital library interaction regarding metacognitive skills.

This research examines the relationship between demographic factors and metacognition in digital library interaction. The study conducted for this purpose attempts to find answers to the following questions below:

- (a) Is there a significant relationship between demographic factors and metacognition when interacting with digital libraries?
- (b) What metacognitive skills do students have when interacting with digital libraries?

2 RELATED WORK

Through the reviewing the previous research, we understood how various techniques of analysis have spe-

cific strengths and weaknesses. (Gorrell et al., 2009) conducted a research on metacognitively aware IR systems. They classified metacognitive components into five categories: schema-training, planning, monitoring, evaluation, and transfer. They used metacognition information Likert-based knowledge generator, so called MILK, to offer the coverage of the sub-areas in the taxonomy. They found that metacognitive uses vary greatly depending on gender, age, and academic discipline and that older adults, in particular, use metacognitive skills more than younger adults.

(Madden et al., 2012) revealed whether the application of user evaluation standards provides Metacognitive evidence for web reliability, a total of 48 graduate students from various disciplines participated in the study. It was observed that some students encountered some difficulties because of an overly simplistic approach to assessing sites. The students were given different evaluation criteria such as websites (name, description, appearance, references) and contents such as authorship, quality of writing, evidence of research, and contacts. According to the findings, some participants reported dissatisfaction with the lack of information about the author's identity. It was also noted that the students in the study appeared to have received little formal assistance on the evaluation, which provides the most direct indication of metacognition behavior, according to (Bonds et al., 1992), will not only hold information in ways that vary depending on the task. In this study, some volunteers engaged in such conduct.

(Cadamuro et al., 2019) investigated the relationship between metacognition and information and communication technology (ICT) in undergraduate students. Metacognitive reflection during an online course was found to build advanced epistemic agency in students and increase decision-making skills through metacognitive tips in the e-learning environment. Another observation during the research is that metacognition self-regulation and organization predicted students' performance and information searching behavior. The ICT and Metacognition are likely in a bi-directional relationship. On the other hand, working in technology-mediated contexts supports the development of metacognitive skills, leading to better learning outcomes. However, metacognitive skills are necessary to take advantage of web-based training. The use of metacognitive skills in ICT has contributed to creating powerful learning and information search.

Another previous research is the think-aloud procedure focusing on comparing the use of these strategies by people with different levels of critical thinking in searching behavior; participants were asked to

think aloud everything they thought of when completing the task that attracted favorable attention. (Ericsson and Simon, 1984) argue that it reflects the thinking process more directly than other oral reports that require participants to express their ideas in words during the task. It is also considered to be a straightforward method to gain insight into human knowledge and techniques of problem-solving,” similarly to (Crain-Thoreson et al., 1997). The study examines think-aloud procedures to check the reading comprehension process and found that only the conversion response represents after modifying the coding scheme. The variance in comprehension score accounted for by the usage of comprehension strategy increased from 20% to 40% after adjusting the coding system to include only transformed responses that demonstrated active and correct integration of passage content.

(Reisoğlu et al., 2020) analyzed 20 university students on online information searching and metacognitive abilities during argumentation activities, focusing on three phenomena: argumentation activities, metacognitive skills, and the link between them. By examining screen recordings and the online information searching strategy inventory (OISSI) created by (Tsai, 2009), different results were obtained in terms of metacognition awareness and online information searching methods. The findings of the interviews demonstrated a link between participants’ online information searching tactics used in argumentation activities and all dimensions of metacognition. Furthermore, the data shows that metacognition skills are frequently linked to multiple sub-dimensions in (Tsai, 2009) framework and might readily encompass a variety of techniques. Again, online argumentation activities can be argued to allow for extensive use of metacognitive abilities in planning, monitoring the information gathering process, and applying appropriate strategies. According to (van Geel et al., 2015) metacognition is essential in solving complicated information problems.

However, given that information search is a problem-solving process, metacognition has always been a missing piece in the puzzle to understanding users better. The current study used a metacognitive paradigm from educational psychology to illuminate the research problem. However, further research is needed to categorize users’ difficulties during information searching through digital libraries and develop adequate taxonomy, which are five core metacognition skills to address these challenges.

3 METHODOLOGY

This section will present the study design, instrument, participants, and data analysis information.

3.1 Research Design

The quantitative research approach was employed to identify the demographic differences in metacognition in digital library interaction. This strategy was used to gather data on the subject and contribute to a more comprehensive understanding.

3.2 Participants

The study was conducted to gather adequate information from Kyungpook National University students (Koreans and international students) and Nigerian university students. This research tends to employ a simple random sampling technique in selecting one hundred and twelve (112) respondents for this research. This includes 20 Knu-Korea students, 35 Knu-International students, and 57 Nigeria university students. The bulk of the 112 participants was male (57%) and female (43%) respectively. The participant’s ages range from 18 to 43 years above. Participants have different qualifications ranging from bachelor’s degrees, master’s degrees, doctorate degrees (Ph.D.), and post-doc.

3.3 Metacognition Components

The existing concept of five metacognitive skills used in the study Schema-training, planning, monitoring, evaluation, and transfer were developed by (Brown, 1975). Schema-training: This involves the development of cognitive structures that provides a conceptual framework for comprehension (Gordon and Braijn, 1985). Planning: This refers to carefully evaluating the current situation in order to determine potential causes of action to implement solutions to achieve results and make efficient use of the time and resources (Erenler and Cetin, 2019). Monitoring: This is about the person’s awareness of how she performed in relation to the process that was planned. (Schraw, 1998) Evaluation: This refers to how the person evaluates both the organization process and the outcomes of her own learning (Schraw and Moshman, 1995) Transfer: It refers to the process of moving one’s skills and knowledge from one problem-solving scenario to another. The low and high-road theory on learning transfer, developed by (Perkins et al., 1992) However, (Wallace and Kupperman, 1997), who also explores web searching,

Table 1: Metacognition components.

	Memory	Comprehension	Task	Technology
Schema-training	[S-M]	[S-C]	[S-Ta]	[S-Te]
Planning	[P-M]	[P-C]	[P-Ta]	[P-Te]
Monitoring	[M-M]	[M-C]	[M-Ta]	[M-Te]
Evaluation	[E-M]	[E-C]	[E-Ta]	[E-Te]
Transfer	[T-M]	[T-C]	[T-Ta]	[T-Te]

recommends breaking down the rest of the required skills into two parts task and technology. In addition, metamemory and meta-comprehension are subareas of Metacognition (Brown, 1975). All these make it four elements applied to five components of the task. As shown in Table 1, there are twenty subareas.

Based on the subareas of Table 1, a 32-item questionnaire was developed by the researcher using the Google forms feature. The questionnaire aims to cover all the domains outlined in the taxonomy and explore participants' perceptions of their usage of selected metacognitive skills while searching on the web and digital libraries. The measurement used is a 5-point Likert type scale range and was set out and scored in the following way 5= Strongly agreed, 4= Agree, 3=Neutral, 2= Disagree, 1= Strongly disagree.

3.4 Data Analysis

The data collected was coded organized and processed by the researcher using the statistical package for the social sciences (SPSS version 26). Descriptive statistics of frequency count were used to provide information on the number of males and females that participated in the research. It was also used to analyze the age, qualification, academic background, searching skills, and experience with a digital library.

The mean and standard deviation were used to analyze the research question to determine the metacognitive skills possessed when searching through the digital library. Bivariate correlation was used to analyze the relationships between demographics and metacognition awareness.

4 RESULTS

The results of descriptive statistical analyses were used to determine the university student's metacognitive competence scores. The overall mean score is ($\bar{x} = 3.62$) and the criterion mean is 3.0 on a five-point scale of metacognitive skills which was divided into five dimensions namely: schema-training, planning, monitoring, evaluation, and transfer. The group

means for each metacognitive skills dimension was also calculated. The result revealed that schema training ($\bar{x} = 3.48$), planning ($\bar{x} = 3.64$), monitoring ($\bar{x} = 3.63$), evaluation ($\bar{x} = 3.54$) and transfer ($\bar{x} = 3.81$). Full results are listed in Table 3.

Bivariate correlation analyses were conducted to discover relationships between demographic information (gender, age, qualification, discipline, searching skills, and experience with digital library) and scores for the various metacognitive components. A further correlation analysis sought relationships between demographic information and metacognition. The finding revealed the Pearson correlation analysis which shows that there is negative significant relationship between gender and task ($r = -0.265, p > 0.05$) as well as on schema ($r = -0.235, p > 0.05$) and planning ($r = -0.256, p > 0.05$). It also showed that there is negative significant relationship between age and comprehensive ($r = -0.190, p > 0.05$) as well as on schema ($r = -0.198, p > 0.05$). Furthermore, there is negative significant relationship between academic background and comprehensive ($r = -0.224, p > 0.05$) as well as on planning ($r = -0.270, p > 0.05$). Finally, there is negative significant relationship between experience with digital library and memory ($r = -0.249, p > 0.05$) as well as on comprehension ($r = -0.199, p > 0.05$), task ($r = -0.279, p > 0.05$), technology ($r = -0.234, p > 0.05$), schema ($r = -0.395, p > 0.05$) and planning ($r = -0.242, p > 0.05$).

5 DISCUSSION

Table 2 shows that there is a negative significant relationship between task, schema, and planning based on gender. This implies that males have a strong negative significance in the metacognitive skills in the main component of schema training $p = 0.013$ and planning $p = 0.006$ and subareas task $p = 0.005$ than females to a greater extent. Furthermore, age has a negative significance with older people having better metacognition skills in schema training $p = 0.037$ and comprehension $p = 0.045$. This implies that older

Table 2: Pearson correlation coefficient (Pearson's r) and p -value between Demographics and Metacognition awareness.

	Pearson's r (p -value)					
	Gender	Age	Qualification	Academic Background	Searching Skills	Experience with digital library
Schema	-0.235* (< 0.013)	-0.198* (< 0.037)	-0.021 (< 0.825)	-0.137 (< 0.149)	0.143 (< 0.133)	-0.395** (< 0.000)
Planning	-0.256** (< 0.006)	-0.112 (< 0.239)	-0.055 (< 0.563)	-0.270** (< 0.004)	0.063 (< 0.509)	-0.242* (< 0.010)
Monitoring	-0.032 (< 0.739)	-0.064 (< 0.501)	-0.012 (< 0.904)	-0.172 (< 0.070)	-0.031 (< 0.744)	-0.077 (< 0.419)
Evaluation	-0.089 (< 0.353)	-0.076 (< 0.426)	-0.134 (< 0.159)	0.169 (< 0.075)	0.017 (< 0.861)	-0.139 (< 0.143)
Transfer	-0.126 (< 0.186)	0.076 (< 0.424)	-0.062 (< 0.515)	0.044 (< 0.646)	0.095 (< 0.320)	-0.149 (< 0.116)
Memory	-0.146 (< 0.125)	-0.074 (< 0.437)	-0.070 (< 0.463)	0.064 (< 0.506)	0.119 (< 0.210)	-0.249** (< 0.008)
Comprehension	-0.111 (< 0.243)	-0.190* (< 0.045)	-0.080 (< 0.404)	-0.224* (< 0.018)	-0.009 (< 0.923)	-0.199* (< 0.036)
Task	-0.265 ** (< 0.005)	-0.100 (< 0.296)	-0.040 (< 0.678)	-0.123 (< 0.198)	0.110 (< 0.248)	-0.279** (< 0.003)
Technology	-0.150 (< 0.114)	0.004 (< 0.967)	0.003 (< 0.975)	-0.033 (< 0.726)	-0.024 (< 0.798)	-0.234* (< 0.013)

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

people have metacognitive accuracy for their capacity and older adults can accurately assess their ability to selectively remember high-value information. There is a negative significant relationship between the main component of planning $p = 0.004$ and sub-areas comprehension $p = 0.010$ based on the academic background which indicates that information technology students possessed high metacognition skills which comprise the highest percentage of students that participated in the research. Experience with the digital library is very much influenced and an important factor in both schema training, $p = 0.000$ planning, $p = 0.010$ memory, $p = 0.008$ comprehension $p = 0.036$ task, $p = 0.033$ technology $p = 0.013$. Therefore, gender, age, academic background, and experience with the digital are compelling factors that determine the level of the metacognitive skills of the students.

Table 3 shows that University students reported higher use of metacognition in general, with the highest score of ($\bar{x} = 3.81$) in transfer and in most of the subareas of metacognition skills. Under schema training majority of the students possessed high metacog-

nitive skills in schema training-memory. This shows that a higher percentage of university students have a number of techniques at their disposal to help them remember what they learned and where they found it. It was observed that 52% of respondents lack techniques that can help them understand the information they find while using a digital library.

Under planning, it could be concluded that planning comprehension is significantly high, this shows that the majority of the students possessed skills in planning to summarize in order to monitor and improve their comprehension. While they lack plan search task strategies such as verifying how trustworthy the information from digital libraries is from different academic journals. Importantly, a lack of planning doesn't necessarily mean an ineffective search. In other words, the high potential in planning which in some cases inhibits efficient retrieval.

A higher percentage of students under monitoring show strong metacognition skills in the monitoring search task which indicates when using the digital library for a learning task, they ask themselves ques-

Table 3: Types of metacognitive skill.

S/n	Types of metacognitive skills	$\bar{x}$	σ
Schema-Training			
[S-M]	I have techniques that help me remember any information I come across on digital library.	3.84	0.94
[S-C]	I have techniques that help me understand the information I find while using digital library.	3.21	1.19
[S-Ta]	I have developed ways of identifying the type of information I need for my learning tasks while using digital library.	3.33	1.24
[S-Te]	I am good at choosing keywords on user interface of digital library to get optimized feedback.	3.52	1.18
Average mean		3.48	1.14
Planning			
[P-M]	I plan ways to remember the information I find in the digital library.	3.88	0.84
[P-C]	I may use the digital library to start a search task to increase my understanding on a subject area.	3.87	1.02
[P-Ta]	I decide in advance exactly what type of information I am looking for on digital library.	3.17	1.33
[P-Te]	I tend to work out my search skills before using a digital library.	3.65	1.02
Average mean		3.64	1.05
Monitoring			
[M-M]	Sometimes when using the digital library, I am aware that I might for get the information I find.	3.68	1.01
[M-C]	Sometimes when using the digital library, I may have misunderstood information I read earlier in my search.	3.04	1.14
[M-Ta]	When using the digital library for a learning task, I find myself asking questions along the lines of: "Is this search providing the type of in formation I need?".	3.94	0.80
[M-Te]	At times I am cautions as I search (e.g. the words I put into the search box) in the digital library.	3.85	0.91
Average mean		3.63	0.97
Evaluation			
[E-M]	It is clear to me when I am failing to remember what I learned in the digital library.	3.49	0.93
[E-C]	While using the digital libraries I get convinced that the feedback of my search is valid.	3.18	1.15
[E-Ta]	I spend a lot of time judging how well the information I find in digital libraries matches my learning needs.	3.70	1.02
[E-Te]	It is usually obvious to me whether I am using a good search strategy on a digital library or not.	3.80	0.92
Average mean		3.54	1.01
Transfer			
[T-M]	I use different approaches to recall the information that I learned in other domains of digital library.	3.77	0.96
[T-C]	My experience with different learning tasks has helped me monitor how well I am understanding what I read in the digital library.	3.79	0.83
[T-Ta]	My experience in other areas helps me to work out exactly what type of information I need for my learning task in the digital library.	3.99	0.93
[T-Te]	The skills I apply when using a digital library are useful in other areas of my information searching.	3.70	0.94
Average mean		3.81	0.92
Overall mean = 3.62			

tions along the line if the search provides the type of information they need. It shows that 74% of the participants are aware of their progress towards completing the task and the reliability of the information they found. But possessed weak metacognition skills in monitoring comprehension which shows that when using a digital library they must have misunderstood information they read in the past when searching.

Under evaluation, it could be concluded that users possessed strong metacognition skills in evaluation technology which implies that it is usually obvious to the majority of university students whether they are using a good searching strategy in a digital library or not. It shows that participants critically evaluate their search method and use of technology and draw conclusions about the use. However, they lack evaluation comprehension skills while using digital libraries they get convinced if the feedback of their search is valid.

Under transfer, it shows that users possessed a higher level of metacognition skills in the transfer search task with 74% of participants who have experience from other areas helping them to work out precisely what type of information they need for their learning task. On the contrary, there is low metacognition awareness in transfer technology which indicates that the use of carrying learning from other tasks into their use of sources and technology is low.

6 CONCLUSION AND RECOMMENDATION

This study's findings imply that, regarding the issue with digital libraries, metacognition awareness is essential where it is lacking. Based on the conclusions of this paper, metacognitive skills were observed from different components and subareas of the taxonomy. The result shows that students demonstrated strong metacognition skills on transfer which is a result of metacognitive knowledge that helps them to select previously learned strategies to achieve searching goals or to deal with problems encountered during the task. According to (Groteluschen et al., 1990), if the chosen strategy is effective and searchers believe it improves learning quality and their specific strategy is stronger. The nature of this study cannot be considered indicative of the entire student body. Given the fact that there are 122 participants in the study, it seems reasonable that it doesn't cover larger numbers of students with different academic backgrounds. It is possible to argue that teaching students how to apply metacognitive methods improves their academic performance (1988). Students with advanced skills in metacognition may keep track of their own learning,

express their opinions on it, keep up to speed on their knowledge, and develop and implement new learning procedures. Students that effectively use their metacognitive skills are more aware of their strengths and weakness and seek to develop their skills further than other students (Bransford et al., 2000). The results show improvement in planning which shows that majority of the students demonstrate the use of implementing metacognitive strategies that require careful planning and a better understanding of metacognitive skills should provide students with potential strategies to use when searching for information in digital libraries.

According to (Jones et al., 1995) the greater a student's understanding of metacognition, the greater his or her efficiency. In the twenty-first century, metacognition strategies are needed. This will allow students to successfully cope with new situations when searching, and school library media specialists to capitalize on their skills and gain access to a wealth of resources, fostering the development of good thinkers who are successfully problem-solvers, searching skills, and information retrieval experts. Additionally, students with good metacognitive skills are better critical thinkers, problem solvers, or decision-makers than students without metacognitive skills, and also metacognitive training can increase students' self-confidence and personal responsibility for their own development.

The implication is that metacognition skills help individuals process and retain information through self-recognition and reflection. Metacognition skills are essential because they help students understand their learning process and how they learn effectively, and it also assists users in learning to control their searching by defining search goals and tracking their progress towards them. The taxonomy used in this research to identify the metacognition awareness from different nationalities can be used for future research on information searching through a web search. Its feature can be used not just for digital libraries.

ACKNOWLEDGEMENTS

This Research was supported by Kyungpook National University Research Fund, 2021.

REFERENCES

- Akyol, Z. and Garrison, D. R. (2011). Assessing metacognition in an online community of inquiry. *The Internet and higher education*, 14(3):183–190.

- Bonds, C. W., Bonds, L. G., and Peach, W. (1992). Metacognition: Developing independence in learning. *The clearing house*, 66(1):56–59.
- Bowler, L. (2009). Adolescent metacognitive knowledge during the information search process. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–4.
- Bransford, J. D., Brown, A. L., Cocking, R. R., et al. (2000). *How people learn*, volume 11. Washington, DC: National academy press.
- Brown, A. L. (1975). The development of memory: Knowing, knowing about knowing, and knowing how to know. *Advances in child development and behavior*, 10:103–152.
- Cadamuro, A., Bisagno, E., Pecini, C., and Vezzali, L. (2019). Reflecting a...“bit”. what relationship between metacognition and ict? *Journal of E-Learning and Knowledge Society*, 15(3):183–195.
- Case, D. O. (2002). Looking for information: A survey of research on information seeking. *Needs, and behavior*, pages Fourth Edition: 146–48.
- Chowdhury, G. G. and Chowdhury, S. (2003). *Introduction to digital libraries*. London: Facet publishing.
- Crain-Thoreson, C., Lippman, M. Z., and McClendon-Magnuson, D. (1997). Windows on comprehension: Reading comprehension processes as revealed by two think-aloud procedures. *Journal of Educational Psychology*, 89(4):579.
- Erenler, S. and Cetin, P. S. (2019). Utilizing argument-driven-inquiry to develop pre-service teachers’ metacognitive awareness and writing skills. *International Journal of Research in Education and Science*, 5(2):628–638.
- Ericsson, K. A. and Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. the MIT Press.
- Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.) *The nature of intelligence*.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.
- Gordon, C. J. and Braijn, C. (1985). Metacognitive processes: Reading and writing narrative discourse. In D.L Forrest-Pressley, G.E. Mackinnon, T.G. Waller (eds.) *Instructional Practices: 1-75*. Academic press, pages 1–75. Academic Press.
- Gorrell, G., Eaglestone, B., Ford, N., Holdridge, P., and Madden, A. (2009). Towards “metacognitively aware” ir systems: an initial user study. *Journal of documentation*, 65(3) 446-469.
- Groteluschen, A. K., Borkowski, J. G., and Hale, C. (1990). Strategy instruction is often insufficient: Addressing the interdependency of executive and attributional processes. In *Intervention research in learning disabilities*, pages 81–101. Springer.
- Jones, M. G., Farquhar, J. D., and Surry, D. W. (1995). Using metacognitive theories to design user interfaces for computer-based learning. *Educational Technology*, 35(4):12–22.
- Madden, A. D., Ford, N., Gorrell, G., Eaglestone, B., and Holdridge, P. (2012). Metacognition and web credibility. *The electronic library*, 30(5):671–689.
- McCormick, C. B. (2003). Metacognition and learning. *Handbook of psychology*, 7:79–102.
- Muhali, M. (2018). Pengembangan model pembelajaran reflektif-metakognitif untuk meningkatkan kemampuan metakognisi siswa sma (development of reflective-metacognitive learning model to improve high school students’ unpublished ph.d. thesis, universitas pendidikan mandalika.).
- Perkins, D. N., Salomon, G., et al. (1992). Transfer of learning. *International encyclopedia of education*, 2:6452–6457.
- Reisoğlu, İ., Toksoy, S. E., and Erenler, S. (2020). An analysis of the online information searching strategies and metacognitive skills exhibited by university students during argumentation activities. *Library & Information Science Research*, 42(3):101019.
- Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional science*, 26(1):113–125.
- Schraw, G. and Moshman, D. (1995). Metacognitive theories. *Educational psychology review*, 7(4):351–371.
- Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences*, 7(2):269–287.
- Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills*. Harcourt Brace Jovanovich, ISBN: 0155414704.
- Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem-solving. *Journal of educational psychology*, 82(2):306.
- Tsai, M.-J. (2009). Online information searching strategy inventory (oissi): A quick version and a complete version. *Computers & Education*, 53(2):473–483.
- van Geel, M., Goemans, A., and Vedder, P. (2015). A meta-analysis on the relation between peer victimization and adolescent non-suicidal self-injury. *Psychiatry research*, 230(2):364–368.
- Wallace, R. and Kupperman, J. (1997). On-line search in the science classroom: Benefits and possibilities. page 31 <https://files.eric.ed.gov/fulltext/ED407252.pdf>.

Effect of Environment Size on Spatial Perception in Virtual Reality

Morten Trier Bach, Poul Anker Jensen, Andrei Iuliu Lucaci, Damian Grzegorz Pupczyk
and Claus Brøndgaard Madsen ^a

Computer Graphics Group, Department of Architecture and Media Technology, Aalborg University, Aalborg, Denmark
bach97@live.dk, poulankerjensen@yahoo.dk, lucaci.andrei@ymail.com, damianpupczyk@gmail.com, cbm@create.aau.dk

Keywords: Virtual Reality, Virtual Environments, Spatial Perception, Blind Walking.

Abstract: A trend of distance underestimations in Virtual Reality (VR) is well documented, but the reason still remains unclear. Therefore, this paper investigates the effect of differently sized Virtual Environments (VEs) on ego-centric distance perception in VR as a potential influence. Verbal assessment, blind walking, and our own proposed method: walk and assess, were compared in an experiment, and blind walking was found to be the most accurate. A virtual replica of a real-life location was created as a transitional environment, while small (15m²), medium (35m²) and large (95m²) rooms were created to investigate the effect of VE size on spatial perception in VR. To establish the differences in estimations between the real world and VR when using blind walking, a study was conducted with the virtual replica and its real life counterpart at distances between 1 and 10 meters. Following this, the three distinct room sizes were used in an experiment to investigate the effect of the size of rooms on spatial perception in VR. The findings showed consistent underestimates of distances, and a trend for underestimation to grow as the distance grows was observed. Similarly, underestimates grew with the size of the environment.

1 INTRODUCTION

Virtual Reality (VR) as a technology has been shown to be influential within the world of architecture, where it is used to represent and simulate architectural spaces Loyola (2018). Unfortunately, quite a few researchers have found that distances in Virtual Environments (VEs) are generally underestimated by 10 to 26 percent (Loyola (2018), Witmer and Sadowski Jr. (1998), Ahmed et al. (2010), Renner et al. (2013)), which threatens the validity of the platform as a medium for representation of spaces, and makes it a subject worth exploring.

In an experiment done by Lucaci et al. (2022), findings indicated general overestimation of distances. They hypothesized that it might have been due to the relatively small size of the environment (12.6m²) compared to previous research which used larger environments ranging from 13m² to 233m² (Simpson et al. (2018), Moscoso et al. (2021), Zhao et al. (2019)).

Lucaci et al. (2022) used *Verbal Assessment* to measure perceived distances, however, Thompson et al. (2004) have criticised the method for being subjected to bias and noise.

With this in mind, the aim of this study was to *identify a reliable method of quantifying perceived distances, and use that to investigate the effect of different sizes of VEs on spatial perception in VR.*

2 SPATIAL PERCEPTION IN VIRTUAL REALITY

Spatial perception is a combination of physiological and cognitive processes which utilize a variety of perceived depth cues to turn a two-dimensional image into a three-dimensional representation of the world (Renner et al., 2013).

Perceived distances can be either ego-centric or exo-centric. Ego-centric distances refer to distances between an observer and a point, while exo-centric distance is the distance between two external objects. (Maruhn et al. (2019), Renner et al. (2013))

Underestimation of ego-centric distances in Virtual Reality (VR) was found to be 19% by Loyola (2018) and 15% by Witmer and Sadowski Jr. (1998), while real-world distances were found to be underestimated by 8% by Witmer and Sadowski Jr. (1998), and 4% by Messing and Durgin (2005). These underestimations are particularly evident with egocen-

^a  <https://orcid.org/0000-0003-0762-3713>

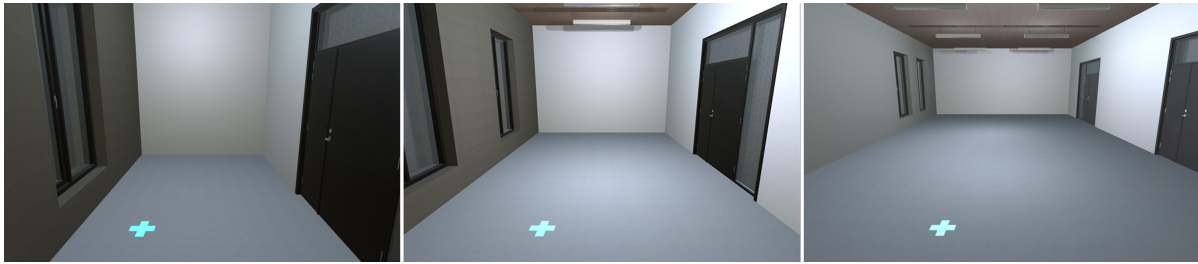


Figure 1: The three differently sized rooms used in the *Room Size Experiment* in Section 6.

tric distances above 1 meter. On the other hand, some studies have indicated that exo-centric distances in VR may be overestimated in VR. (Maruhn et al., 2019)

However, the focus of this paper is only on ego-centric distances to emphasize the room size effect. It is also applicable to more methods of distance estimation than exo-centric distances (see Section 4).

2.1 Depth Cues

To properly understand spatial perception, a basic understanding of the depth cues which allow us to perceive depth is needed. Such an understanding can aid in identifying what might lead to errors in distance estimations.

The depth cues used to inform the three-dimensional understanding of the world can be divided into two categories; pictorial and non-pictorial cues. Pictorial cues are cues that can be obtained from a motionless image, while non-pictorial cues are largely from motion in the scene or the human visual system (Renner et al., 2013). The pictorial cues include occlusion, relative and familiar size, horizon ratio, shading and texture gradients (Proffitt and Caudek, 2012).

Non-pictorial cues include binocular disparity, motion parallax and ocular motor (accommodation and vergence) (Proffitt and Caudek, 2012). Maruhn et al. (2019) suggest that visual cues might differ between physical environments and VR in a multitude of ways, such as by introducing the vergence-accommodation conflict, which is a result of presenting different images to each eye using stereoscopic displays.

This phenomenon can be a source of visual discomfort and dizziness.

When discussing distance perception of the world, one should also consider the circular segmentation of egocentric distances around a moving observer as presented by Cutting and Vishton (1995). The segments and their associated depth cues are shown in Figure 2. The first of these segments is referred to as *Personal Space*, and is the area closest to an individual

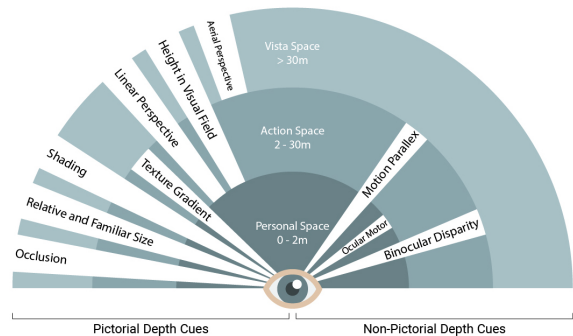


Figure 2: The interplay between segmentation of egocentric distances and most important visual cues.

(within arm's reach) delimited by Cutting and Vishton (1995) to 2 meters. The second segment is called *Action Space* which extends from the bounds of the personal space out to about 30 meters - one can move relatively quickly, speak, teach and toss objects with relative ease in this segment. The final segment is called the *Vista Space* which extends beyond 30 meters, at these distances the benefits of binocular vision and movement of the observer significantly diminishes. (Cutting and Vishton, 1995)

3 ROOM SIZES IN VIRTUAL REALITY

When acting in an environment, the size and layout of it profoundly impacts one's perceived personal, action, and vista spaces, as distances to objects and walls might limit the area one can interact in and the available depth cues. To our knowledge, no prior research has been done with a particular focus on the effect of room sizes on spatial perception. Nevertheless, the sizes of Virtual Environments (VEs) used in Virtual Reality (VR) research can be incorporated to inform the design of our VEs.

Sizes of rooms used in such studies vary from just 13.46m^2 in a study by Zhao et al. (2019) all the way to 233.26m^2 ($21.44\text{m} \times 10.88\text{m}$) in a study by Moscoso et al. (2021).

While Moscoso et al. (2021) only used two different rooms in their studies, Simpson et al. (2018) incorporated 7 square rooms with the sizes 16m^2 , 18.75m^2 , 21.34m^2 , 25m^2 , 28.40m^2 , 32.14m^2 , 36m^2 , and uniform heights of 2.74m to study the impact of visual design elements on perception of spaces.

Zhao et al. (2019) used 4 square rooms of sizes 16m^2 , 21.80m^2 , 28.40m^2 , and 36m^2 to the influence of what is seen through windows of a room on perceived spaciousness of rooms. Zhao et al. (2019) also presented a room of 13.46m^2 size as a reference to spaciousness of 1, and a room of size 40.06m^2 as a reference for spaciousness 10.

3.1 Categorization of Room Sizes

Based on the rooms used by Simpson et al. (2018), and Zhao et al. (2019), a survey was designed and distributed to 45 people. The survey consisted of 10 rectangular floor plans (9m^2 , 19m^2 , 29m^2 , 39m^2 , 49m^2 , 59m^2 , 63m^2 , 75m^2 , 86m^2 , and 97m^2), which were presented to participants in random order. Participants were asked to rate whether they found each room to be *small*, *medium* or *large*.

The weighted average of the results were computed for each group; the boundary between *small* and *medium* was $\sim 25\text{m}^2$ and between *medium* and *large* was $\sim 52\text{m}^2$.

From these results the intervals were defined as: $0\text{--}25\text{m}^2$ for *small*, $25\text{--}50\text{m}^2$ for *medium*, and above 50m^2 for *large*. This categorization was used to inform the sizes of rooms used in the *Room Size Experiment* presented in Section 6.

4 METHODS FOR QUANTIFYING PERCEIVED DISTANCES

To investigate and compare perceived distances in a set of virtual environments, one must be able to quantify them. Two of the most prominent methods for quantifying perceived distances are:

- **Verbal Assessment:** (Loyola (2018), Peer and Ponto (2017), Armbrüster et al. (2008), Gagnon et al. (2020), von Castell et al. (2018), Kelly et al. (2017), Klein et al. (2009))
- **Blind Walking:** (Interrante et al. (2006), Kelly et al. (2018), Li et al. (2015), Kelly et al. (2017), Ahmed et al. (2010))

It was decided to evaluate *Verbal Assessment*, *Blind Walking*, and *Walk and Assess* (a method where participants walk a distance of their choice from a starting point, and then estimate it verbally, devised by

us) with participants to identify the method that performed most similarly between real life and VR with the highest accuracy. The experiment was conducted on 30 participants with perfect 20/20 or corrected vision, who would estimate three distinct distances for each chosen method. *Verbal Assessment* resulted in 22% average error in VR and 13% in the Real World (RW), *Walk and Assess* showed 25% average error in VR and 18% in RW, and *Blind Walking* showed 12% average error in both VR and RW. Due to the fact that *Blind Walking* performed most similarly between VR and RW, and resulted in the least average error in both VR and RW, we chose this method to be used in the succeeding experiment.

5 BLIND WALKING BASELINE EXPERIMENT

To establish what difference might exist in estimates between the Real World (RW) and Virtual Reality (VR) when using *Blind Walking*, an experiment to evaluate the method was designed. The purpose was the establishment of a baseline difference between VR and RW, as well as it being a pilot test of the method for the final investigation of room sizes effect on spatial perception in VR.

5.1 Participants

40 participants were recruited with normal (20/20) or corrected vision. All participants were university students recruited by convenience, and had an average Interpupillary Distance (IPD) of 65mm.

5.2 Apparatus

A Quest 2 Head Mounted Display (HMD) was used, untethered to allow free movement, connected to a laptop with Oculus Airlink. The laptop had a Nvidia Rtx 3060 graphics card and ran Windows 11.

In order to minimise the blurriness that participants might experience from an uncalibrated HMD, each participant's Interpupillary Distance (IPD) settings were measured and adjusted to one of the three physical settings on the Quest 2 (58mm, 63mm, and 68mm).

The Field of View (FOV) of the Quest 2 was also measured, and found to be 94° (IPD 58mm), 93° (IPD 63mm) and 95° (IPD 68mm) vertically, with the horizontal FOVs being 93° (IPD 58mm), 94° (IPD 63mm) and 86° (IPD 68mm).

5.3 Procedure

The participants would enter a large empty room (130m²) assigned as a testing area (as shown on Figure 3) and complete a consent form.

A facilitator (facilitator 2 in Figure 3) shortly briefed each participant of the procedure of the experiment, and would keep the participants informed of what would happen next during the experiment. A Latin square was utilized to counterbalance the order of the conditions and distances between participants.

In the real world condition participants were first shown a target point, after which they were asked to blindly walk to the target point. The participants were asked to close their eyes and wear a sleeping mask to prevent them from seeing the target point while walking. After assessing a distance by walking to the point, the distance was measured by a facilitator with a laser range finder, and the participants were asked to return to their starting position which was marked on the floor with a cross. The process was repeated until all distances were assessed.

In the VR condition, participants entered a virtual replica of the empty room and went through the same process as described above. Instead of a sleeping mask, the screen in the HMD was faded to black after participants were informed to close their eyes. The participants' walk was assessed by the software instead of a laser range finder. The changing of targets, fading to black, and noting of measurements was done by facilitator 1 as seen in Figure 3.

To ensure participants' safety during their *Blind Walking*, facilitator 3 (in Figure 3) would prevent them from colliding with walls should they have walked too far.



Figure 3: Visualization of the test setup for the *Blind Walking* baseline experiment.

5.4 Results

To compare the distance assessments across conditions and distances, the amounts of error in the assessments were expressed in percentage. The absolute values of results were analyzed in the context of accuracy, defined as *"how much the assessments deviate from the actual distances"* (Loyola, 2018), while the standard deviation of relative values were used to represent the precision of the assessments, defined as *how similarly participants within a certain group assessed the distances* (Loyola, 2018).

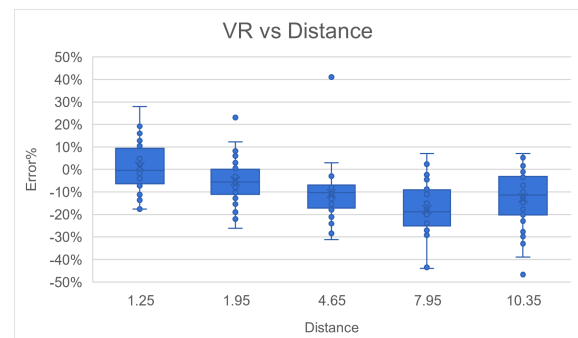


Figure 4: Participants' precision in VR at given distances reveals a tendency to underestimate as distances gets larger.

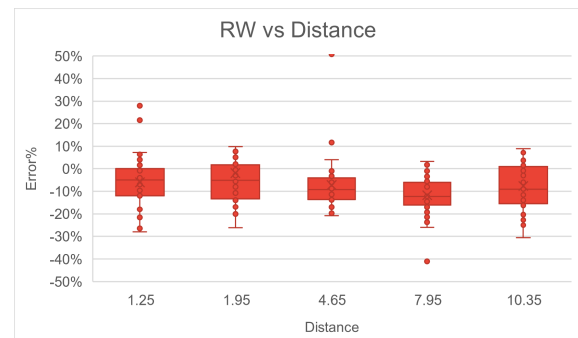


Figure 5: Participants' precision in RW at given distances shows a similar trend of underestimation as in the VR condition in Figure 4.

The mean relative error for the VR condition was -9.21%, and the mean relative error for the RW condition was -6.84% - this shows that on average, participants underestimated distances in both conditions, with underestimations being greater by 2.37% in VR than RW. The standard deviation of the relative errors in VR was 7.42%, while the standard deviation for the real world condition was 8.9%. As the VR condition had a slightly lower standard deviation (1.51% difference), it shows that the VR condition had slightly higher precision than in the RW condition. This was shown to not be a statistically significant difference ($p\text{-value} = 0.22 > 0.05$).

Plots showing the relative errors at each distance for the VR condition can be seen on Figure 4, and the relative errors at each distance for the RW condition can be seen on Figure 5. The plots show a clear trend towards greater degrees of underestimation as the distance increases, which is both observed in VR and RW. Also noteworthy is that the underestimation gets slightly smaller in the final estimate (10.35m) potentially due to it being close to the physical end of the room used for testing.

The mean absolute error in the VR environment was found to be 12.26%, and 10.76% in the RW environment, suggesting that the estimates in the RW were more accurate than in VR, and the difference between the two conditions were found to be statistically significant ($p\text{-value} = 0.04 < 0.05$).

This study of *Blind Walking* as a method, establishes that the degrees of underestimation between VR and RW are fairly similar ($\sim 2\%$ difference), with less underestimation in RW than VR. The RW underestimation were in line with the prior research presented in Section 2, while the degree of underestimation in VR was also lower than what was found by prior research. As the method performed similarly to prior work in the field, with little difference between VR and RW, it was chosen to be utilized in the final experiment regarding the effect of room sizes on spatial perception in VR (described in Section 6).

6 ROOM SIZE EXPERIMENT

As a baseline for the relationship between Virtual Reality (VR) and the Real World (RW) was already established in Section 5, the rest of the experiments were conducted purely in VR.

To investigate the influence of room size on spatial perception three different rooms were designed: a small room ($\sim 15\text{m}^2$), a medium room ($\sim 35\text{m}^2$) and a large room ($\sim 95\text{m}^2$). The sizes of the rooms were based on the research described in Section 3.1 and were all chosen to have the same 3:5 aspect ratio. All the rooms give access to the personal and action spaces, while the vista space is never reached, which afforded the depth cues presented in Figure 2. The rooms can be seen side by side in Figure 1.

6.1 Participants

36 participants were recruited with normal (20/20) or corrected vision. Participants consisted of university students recruited by convenience. The average IPD of the participants was measured to be $\sim 65\text{mm}$.

6.2 Apparatus

The apparatus used for the *Room Size Experiment* was the same as for the *Blind Walking Baseline Experiment*, as presented in Section 5.2.

6.3 Procedure

The participants would enter the same large empty room as in the *Blind Walking Baseline Experiment* with an area of 130m^2 , where they were informed of the procedure of the experiment and filled out a consent form.

Participants then entered a transitional environment in Virtual Reality (VR), which resembled the large empty room they were present in. After a brief delay, they would then be transported to the first room of the experiment (either the small, medium or large one).

In each of the rooms of the experiment, participants would be presented with blue crosses on the ground which served as targets, and once they said that they felt ready to walk there without vision, they were asked to close their eyes and the screen would fade to black. This procedure would be repeated for 4 targets in the small room, 6 targets in the medium room, and 8 targets in the large room as indicated on Figure 6.

All targets and rooms were presented in random orders (based on a Latin square) for each participant, and since the targets 0.9m, 1.7m, 2.6m and 3.2m were presented to participants in both the small, medium and large rooms they would serve as the basis of later comparisons. We refer to these distances as the *universal distances*.

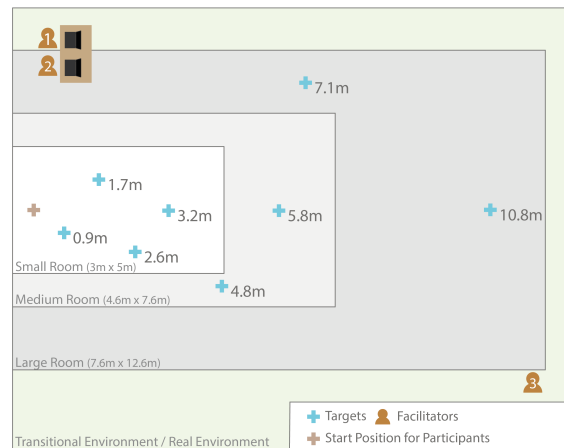


Figure 6: Visualization of the transitional environment setup with the three overlapping rooms and distances.

6.4 Results

The mean relative error for the *universal distances* was found to be -4% in the small room, -5% in the medium room, and -6% in the large room. This shows a consistent tendency towards underestimation of distances, regardless of the room size, which seems to grow slightly larger as the room size does. The relative error for *universal distances* was found to be normally distributed, and no statistically significant difference was found between the small and medium rooms ($p = 0.49 > 0.05$), however, a statistically significant difference was found between the small and large room ($p = 0.03 < 0.05$). A statistically significant difference was also found between the error estimates for the medium (8%) and large room (9%) ($p = 0.04 < 0.05$) when comparing all overlapping distances (*universal distances*, 4.8m and 5.8m).

The standard deviation for the *universal distances* was found to be 12% for the small room, 13% for the medium room and 11% for the large room. This indicates that the *precision* is not linearly influenced by the room size.

The relative error estimates for each distance in each room size are presented in Table 1. The table shows a general trend towards greater degrees of underestimation as the distance increases. It is also worth noting that the first distance, 0.9 meters, was consistently overestimated regardless of room size.

A statistically significant difference was found between all distance estimates of consecutive distances ($p < 0.05$), except for between 2.6m and 3.2m in the small room, 1.7m and 2.6m in the medium room, and 3.2m and 4.8m in the large room. In most cases the statistical significance shows growing underestimation, with the exception of the cases 3.2m to 4.8m in the medium room, and 5.8m and 7.1m in the large room.

Table 1: Relative error of each distance and room size in percentage. *Universal distances* are highlighted in green. Light blue shows overlapping distances between medium and large, and light purple are only for the large room.

Average Error in Estimates			
Distance (meters)	Small Room	Medium Room	Large Room
0.9	+4%	+3%	+3%
1.7	-4%	-4%	-6%
2.6	-7%	-7%	-11%
3.2	-11%	-12%	-11%
4.8		-10%	-11%
5.8		-16%	-20%
7.1			-14%
10.8			-19%

The mean absolute error, representing the *accuracy*, was found to be 10% for the small room, 11% for the medium room, and 11% for the large room. The small room showed the lowest absolute error, but no statistically significant difference was found between any of the room sizes (small-large [$p = 0.21 > 0.05$], small-medium [$p = 0.58 > 0.05$], medium-large [$p = 0.31 > 0.05$]). The absolute error data was found to be non-normally distributed.

7 DISCUSSION

The mean relative errors for the *universal distances* were found to be -4% in the small room, -5% in the medium room, and -6% in the large room. The large room was found to have statistically significantly different estimates from the small and medium rooms. However, this was not the case between the small and medium rooms. This observation indicates that room size affects participants' ability to perceive distances, with a tendency of larger rooms causing greater underestimations.

The small and medium rooms showed no statistically significant differences in participants' error degree, which suggests that the difference in size between the small and medium room has to be bigger for the effect to have a statistically significant influence on the results.

The results of the experiments also show a tendency of longer distances being underestimated more than shorter distances. This is shown by most consecutive distance estimates being statistically significantly different from each other, while also maintaining a trend of increasing underestimation.

In the *Blind Walking Baseline Experiment* as a method for quantifying perceived distances, the average error was roughly -9% for Virtual Reality (VR) and roughly -7% for Real World (RW). These degrees of underestimation are largely consistent with previous studies investigating spatial perception in VR, using *Blind Walking* as a method. Ahmed et al. (2010) found underestimations to be 10% and Witmer and Sadowski Jr. (1998) reported 15% underestimation in VR. The 7% underestimation found for the RW condition is also in line with previous research, especially insofar as being lower than the VR underestimations. Loyola (2018) found 10% underestimation, Witmer and Sadowski Jr. (1998) found 8% underestimation in RW.

Based on the results from the *Blind Walking Baseline Experiment*, it can also be theorized that the presence of a physical wall close to the furthest distance affected the estimations. As can be seen in Figures 4

and 5, in both conditions the accuracy decreased the longer the distance - until 10.35, which was the distance closest to the wall. A reason for this might have been some participants putting out their hands in front of them to try to feel the wall, which was a flaw of the experiment.

Similarly, in the *Room Size Experiment* we observed that estimation points close to the side walls had less underestimation, with 4.8m being the closest to a wall in the medium room and 7.1m in the large room as seen in Figure 6 than the furthest point, and the point before them. This was not seen in the small room, likely because the points were similarly spaced from the walls.

A noteworthy deviation from prior research is the advancement of VR technology in the recent years allowing the display of images in a higher resolution and at higher refresh rates. Another difference is that most previous studies used tethered HMDs for their experiments, while this study used an untethered HMD. This might be a reason for the lower degree of underestimations found in this study, since being untethered affords more freedom of movement.

7.1 Validity and Reliability Concerns

The implications of this study should be evaluated with caution due to some threats to the validity and reliability of the results. One of them being the fact that the experiments were carried out on a convenience sample of university students, opposed to a random sample. This could affect how adequately they represent the general population. Nevertheless, as mentioned in Section 6.4, the relative error data was normally distributed. Additionally, since most of the participants who took part in the experiments were students pursuing technology and design related degrees, it is also likely that some of them were experienced with VR. This could affect the results of the experiments, as it could translate into better performance due to being more used to acting in a virtual space.

Another potential threat affecting the results could be the sound cues caused by the movement of the facilitator, who ensured that participant's would not collide with walls, as it might have affected their spatial awareness.

Additionally, the distances used in all experiments were consistent across conditions, with the same groups of participants testing in all conditions per experiment. This could entail some carryover effects that could affect the results, such as practice and visual fatigue. To account for these effects, the experiments were designed with the use of Latin squares, randomizing the order of distances. To further pre-

vent visual fatigue, the experiments were designed to take up a short amount of time (on average 10 minutes per participant).

8 CONCLUSION

The aim of this study was two-fold: to identify a reliable method of quantifying perceived distances, as well as to use the identified method to investigate the effect of different sizes of Virtual Environments (VEs) on spatial perception in Virtual Reality (VR).

To investigate the reliability of the methods, they were compared between VR and the Real World (RW). *Blind Walking* was identified as a promising method, as it resulted in the lowest degree of underestimation in both VR (12%) and RW (12%) and it also performed the most similarly between RW and VR compared to verbal assessment and walk and assess. The method was further investigated in a large (130m²) empty room to establish a baseline for the difference between error estimation in VR and RW, where underestimation of distances were found to be ~9% in VR and ~7% in RW.

The effect of different VE sizes on spatial perception was then investigated using VR and three differently sized rooms: a small room (15m²), a medium room (35m²) and a large room (95m²). Underestimations were found to be 4% in the small room, 5% in the medium room, and 6% in the large room for the *universal distances* (distances between 0.9m and 3.2m) in the *Room Size Experiment*. This indicates that distances become more underestimated the larger the area of the VE. It is worth noting that no statistically significant difference in estimates was found between the small and medium rooms, however, the large room had a statistically significantly greater degree of underestimation than the small and medium rooms.

In general, the underestimations from the *Blind Walking Baseline Experiment* and *Room Size Experiment* are in line with previous literature as presented in Section 2. Additionally a tendency for underestimation to grow with the distance was found.

More importantly, the *Room Size Experiment* indicated that room size could influence spatial perception in VR, however, more research is needed to validate this and find the exact threshold where room size begins to have a clear effect.

In conclusion, this study contributes to a better understanding of spatial perception in VR, how it is affected by the size of the environment, as well as the performance of *Blind Walking* as a method of quantifying perceived distances.

REFERENCES

- Ahmed, F., Cohen, J. D., Binder, K. S., and Fennema, C. L. (2010). Influence of tactile feedback and presence on egocentric distance perception in virtual environments. In *2010 IEEE Virtual Reality Conference (VR)*, pages 195–202. IEEE.
- Armbrüster, C., Wolter, M., Kuhlen, T., Spijkers, W., and Fimm, B. (2008). Depth perception in virtual reality: Distance estimations in peri- and extrapersonal space. *CyberPsychology & Behavior*, 11(1):9–15. PMID: 18275307.
- Cutting, J. E. and Vishton, P. M. (1995). Chapter 3 - perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth*. In Epstein, W. and Rogers, S., editors, *Perception of Space and Motion*, Handbook of Perception and Cognition, pages 69–117. Academic Press, San Diego.
- Gagnon, H. C., Buck, L., Smith, T. N., Narasimham, G., Stefanucci, J., Creem-Regehr, S. H., and Bodenheimer, B. (2020). Far distance estimation in mixed reality. In *ACM Symposium on Applied Perception 2020, SAP '20*, New York, NY, USA. Association for Computing Machinery.
- Interrante, V., Ries, B., and Anderson, L. (2006). Distance perception in immersive virtual environments, revisited. In *Proceedings of the IEEE Conference on Virtual Reality*, VR '06, page 3–10, USA. IEEE Computer Society.
- Kelly, J. W., Cherep, L. A., Klesel, B., Siegel, Z. D., and George, S. (2018). Comparison of two methods for improving distance perception in virtual reality. *ACM Trans. Appl. Percept.*, 15(2).
- Kelly, J. W., Cherep, L. A., and Siegel, Z. D. (2017). Perceived space in the htc vive. *ACM Trans. Appl. Percept.*, 15(1).
- Klein, E., Swan, J. E., Schmidt, G. S., Livingston, M. A., and Staadt, O. G. (2009). Measurement protocols for medium-field distance perception in large-screen immersive displays. In *2009 IEEE Virtual Reality Conference*, pages 107–113.
- Li, B., Zhang, R., Nordman, A., and Kuhl, S. A. (2015). The effects of minification and display field of view on distance judgments in real and hmd-based environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, SAP '15, page 55–58, New York, NY, USA. Association for Computing Machinery.
- Loyola, M. (2018). The influence of the availability of visual cues on the accurate perception of spatial dimensions in architectural virtual environments. *Virtual Reality*, 22:235–243.
- Lucaci, A., Bach, M., Jensen, P., and Madsen, C. (2022). Influence of texture fidelity on spatial perception in virtual reality. *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - GRAPP*, 17:244–251.
- Maruhn, P., Schneider, S., and Bengler, K. (2019). Measuring egocentric distance perception in virtual reality: Influence of methodologies, locomotion and translation gains. *PLOS ONE*, 14(10):1–24.
- Messing, R. and Durgin, F. H. (2005). Distance perception and the visual horizon in head-mounted displays. *ACM Trans. Appl. Percept.*, 2(3):234–250.
- Moscato, C., Chamilothon, K., Wienold, J., Andersen, M., and Matusiak, B. (2021). Window size effects on subjective impressions of daylight spaces: Indoor studies at high latitudes using virtual reality. *LEUKOS*, 17(3):242–264.
- Peer, A. and Ponto, K. (2017). Evaluating perceived distance measures in room-scale spaces using consumer-grade head mounted displays. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 83–86.
- Proffitt, D. R. and Caudek, C. (2012). Depth perception and the perception of events. In *Handbook of Psychology, Experimental Psychology*, pages 212–235. John Wiley & Sons, Incorporated.
- Renner, R. S., Velichkovsky, B. M., and Helmert, J. R. (2013). The perception of egocentric distances in virtual environments - a review. *ACM Comput. Surv.*, 46(2).
- Simpson, G., Sinnis-Bourozikas, A., Zhao, M., Aseeri, S., and Interrante, V. (2018). A virtual reality investigation of the impact of wallpaper pattern scale on qualitative spaciousness judgments and action-based measures of room size perception. In Bourdot, P., Cobb, S., Interrante, V., kato, H., and Stricker, D., editors, *Virtual Reality and Augmented Reality*, pages 161–176, Cham. Springer International Publishing.
- Thompson, W. B., Willemsen, P., Gooch, A. A., Creem-Regehr, S. H., Loomis, J. M., and Beall, A. C. (2004). Does the quality of the computer graphics matter when judging distances in visually immersive environments.
- von Castell, C., Hecht, H., and Oberfeld, D. (2018). Bright paint makes interior-space surfaces appear farther away. *PLOS ONE*, 13(9):1–15.
- Witmer, B. G. and Sadowski Jr., B. G. J. (1998). Nonvisually guided locomotion to a previously viewed target in real and virtual environments. *Human Factors*, 40(3):478–488.
- Zhao, M., Sinnis-Bourozikas, A., and Interrante, V. (2019). Exploring the use of immersive virtual reality to assess the impact of outdoor views on the perceived size and spaciousness of architectural interiors. In Bourdot, P., Interrante, V., Nedel, L., Magnenat-Thalmann, N., and Zachmann, G., editors, *Virtual Reality and Augmented Reality*, pages 312–319, Cham. Springer International Publishing.

An Analysis of Correlations between Empathy and Both EEG and HEG during Text Chat

Masaki Omata¹ and Kana Watanabe²

¹Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu, Yamanashi, Japan

²Department of Computer Science and Engineering, University of Yamanashi, Kofu, Yamanashi, Japan
{omata, nabe19}@hci.media.yamanashi.ac.jp

Keywords: Text chat, Emotion, Electroencephalography, Hemoencephalography.

Abstract: We have addressed a problem that emotions associated with texts are not correctly conveyed in text chat. In this study, we conducted two experiments to analyze whether electroencephalography (EEG) and hemoencephalography (HEG) of a receiver can be used to identify the receiver's empathic state when the receiver is empathizing with the emotion associated with the text sent by a sender. As the results, we found that emotional valence was more likely to be empathized with in text-based chat, but emotional arousal was less likely to be empathized with. We also found that the power of theta waves at O1 (in the occipital region) of the empathic receivers was significantly lower than that of the non-empathic receivers.

1 INTRODUCTION

People communicate using a variety of modalities such as text, voice, and video with increasing popularity of smartphones and social networking services (IICP of MIC, 2019). Among the modalities, text-based communication, such as chat and email, has an advantage of being less constrained in time and space, and is easier to communicate anytime and anywhere. On the other hand, a disadvantage of text-based communication is that it lacks visual information (such as facial expressions and body language) and auditory information (such as tone and volume of voice) compared to voice, video, and face-to-face communication, and thus may not correctly convey emotions to a receiver. If the emotions are not conveyed correctly to the receiver, there is a possibility that the receiver will not understand the emotions and will not sympathize with them.

To address the problem, we have proposed use of physiological signals of the receiver to estimate the receiver's empathy with regard to emotions of conversational content in the text. The reason for using physiological signals is that physiological signals can be measured continuously and unconsciously, and can be used without burdening users with interruptions or interventions during a text conversation. If our proposed system is able to estimate empathy from physiological signals, it will

be able to feed back the receiver's empathy to the sender immediately during text chatting, which will make communication smoother in the future.

In this paper, we introduce related studies and indicate the position of our study in the next section. After that, we explain the first experiment, which examines correlations between the empathizer's emotions and both electroencephalography (EEG) (Nunez et al., 2007) and hemoencephalography (HEG) (Tinius, 2004) during text chatting. Then, we explain the second experiment, in which we added data based on the results of the first experiment. Our contribution in this paper is that we mentioned the possibility of using the theta waves of the EEG on the occipital region of the empathizers (receivers) to estimate their empathies generated from the text chat.

2 RELATED WORK

Kinoshita et al. measured affective sharing from EEG signals and conducted an experiment in which participants communicated using facial expressions of joy, sadness, and neutrality (Kinoshita et al., 2019). The results showed that correlations of EEG powers were significantly higher under the high affective sharing condition compared to the low affective sharing condition, and that the correlations of the EEG powers were significantly higher under the joy

and sadness conditions in the alpha-mu band. The results have suggested possibility to measure affective sharing in response to emotional faces from the correlation of EEG powers.

Vanderhaegen et al. studied synchronization between dynamic events with heartbeats on non-conscious errors in the control of dynamic events by comparing two groups of subjects: a group for which alarms were activated synchronously with the current heart rate of the subjects and a group for which the alarms were activated without being synchronized with the current heart rates of the subjects (Vanderhaegen et al., 2020). The results showed that there was a significant impact of such a synchronization of events with heartbeat.

Jain et al. looked at and analyzed how messages sent on instant messages or posts on social networks are interpreted by readers in terms of the emotional state of the sender (Jain et al., 2016). As the results of Spearman's rank-correlation coefficient about the Self-Assessment Manikin (SAM) (Bradley et al., 1994), they found that a high correlation ($\rho = 0.80$) was found between valence of the sender for each message and perceived valence of the sender, and that no correlation ($\rho = 0.11$) was found between arousal of the sender for each message and perceived arousal of the sender. These indicated how easily and accurately valence is conveyed while arousal is almost never conveyed accurately.

Ghosh et al. designed and implemented an Android application TapSense which traced smartphone typing and records self-reported emotion state and conducted an online survey to understand the typing habits in smartphones and collect feedback on multiple measurable parameters that affect their emotion while typing (Ghosh et al., 2017). As the results, they observed that using only typing features, it was possible to identify four emotion states (happy, sad, stressed and relaxed) with an average accuracy of 73% and a maximum of 94%.

We believe that if it is found that physiological signals can be used to estimate the receiver's empathy as in Kinoshita et al. (Kinoshita et al., 2019) in text chatting, which is our research target, we will be able to provide a user interface that supports smoother communication with emotions in text chatting than before. Specifically, we have been designing a user interface that sequentially estimates the receiver's empathy from the physiological signals and provides feedback to the sender on a receiver's level of the empathy, and a user interface that automatically decorates text according to the receiver's level of empathy. The reason for using physiological signals in this way is that physiological signals are not as

subjectively controllable by the user intentionally as questionnaire surveys, can be measured continuously, and do not interfere with the user's operation. For this purpose, in this paper, we asked senders and receivers to answer their emotions in text chat using SAM, as in Jain et al. and then recorded the receivers' EEG and HEG, as in Kinoshita et al. (Kinoshita et al., 2019) and analyzed the correlation between the receivers' emotions and the physiological signals.

3 EXPERIMENT FOR ANALYSIS OF PHYSIOLOGICAL SIGNALS DURING EMPATHY

In this experiment, we analyzed correlation between physiological signals and empathy for chat text by recording physiological signals of receivers during text chat and comparing them with their normal conditions.

3.1 Physiological Signals

Electroencephalography (EEG) and hemoencephalography (HEG) were measured in this experiment. The EEG and HEG sensors were connected to an encoder (Thought Technology Ltd., ProComp INFINITI) (Thought Technology Ltd., 2022). In this section, the properties of the physiological signals are described.

3.1.1 EEG

EEG is a record of the oscillations of brain electric potentials recorded from electrodes attached to the human scalp (Nunez et al., 2007). The frequency ranges are categorized as delta (0.5 to 3 Hz), theta (4 to 7 Hz), alpha (8 to 13 Hz) and beta (14 to 30 Hz). Very high frequencies (typically over 40 Hz) are referred to as gamma activity. In general, theta waves are seen in deep meditation and slumber, alpha waves are seen in relaxation, and beta waves are seen in daily life, tension, and mental excitement. The power values of theta, alpha, and beta waves are used as the physiological indexes of this experiment. The BioGraph INFINITI and BioGraph Infiniti Software Platform of Thought Technology Ltd. were used to calculate the power values.

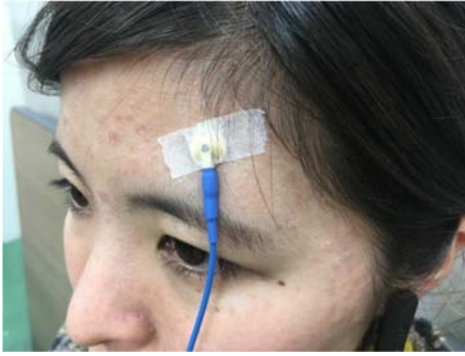


Figure 1: The EEG sensor and its installation.



Figure 2: The HEG sensor and its installation.

The EEG-Z sensor from Thought Technology, shown in Figure 1, was used as the electrodes for the EEG measurement. Three probe electrodes were placed at Fp1 (left front polar), O1 (left occipital) and O2 (right occipital) according to the International 10–20 system (American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature, 1991).

3.1.2 HEG

A HEG is relative ratio of oxidized hemoglobin to deoxygenated hemoglobin with blood flow dynamics and cellular metabolism in localized parts of the brain cortex (Tinius, 2004). The measurements are closely linked with brain activation due to the phenomenon of neurovascular coupling. The HEG ratio (Serra-Sala et al., 2012 and Skalski et al., 2021), calculated from the increase or decrease of oxidized hemoglobin and deoxidized hemoglobin in the blood flow to the location of the frontal cortex, was used as a physiological index in our study.

We used MediTechElectronic's HEG-Sensor shown in Figure 2 in order to measure the ratio. The measurement point was Fp2 (right forehead) of the International 10–20 system.

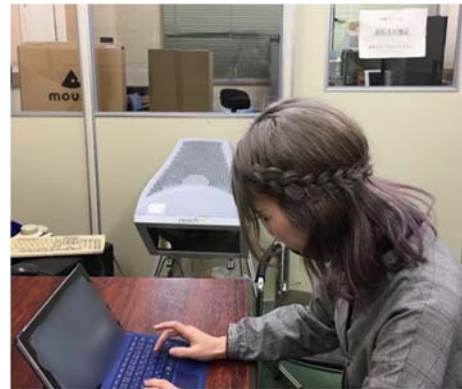
3.1.3 Standardization

We standardize the recorded physiological indexes by referring to the method of Omata et al (Omata et al., 2014) to analyze the variation of the indexes during an experimental task from the indexes in resting state. EEG and HEG-ratio are standardized as shown in Equation (1),

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where X is the data of each physiological index during the experiment, μ is the mean value of the normal state, and σ is the standard deviation.

3.2 Environment and Participants' Roles



(a)



(b)

Figure 3: An emotional generator (a) and an empathizer (b) during the experiment task.

As shown in Figures 3a and b, two of the participants chatted using laptop computers in two rooms separated by a barrier that prevented them from directly seeing and hearing each other. We chose slack (Slack Technologies, 2022) as the chat system for this experiment because it is intuitive and easy to understand for the participants, and because it is easy

for the participants to reproduce emotional expressions more richly by using pictorial reactions (Slack Technologies, 2022).

The two participants were asked to chat with each other mainly using text, and each was assigned a role for this experiment. One of the two was assigned the role of transmitting emotional experiences through the chat (hereinafter called the “emotion generator”), and the other was assigned the role of empathizing with the received emotional experiences (hereinafter called the “empathizer”). The empathizer was equipped with the physiological signal sensors as described above.

3.3 Procedure and Task

We asked the participants to participate in the chat experiment in pairs (emotion generator and empathizer). First, we gave informed consent to these two participants and obtained their consent to participate in the experiment. After that, we took them to separate rooms, sat them in front of a laptop computer for chatting, and explained how to use slack.

Before chatting, the emotion generator was asked to recall his/her own happy and sad experiences, and to rate the experiences on 9-point scale of emotion valence and 9-point scale of arousal of the Self-Assessment Manikin (SAM) (Bradley et al., 1994). On the other hand, the empathizer was equipped with physiological signal sensors. Then, the physiological signals were recorded for one minute while the sympathizer was resting and doing nothing, to be used as baseline data for analyses.

The experimental task for both the participants was to have a chat conversation about the happy and sad events that the emotion generator remembered. Before starting the chat, both the participants were asked to familiarize themselves with slack for two minutes (or more if either of the two requested it). At the start of the chat task, via the chat system, the experimenter instructed the emotion generator “Please tell the other participant about the experience you recalled and try to convey the emotion of the experience as much as possible,” and instructed to the empathizer “Please be a collocutor of the other participant’s story and try to capture the emotion of the story.” After the experimenter signaled the start of the chat, both the participants chatted for about three to five minutes based on the instructions. During the chat, the physiological signals of the empathizer were continuously recorded. The experimenter read the content of the chat, and signaled the end of the chat when the conversation was finished. Then, the

physiological signals of the empathizer were recorded again for one minute while she was at rest doing nothing.

After the chat was over, the empathizer was asked to answer the emotion she recognized from the chat text and the emotion she felt as a result of the conversation by using the SAM. After a five-minute break, the participants performed the same task for the other experience that the emotion generator recalled. The number of pairs that started with happy experiences was equal to the number of pairs that started with sad experiences in order to counterbalance the order of happy and sad experiences in the chat contents.

After completing the chat task for the two emotions, both the participants were asked to complete a questionnaire survey about their past experience in using slack and their impressions of using slack to express the emotions in this experiment.

Eight participants (one male and seven females, ranging in age from 21 to 23 years) participated in this experiment. All of them had experience using slack before the experiment. When creating pair combinations from the participants, we paired them with the pairs that they usually communicate with via text chat. The empathizers in each pair were all female. The reason for this was based on the results of Davis’s study that women were more likely to empathize than men (Davis, 1980).

3.4 Results

3.4.1 Correlation Analysis of Subjective Emotions

Figure 4 shows the relationship between the emotional valence (from negative (1) to positive (9)), that the emotion generators responded about the experiments they recalled and the emotional valence (from negative (1) to positive (9)) that the empathizers felt from the chat. Since the participants were asked to chat about happy and sad events, the plot points were divided into two groups. The Spearman correlation coefficient for the plots is 0.948. This indicates that both emotion generators and empathizers had similar emotional valences.

Figure 5 shows the relationship between the arousal (from low (1) to high (9)) that the emotion generators responded about the events they recalled and the arousal (from low (1) to high (9)) that the empathizers felt from the chat. Since the Spearman correlation coefficient for the plots is 0.621, there was no correlation like that of the emotion valence, and there

were differences in the arousal levels held by emotion generators and empathizers.

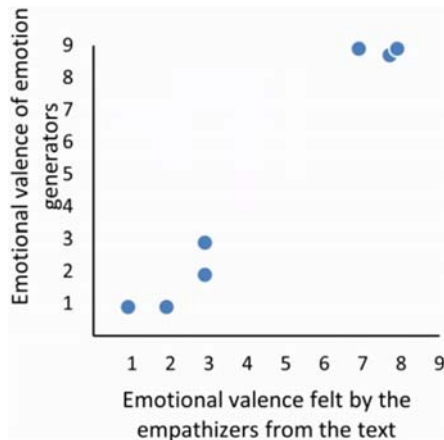


Figure 4: The relationship between the emotional valence (from negative (1) to positive (9)) of the emotion generators and the emotional valence (from negative (1) to positive (9)) felt by the empathizers from the text chat.

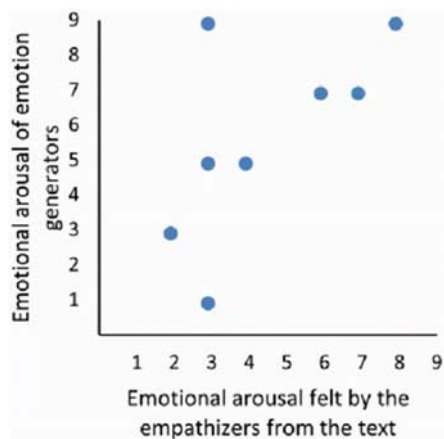


Figure 5: The relationship between the emotional arousal (from low (1) to high (9)) of the emotion generators and the emotional arousal (from low (1) to high (9)) felt by the empathizers from the text.

3.4.2 Correlation Analysis between Physiological Signals and Subjective Emotions

The power values of the three frequency bands (alpha, beta, and theta) in the EEG at three locations (Fp1, O1, and O2) and the HEG ratio at one location (Fp2) were analyzed for significant differences among the three states: resting state before the task, happy empathy, and sad empathy. As the results, there was a significant difference ($p < .05$) between the resting state and the happy empathy state in the power values

of theta waves measured at O1 and O2. There was no significant difference ($p < .05$) in the HEG ratio.

Figures 6 and 7 show the correlations between the power values of theta waves at O1 and O2 and the subjective emotional valence of the empathizers during resting and happy empathy, respectively. Here, the empathizer's valence at rest is set to 5, which represents a neutral emotional state. Therefore,

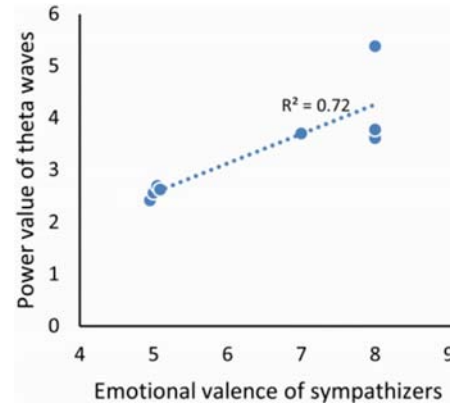


Figure 6: Relationship between the emotional valence of the empathizers and the power value of the theta waves of O1.

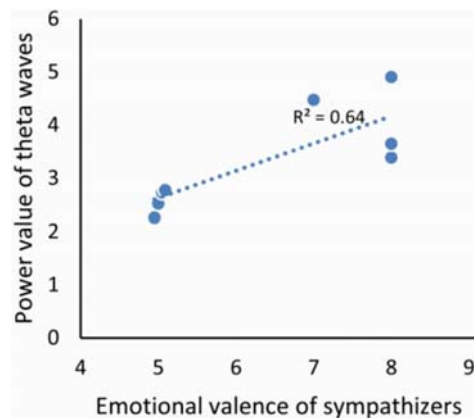


Figure 7: Relationship between the emotional valence of the empathizers and the power value of the theta waves of O2.

these graphs show that the power value of the theta waves increases when the neutral emotional valence at rest is positively changed by empathy for happiness. The R^2 in each of the graphs is its coefficient of determination.

3.5 Discussions

Based on the results of SAM, we found that emotional valence is easily conveyed but arousal is not easily conveyed in text-based chat. This is consistent with the results of the study by Jain et al (Jain et al., 2016). From the analysis of physiological signals, we found

that the power value of the theta wave band of the occipital EEG increased during empathy for chatting about happy events. This result is in line with the results of Knyazev's study that theta waves are related to emotion regulation (Knyazev, 2007). Therefore, we argue that occipital theta waves can be used to estimate a state of positive empathy of chat users toward the chat contents.

4 FURTHER EXPERIMENT

We conducted another experiment to add more participants' data to the aforementioned experiment. The environment, the roles of participants, the procedure, and the task were the same as in the aforementioned experiment. However, this experiment differs from the previous one in the following points.

- The results of SAM in the aforementioned experiment suggest that some participants remembered "Angry" and "Afraid" in Russell's circle model (Russell, 1980). However, we wanted them to remember Sad, so we instructed participants to recall a sad experience, not a bad experience.
- We asked participants to recall the recent experiments in order to generate the arousal level more accurately.
- Men were included in empathizers.

4.1 Results

4.1.1 Correlation Analysis of Subjective Emotions

Figure 8 and 9 show the relationship between the emotional valence (from low (1) to high (9)) and arousal (from low (1) to high (9)) that the emotion generators responded about the experiments they recalled and those that the empathizers felt from the chat. From the plots of the graphs, it can be seen that there is a highly positive correlation (the Spearman correlation coefficient is 0.806.) between the emotional valences of both the roles as the results of Section 3. On the other hand, there is low correlation (the Spearman correlation coefficient is 0.658.) between the emotional arousals of both the roles, although the correlation was slightly stronger due to the difference in instruction.



Figure 8: The relationship between the emotional valences (from low (1) to high (9)) of the emotion generators and the emotional valences (from low (1) to high (9)) felt by the empathizers from the text in the further experiment.

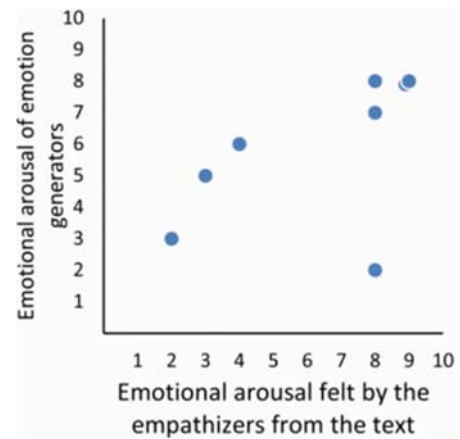


Figure 9: The relationship between the emotional arousal (from low (1) to high (9)) of the emotion generators and the emotional arousal (from low (1) to high (9)) felt by the empathizers from the text in the further experiment.

4.1.2 Correlation Analysis between Physiological Signals and Subjective Emotions

The power values of the three frequency bands (alpha, beta, and theta) in the EEG at three locations (Fp1, O1, and O2) and the HEG ratio at one location (Fp2) were analyzed for significant differences among the three states: resting state before the task, happy empathy, and sad empathy. As the results, there were no significant differences ($p < .05$) in the EEG and the HEG ratio. Although there was no significant difference, the power values of theta waves at O1 and O2 were higher than those at rest in 7 out of 8 trials as in the aforementioned experiment.

4.1.3 Integrated Analysis of Data from Two Experiments

We integrated the data from all 16 trials, including the experiment in Section 3 and this experiment, and divided them into two groups. Specifically, the seven trials in which the sum of the absolute values of the differences between the emotion generators' responses and the empathizers' responses from the text and their own emotions was less than or equal to 3 were classified as empathizable group, and the nine trials as the rest were classified as non-empathizable group, based on the results of the SAM of the participants of the two experiments.

Figure 10 and Figure 11 show the power values of theta waves at O1 and O2 for each of the two groups. The results show that the power values at O1 in the empathizable group was significantly lower than those in the non-empathizable group ($p < 0.05$). There was no significant difference about O2.

4.2 Discussions

From the results of responses to SAM in the two experiments, we found that emotional valence was more likely to be empathized with, even in text-based chat. On the other hand, when we instructed constraints on the degree of arousal in the further experiment, which became easier to empathize to some extent, but in general of the two experiments, arousal was not easily empathized with in text-based chat. We believe that the reason for this is that in text-based chat, there are many words to express positive and negative emotions, but not many words to express level of arousal.

One of the reasons for the lack of significant differences in physiological signals between the resting state and the on-task state in the further experiment may be that, unlike the aforementioned experiment, the data included data from males, but since the number of data is insufficient, further additional experiments are necessary in the future.

Since the power of theta waves at O1 in the empathizers, who were similar to the emotional valence and arousal of the emotion generators, we believe that the theta waves in the occipital region can provide data for estimating that emotion generators and empathizers have similar emotions. However, since it is not a simple correlation that the power value increases when the emotions are similar, it is necessary to conduct further experiments to analyze the relationship between the value and the degree of empathy in more detail.

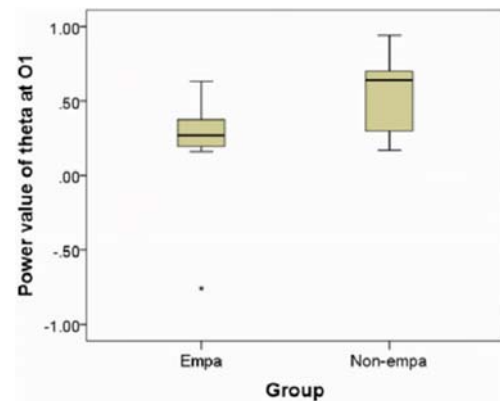


Figure 10: The power value of theta wave at O1 in two groups of emotion similarity difference.

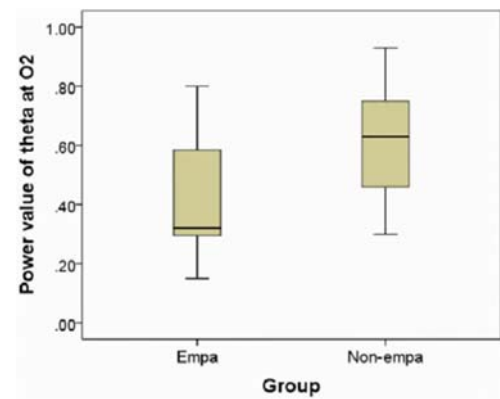


Figure 11: The power value of theta wave at O2 in two groups of emotion similarity difference.

5 CONCLUSIONS

We conclude from the two experiments that the power of theta waves in the occipital region is higher during empathy for the content of a text chat than at rest, but lower when empathizing with the same emotion of an emotion generator. Moreover, we find that we need more data to analyze the differences in the measurement points of physiological signals, the relationship between EEG and HEG, and the individual differences among participants. In addition, we believe that further analysis of the content of the text chat, the degree of empathy, and the fluctuations of the physiological signals during the text chat will show possibility of using the physiological signals in more detail.

REFERENCES

- American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. (1991). *Journal of Clinical Neurophysiology*. Vol. 8, No. 2, 200–202.
- Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, Vol. 25, No. 1, 49–59. [https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9).
- Davis, M. (1980). A Multidimensional Approach to Individual Differences in Empathy. *Journal of Personality and Social Psychology Catalog of Selected Documents in Psychology*, 10, 85.
- Ghosh, S., Ganguly, N., Mitra, B., and De, P. (2017). Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 146–151.
- Jain, M., Seshagiri, S., and Chopra, S. (2016). How do I communicate my emotions on SNS and IMs? In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI'16)*. Association for Computing Machinery, New York, NY, USA, 767–774. DOI:<https://doi.org/10.1145/2957265.2961862>.
- Kinoshita, T., Tanaka, H., Yoshino, K., and Nakamura, S. (2019). Measuring Affective Sharing between Two People by EEG Hyperscanning. In *Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI'19)*. Association for Computing Machinery, New York, NY, USA, Article 3, 1–6. DOI:<https://doi.org/10.1145/3351529.3360657>.
- Knyazev, G. G. (2007). Motivation, Emotion, and Their Inhibitory Control Mirrored in Brain Oscillations. *Neuroscience & Biobehavioral Reviews*, Vol. 31, Issue 3, 377–395.
- Nunez, P. L. and Srinivasan, R. (2007). *Electroencephalogram*. *Scholarpedia*, Vol. 2, No. 2:1348. Retrieved Feb. 15, 2022 from <http://www.scholarpedia.org/article/Electroencephalogram>.
- Omata, M., Kanuka, D., and Mao, X. (2014). Experiments for Emotion Estimation from Biological Signals and Its Application. *Transactions on Computational Science XXIII Lecture Notes in Computer Science* Vol. 8490, 178–198.
- Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39, 1161–1178. DOI:<https://doi.org/10.1037/h0077714>.
- Serra-Sala, M., Timoneda Gallart, C, and Pérez-Alvarez, F. (2012). Evaluating Prefrontal Activation and Its Relationship with Cognitive and Emotional Processes by Means of Hemoencephalography (HEG). *Journal of Neurotherapy*. 16. 183–195. DOI:<https://doi.org/10.1080/10874208.2012.705754>.
- Skalski, S., Pochwatko, G., and Balas, R. (2021). Effect of HEG biofeedback on selected cognitive functions—Randomized study in children with ADHD and neurotypical children. *Infant and Child Development*, vol. 30, No. 4, e2242. DOI:<https://doi.org/10.1002/icd.2242>.
- Slack Technologies. *Slack Features*. Retrieved Feb. 15, 2022 from <https://slack.com/features>.
- Slack Technologies, *Use emoji and reactions*. Retrieved Feb. 15, 2022 from <https://slack.com/help/articles/202931348-Use-emoji-and-reactions>.
- The Institute for Information and Communications Policy of the Ministry of Internal Affairs and Communications (IICP of MIC) (2019). *FY2018 Survey Report on Usage Time of Information and Communications Media and Information Behavior*. Retrieved February 15, 2022 from https://www.soumu.go.jp/iicp/research/results/media_usage-time.html.
- Thought Technology Ltd.. *ProComp Infiniti System*. Retrieved February 15, 2022 from <https://thoughttechnology.com/procomp-infiniti-system-w-biograph-infiniti-software-t7500m/>.
- Tinius, T. (2004). *New Developments in Blood Flow Hemoencephalography*, Hawthorne Press.
- Vanderhaegen, F., Wolff, M., Mollard, R. (2020). Non-conscious errors in the control of dynamic events synchronized with heartbeats: A new challenge for human reliability study. *Safety Science*, Vol. 129, 104814. DOI: <https://doi.org/10.1016/j.ssci.2020.104814>.

UX- for Smart-PSS: Towards a Context-aware Framework

Angela Carrera-Rivera¹^a, Felix Larrinaga¹^b, Ganix Lasa²^c and Giovanna Martinez-Arellano³^d

¹Faculty of Engineering, Mondragon Unibertsitatea, Loramendi 4, 20500 Arrasate, Spain

²Design Innovation Center (DBZ), Mondragon Unibertsitatea, Loramendi 4, 20500 Arrasate, Spain

³Faculty of Engineering, University of Nottingham, Nottingham NG7 2RD, U.K.

{aicarrera, flarrinaga, glasa}@mondragon.edu, giovanna.martinezarellano@nottingham.ac.uk

Keywords: S-PSS, Context-awareness, User Experience, Recommendation System.

Abstract: Smart-product service systems are a business strategy that combines product and service into one value proposition. The user experience of digital services and the smart product can be a clear differentiator among competitors to achieve economically sustainable solutions. Hence, offering a more personalized experience is an important aspect of S-PSS. This paper aims to provide a theoretical framework for a context-aware user experience in S-PSS by providing adaptive and personalized services to the users according to their needs in a given context, by exploiting the digital capabilities of smart products and referring to the use of recommendation systems. The paper presents an application scenario using a smart-wearable as an example of a product-oriented PSS to better describe the framework and each component while stating the future challenges.

1 INTRODUCTION

Smart-Product Service Systems (S-PSS) is an ecosystem that proposes the combination of both product and service into one value proposition. The design of S-PSS has received increased attention in recent years (Carrera-Rivera et al., 2022; Cong et al., 2020; Dou and Qin, 2017) and implies several challenges. From the service perspective, S-PSS offers digital services that can be product independent or product dependent, powered by the integrated sensors in smart connected products or by technologies like digital twin or augmented reality (AR) (Zheng et al., 2019b). Zheng et al. (2019b) defined S-PSS as “an IT-driven value co-creation business strategy consisting of various stakeholders as the players, intelligent systems as the infrastructure, smart, connected products as the media and tools, and their generated e-services as the key values delivered that continuously strives to meet individual customer needs in a sustainable manner”. However, the design of S-PSS in multiple studies is only limited to the early phases of the product development and they do not address the evolution and adaptability that S-PSS should have.

S-PSS follows the Service-Dominant (S-D) logic, a fundamental pillar of Service Design, that characterizes for a “value co-creation” view that uses experiences, context and multi-stakeholder participation to create innovative business value propositions (Wetter-Edman et al., 2014). Users may have different roles as co-creators in the different phases of S-PSS lifecycle: co-ideators, co-innovators, co-evaluators, co-testers or experience creators (Pezzotta et al., 2017). Users as *experience creators* can be involved in the solution conceptualization supporting better customization. Pezzotta et al. (2017) stated that providers can generate richer experiences for customers by understanding their preferences. A deep understanding of the customer experiences could help companies in defining and re-defining better value propositions. User Experience (UX) can be a clear differentiator among competitors to achieve economically sustainable PSS. Cong et al. (2020) highlighted that “user preferences should be associated with different design elements of Smart PSS in specific usage contexts”. Previous studies have focused their attention on the connection between some user preferences and design elements. Misaka and Aoyama (2018) use Knowledge Engineering and machine learning for correlating users based on KANSEI engineering, a methodology to transform customer feelings into design as-

^a <https://orcid.org/0000-0001-8593-5961>

^b <https://orcid.org/0000-0003-1971-0048>

^c <https://orcid.org/0000-0002-2424-5526>

^d <https://orcid.org/0000-0003-3105-4151>

pects of a new product. However, the studies in this research area are still limited to user-specific preferences in the usage stage of Smart PSS. A current challenge is a way to collect and process user behaviour data in real time and define how it will be used. Existing research focused primarily on customer data from surveys or data from online reviews on social media. Further work on adaptation of the means of interaction with S-PSS needs to be done in order to improve the user experience to meet individual user needs. This paper proposes a framework to provide an adaptive UX according to user context, considering the interactions and data from devices and user profiles. The structure of this work is as follows: Section 2 starts with the fundamental background related to UX and Context-Awareness. Section 3 presents the framework and provides a detailed explanation of each phase. Section 4 provides a proof of concept of the framework using a wearable S-PSS. Finally, Section 5 presents the description of challenges, future work and conclusion of this work.

2 BACKGROUND

UX is a result of the internal state of the user, the characteristics of the designed entity and the **context features** where the interaction occurs (Hassenzahl and Tractinsky, 2006). According to Valencia et al. (2015), in S-PSS, the UX is characterized by feelings of *customer empowerment, the individualization of services, the sense of ownership, and an individual and shared experience*. Non-engineering approaches and qualitative tools (i.e., observation, interviews, experience maps, etc.) are broadly used by UX designers at the very early stage of design for user research, to understand the context of use and users involved. Chang et al. (2019) presented a case study for a smart pillbox directed at the elderly; using behavioural analysis and interviews they perform a user analysis not only to capture their specific requirements but also to understand the physiological and cognitive dimensions. Similarly, Jia et al. (2021) used behavioural analysis and experience maps for the development of a smart rehabilitation assisting device. In both cases, current users' problems, popularly denominated 'pain points' usually highlight service opportunities.

However, there is a need to exploit the digital capabilities of S-PSS to generate an adaptive UX with the collected data available. A data-driven UX is a particular characteristic of S-PSS where data from smart products and digital services is easily accessible (Carrera-Rivera et al., 2022). Furthermore, the ability to capture user-generated data in real-time should be

an important part of the design process, and it is a way of identifying and evaluating users' needs. For instance, methods to collect and process real-time user behaviour data and define in what ways it will be used. Existing research mainly focuses on online reviews data on social media or customer data from surveys. Wang et al. (2019) used user ratings and comments related to smart bicycles service from a website, and with the use of Natural Language Processing (NLP) techniques were able to capture implicit requirements that users had. Similarly, Mourtzis et al. (2018) presented a framework to evaluate the PSS services using the feedback received from shop floor experts and business customers through social media platforms. They also considered the information from machines and manufacturing execution systems to obtain KPI values. Therefore, user behaviour data from devices and interactions with service applications are typically not discussed.

In S-PSS, the delivery of personalized services can have a great impact on UX to affect the level of user satisfaction. Multiple approaches beyond PSS have used adaptive user interfaces as a way to offer personalization. Reguera-Bakhache et al. (2021) captured the interaction sequences, consisting of a series of click events from the interface of a mixing machine to understand the patterns of interaction from operators through a clustering analysis, and in this way reduce the time and interaction sequences of operators to perform a task. Todi et al. (2021) presented a system for adaptive menus using the data from user clicks on menu items and reinforcement learning to approximate menus to the user's expertise and interest. However, the use of context can further improve the personalization and adaptation by capturing the real needs of the user at a given time.

2.1 Context-awareness

Context is defined by Abowd et al. (1999) as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves". In UX, the context of use represents the users, tasks, equipment (i.e. hardware, software), and the physical and social environments in which S-PSS is used (Iso, 2010).

The term "context-aware" was defined by Schilit and Theimer (1994) as the "ability to discover and react to changes in the environment". In general, a context-aware system uses context to deliver information and/or services relevant to a user's task (Abowd et al., 1999). Thus, a context-aware system does

not necessarily imply automatization or real-time processing, instead, it refers to the ability to respond to context. For instance, context can be used to resolve what services or information need to be presented to the user. Therefore, context-awareness capability can be used to personalize the UX by providing services for a specific user situation.

In the design of S-PSS, multiple context-aware approaches have been proposed for the requirement-elicitation phase. For instance, Wang et al. (2019, 2021) proposed a graph-based context-aware framework for the requirement elicitation process in a data-driven manner. The framework includes a knowledge management layer that uses domain ontologies to model the product's components, services, and context; and a requirement elicitation layer, which uses graph algorithms to discover implicit requirements that user's had by analyzing reviews and comments from social media. For UX, context-awareness has been used to create realistic prototypes. For instance, Seo et al. (2016) proposed a hybrid evaluation method for smart home services based on VR and AR prototypes using a UX ontology for smart home services and semantic reasoning using the open source framework Jena. On the other hand, Dou and Qin (2017) obtained physiological data from several sensors including electrocardiography (ECG) and facial electromyography (fEMG) from elderly people combined with environmental information, to build a user mental model to better understand the experience of a smart TV S-PSS. The authors used clustering analysis as a way to identify patterns and build user profiles according to the context. Such approaches, however, do not address the need of adaptation in the means of interaction with the smart products in the usage stage of S-PSS.

Users' short term memory is limited to a few items of information. Cognitive load refers to the total amount of mental activity on working memory at an instance in time. If the user interface requires the user to hold more than it can retain, it will produce a cognitive overload (Tracy and Albers, 2006). Therefore, providing customized services recommendations address this problem in applications designed to fully manage or interact with smart products. Context-awareness recommendation systems have been proved to provide better customization since context information is very important in improving recommendation accuracy (Yang, 2018). Tarus et al. (2018) proposed a context-aware recommender system for e-learning, using collaborative filtering (CF) algorithms and learner's goals and study patterns as sources of context in addition to a traditional rating of courses. Yang (2018) presented a recom-

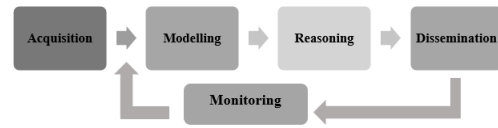


Figure 1: Context-aware life cycle.

mender system model, implemented as a website for movie recommendation. Users are allocated to the corresponding set of similar users using the Top-N algorithm. In the user behaviour analysis step, popular movies are recommended. Then, a matrix of user scores is established by analyzing the results of user-registered movies and using time as context of each score (i.e. season of the year, weekday or weekend).

This section has attempted to provide a brief summary of the literature. As discussed above, the proposed approaches to UX focus their attention only on the design stage, before the use phase. As S-PSS represent a way of continuously striving to meet customer needs, adaptation needs to be reflected also in the interaction with smart products. The section below describes the framework proposed in this paper.

3 FRAMEWORK

Perera et al. (2013) defined four general steps in context life cycle as a way to describe the process of developing context-aware applications (Figure 1). The first step is *Context Acquisition*, which refers to data that needs to be acquired from different sources (i.e. sensors, cyber-physical systems (CPS), databases, etc). In Context Modelling, the collected data needs to be represented in a meaningful way. Then, in Context Reasoning, data is processed to provide useful information and insights (context). Context Dissemination phase distributes context to consumers, which can be end-users or other applications. Lastly, Context Monitoring represents the evolution of the design of S-PSS. It is a stage not broadly represented and should be considered after Dissemination because context may change at some point. Systems or applications have to be able to identify changes and update the models accordingly.

Following the context-aware life cycle, this paper proposes a theoretical framework to provide a personalized UX where services are recommended and shown according to context when the interaction occurs. Figure 2 presents the overall graphic representation of the framework. The following subsections will describe each layer of the framework in detail.

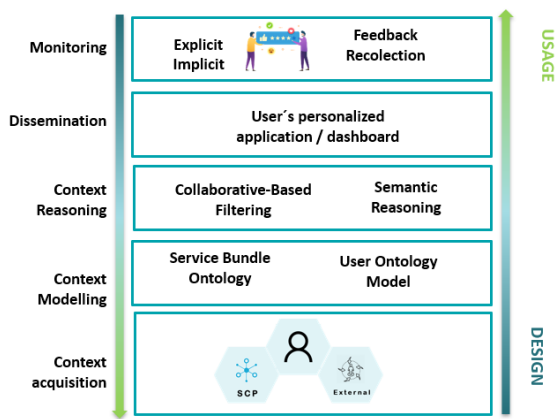


Figure 2: Context-aware framework for UX.

3.1 Context Acquisition

The process of context acquisition will require capturing the data from multiple sources. This process can be classified by the source of context and the nature of the data. Figure 3 presents three main data sources, *Smart Connected Product* (SCP), *user*, and *external*. The nature of the data can be dynamic or static. User context refers to any information related to the user. User static data might include user personal information or user preferences (Liu et al., 2011). The dynamic information might include: the user's current and historical location, user's current and historical activity, user's current emotion, relationships, etc. For instance, capturing the data from users while they interact directly with products or e-services usually delivered by mobile apps or web applications. SCP data is acquired from the device through sensors in smart products that can proportionate relevant information of the user when the smart products are used to monitor the user routine, for instance, wearable devices. Finally, external sources contain dynamic environment physical information (i.e. lighting, noise, temperature, and humidity level, traffic conditions)

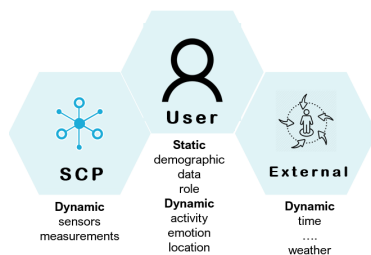


Figure 3: Context data sources.

3.2 Context Modelling

Context modelling allows the representation of the data acquired in terms of attributes, characteristics, and relationships with previously specified context (Liu et al., 2011). Thus, a static model is necessary to represent the services of smart devices where each of them groups low-level services or sub-services related. A sub-service will represent a very specific function of the e-service platform or smart device. According to (Maleki et al., 2018) in an S-PSS architecture, the domain knowledge models function in an intermediary role to integrate the S-PSS services with the cyber-physical components. One approach is the use of domain ontologies. Ontologies are a means to formally model concepts from a particular domain into a detailed specification of entities with properties and relations. Domain ontologies define the vocabulary related to a particular domain (Guarino et al., 2009). In this case, each service is modelled into a pattern that relates the product, service and required information (Maleki et al., 2018) and integrates it with context-specific classes and relationships (see Figure 4). In this way, the interactions with the service can be labelled into multiple categories (i.e. click, like, close) and then linked to the services to understand user behaviour for each of the sub-services.

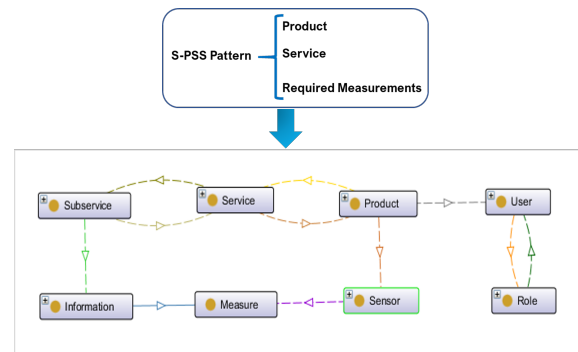


Figure 4: S-PSS Pattern and derived ontology.

3.3 Context Reasoning

Context reasoning is the process of deducing new information that is useful for a task or user, from multiple context-data sources. The relevance of a recommendable service solution not only depends on the users' general preferences but also on their current situation and their short-term interaction and interests (Quadrana et al., 2018). Furthermore, the data from the product itself is a source of context. Collaborative Filtering (CF) is a recommendation technique largely used in e-commerce to provide personalized recommendations on items (i.e. books, movies,

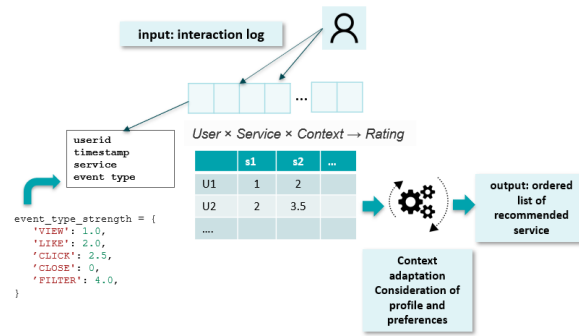


Figure 5: Reasoning process context-aware recommendation systems.

clothes, etc). It uses similarities between users and items simultaneously to provide recommendations. CF exploits the ratings provided to products explicitly (i.e. starts ratings) to measure the similarity and provide adequately recommendations (Elahi et al., 2016). However, is also possible to get implicitly forms of ratings by obtaining and evaluating the interactions that the user had (i.e. page views, clicks, purchases, etc), which is the more adequate method in the case of S-PSS. Each type of interaction could represent a higher level of engagement with a specific service. For example, monitoring events such as *filter changes* and *clicks* would represent a higher rating in comparison with a page view event (Chen, 2005; Núñez-Valdez et al., 2018) as seen in Figure 5. In a context-aware scenario, ratings are modeled as a function of: users, service as well as context. Hence, the rating function can be defined in three dimensions (Tarus et al., 2018): *User x Service x Context Rating*, the process will produce as output an ordered list of recommended services based on the generated ratings of other users with similar preferences in the same context (Figure 5). Although collaborative filtering is the most popular recommendation technique, its major disadvantage is the new user and new item problems (Barjasteh et al., 2016), referred commonly as the cold-start problem, which occurs in scenarios where it is not possible to make reliable recommendations due to an initial lack of ratings (Adomavicius and Tuzhilin, 2005). Semantic reasoning can be used to explicitly infer the preference of implicit users, reducing the sparsity of user information, and alleviating the problems of cold start (Yang, 2018). Considering the use of ontologies in the modelling stage, semantic reasoning could take advantage of this to infer new knowledge based on established relationships.

3.4 Context Dissemination and Monitoring

Context Dissemination is related to the methods to deliver context to clients or users. The purpose of this stage is to provide the individualization of the experience by providing services adequately for each user situation. Digital services should be adaptive to user context and easily accessible through digital platforms provided by the service in the form of apps or websites or the product means of interaction by itself. Furthermore, the process of context monitoring represents the evolution of the design of S-PSS. It is important to monitor user feedback regarding the services delivered by the application. Feedback from users can come through integrated surveys, but also from user behaviour. For instance, if a user is not happy with the service that the application is offering, it is likely they will return to an old state of the application, or do not interact with the information. This can be taken as negative feedback and a negative rating for the service provided. Having discussed each layer, the following section of this paper addresses an application scenario to understand the implementation process.

4 APPLICATION SCENARIO

In order to demonstrate the proposed framework, an activity monitoring scenario is used for a product-oriented PSS. Tukker (2004) classified PSS into three main categories from the business aspect: *Product-oriented*, *Result-oriented*, and *Use-oriented*. Product-oriented is related to services that facilitate the sales of products or add functionality or personalize existing products. In Result-oriented and Use-oriented, the ownership usually remains with the provider. Fitbit is one of the most popular multi-purpose activity wearable trackers and smart-watch and represents a clear example of *Product-oriented* PSS for end-consumers (B2C). An important part of the product

is the e-services platform since it represents a means of interaction with the device and also a source of information to the user which is accessible through the mobile app or web application.

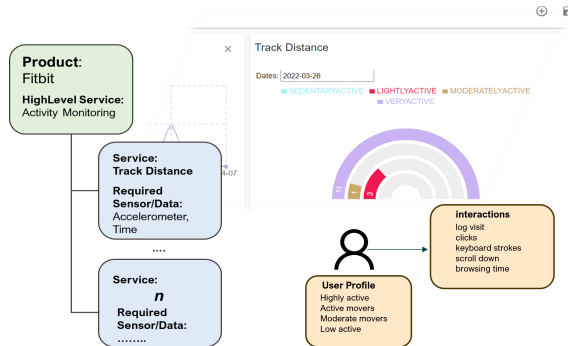


Figure 6: Fitbit application scenario.

In the framework, the context acquisition layer collects the information from the device, which in this case represents data about the user's behaviour, for instance, the number of steps taken by the user in a day. This translates into information such as the distance, the duration as well as the intensity of activity. User interaction with the e-service platform will be captured using Matomo¹, which is an open-source platform capable to capture user interaction (i.e. clicks, page views) from a web application into a Mysql database. Figure 6 represents the context modelling layer. Using an ontology, according to section 3.2, each high-level service provided by the S-PSS (i.e. activity monitoring) has associations with low-level services that represent specific options on the e-service platform. Each interaction that is recorded can be related to a specific service and further enriched with contextual information, for instance, weather, time and device status. Another significant aspect is that the use of an ontology could mitigate the 'cold start' problem when there is a lack of historical interactions. In this scenario, profiles were defined using a dataset from 30 anonymous users (Furberg et al., 2016) by means of a non-supervised machine learning technique, using the clusterization algorithm, K-Means (Figure 7).

Four profiles were identified, in base of their behavioural information, and further used to model an ontology using GraphDB. The ontology can define the relationships among users, profiles and high-level and low-level services as shown in Figure 8, *User1* has an *Active* profile and uses *trackExercise* low-level service. Using semantic reasoning, it can be inferred that users with *Active* profile use the *trackExercise* service. In future work, the reasoning layer will use

¹<https://matomo.org/>

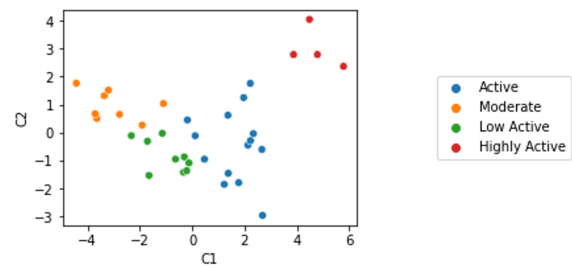


Figure 7: K-means clustering.

collaborative filtering algorithms to produce an ordered list of recommended services. These services will be delivered through the e-service platform application in the form of widgets or personalized notifications in a web application using React framework. One of the challenges will be to achieve a way to deliver the adaptive services in several S-PSS interfaces; software frameworks for application creation are very diverse. *Javascript* is the ultimate web standard with multiple frameworks available (i.e. React, Vue, AngularJS), which have extended to mobile apps. But, forms of interactions might vary in industrial (i.e. manufacturing) settings. This highlights the importance to understand the type of market where the S-PSS is directed. It is key to validate the framework in both, B2C scenario and also in a Business to Business scenario (B2B), where the UX can be perceived differently by users, and the means of interaction and feedback recollection can also change.

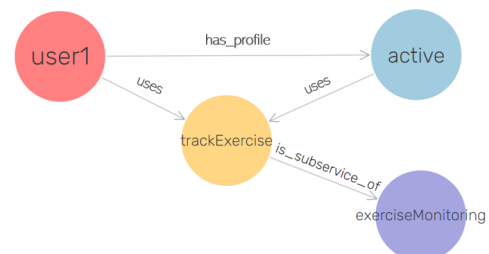


Figure 8: Ontology query in GraphDB.

Another challenge of a data-driven UX is the need for users' personal and behavioural information, which could vary depending on S-PSS' application domains (i.e. health & welfare, manufacturing, smart home, etc.). The services created must help users make informed decisions about their privacy. Similarly, users must have control over their personal information (van Ooijen and Vrabec, 2019). Privacy regulations should be followed (i.e. General Data Privacy Regulation GDPR) and ensure users' consent and awareness of the data collected (Zheng et al., 2019a).

5 CONCLUSIONS

S-PSS emerges as a way to provide a market offer combining both product and service while maintaining a lasting relationship between provider and customer. It is important to comprehend the role of UX to understand the use and expectations of users with smart products and e-service platforms. Continuously delivering personalized experiences for users is challenging. The creation of value with user interactions can allow maintaining the system relevant to the changing needs. This can be accomplished, by implementing technology in a way that it results in a simple and intuitive experience. This work described a theoretical context-aware framework for the UX of S-PSS, following the life cycle of context-aware application and considering the use of context in recommendation systems and the users interactions. The goal is to provide a personalized and adaptive e-service S-PSS platform for each user. An application scenario of a wearable activity tracker device was presented as a product-oriented S-PSS that allows to better showcase the general overview of the framework. The work also presented a description of the challenges that brings S-PSS, such as a multi-business perspective to understand users in different scenarios and industries, and the management of behavioural data and how to handle privacy aspects within the applications. Future work will provide an implementation of the framework and further analyze each of the stages described in this paper.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant No. 814078.

REFERENCES

- Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. (1999). Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*, pages 304–307. Springer.
- Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.
- Barjasteh, I., Forsati, R., Ross, D., Esfahanian, A.-H., and Radha, H. (2016). Cold-start recommendation with provable guarantees: A decoupled approach. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1462–1474.
- Carrera-Rivera, A., Larrinaga, F., and Lasa, G. (2022). Context-awareness for the design of smart-product service systems: Literature review. *Computers in Industry*, 142:103730.
- Chang, D., Gu, Z., Li, F., and Jiang, R. (2019). A user-centric smart product-service system development approach: A case study on medication management for the elderly. *Advanced Engineering Informatics*, 42:100979.
- Chen, A. (2005). Context-aware collaborative filtering system: Predicting the user's preference in the ubiquitous computing environment. In *International Symposium on Location-and Context-Awareness*, pages 244–253. Springer.
- Cong, J.-c., Chen, C.-H., Zheng, P., Li, X., and Wang, Z. (2020). A holistic relook at engineering design methodologies for smart product-service systems development. *Journal of Cleaner Production*, 272:122737. Publisher: Elsevier BV.
- Dou, J. and Qin, J. (2017). Research on User Mental Model Acquisition Based on Multidimensional Data Collaborative Analysis in Product Service System Innovation Process. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 35–44. Springer.
- Elahi, M., Ricci, F., and Rubens, N. (2016). A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50.
- Furberg, R., Brinton, J., Keating, M., and Ortiz, A. (2016). Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016.
- Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.
- Hassenzahl, M. and Tractinsky, N. (2006). User experience—a research agenda. *Behaviour & information technology*, 25(2):91–97. Publisher: Taylor & Francis.
- Iso, D. (2010). 9241–210: 2010: ergonomics of human-system interaction—part 210: human-centred design for interactive systems (formerly known as 13407). *Switzerland: International Standards Organization*.
- Jia, G., Zhang, G., Yuan, X., Gu, X., Liu, H., Fan, Z., and Bu, L. (2021). A synthetical development approach for rehabilitation assistive smart product-service systems: A case study. *Advanced Engineering Informatics*, 48:101310. Publisher: Elsevier.
- Liu, W., Li, X., and Huang, D. (2011). A survey on context awareness. In *2011 International Conference on Computer Science and Service System (CSSS)*, pages 144–147. IEEE.
- Maleki, E., Belkadi, F., Boli, N., Van Der Zwaag, B. J., Alexopoulos, K., Koukas, S., Marin-Perianu, M., Bernard, A., and Mourtzis, D. (2018). Ontology-based framework enabling smart product-service systems: application of sensing systems for machine health monitoring. *IEEE internet of things journal*, 5(6):4496–4505. Publisher: IEEE.

- Misaka, M. and Aoyama, H. (2018). Development of design system for crack patterns on cup surface based on kansei. *J. Comput. Des. Eng.*, 5:435–441.
- Mourtzis, D., Fotia, S., Vlachou, E., and Koutoupes, A. (2018). A Lean PSS design and evaluation framework supported by KPI monitoring and context sensitivity tools. *The International Journal of Advanced Manufacturing Technology*, 94(5-8):1623–1637. Publisher: Springer Nature BV.
- Núñez-Valdez, E. R., Quintana, D., Crespo, R. G., Isasi, P., and Herrera-Viedma, E. (2018). A recommender system based on implicit feedback for selective dissemination of ebooks. *Information Sciences*, 467:87–98.
- Perera, C., Zaslavsky, A., Christen, P., and Georgakopoulos, D. (2013). Context aware computing for the internet of things: A survey. *IEEE communications surveys & tutorials*, 16(1):414–454. Publisher: Ieee.
- Pezzotta, G., Cavalieri, S., and Romero, D. (2017). Collaborative product-service systems engineering: Towards an active role of customers and stakeholders in value co-creation. In *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1247–1254. IEEE.
- Quadrana, M., Cremonesi, P., and Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)*, 51(4):1–36.
- Reguera-Bakhache, D., Garitano, I., Cernuda, C., Uribeetxeberria, R., Zurutuza, U., and Laso, G. (2021). An adaptive industrial human-machine interface to optimise operators working performance. In *2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 1213–1219. IEEE.
- Schilit, B. N. and Theimer, M. M. (1994). Disseminating active map information to mobile hosts. *IEEE network*, 8(5):22–32.
- Seo, D. W., Kim, H., Kim, J. S., and Lee, J. Y. (2016). Hybrid reality-based user experience and evaluation of a context-aware smart home. *Computers in Industry*, 76:11–23. Publisher: Elsevier.
- Tarus, J. K., Niu, Z., and Kalui, D. (2018). A hybrid recommender system for e-learning based on context awareness and sequential pattern mining. *Soft Computing*, 22(8):2449–2461.
- Todi, K., Bailly, G., Leiva, L., and Oulasvirta, A. (2021). Adapting user interfaces with model-based reinforcement learning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Tracy, J. P. and Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual Conference-society for technical communication*, volume 53, page 256.
- Tukker, A. (2004). Eight types of product-service system: eight ways to sustainability? Experiences from SusProNet. *Business Strategy and The Environment*, 13:246–260.
- Valencia, A., Mugge, R., Schoormans, J., and Schifferstein, H. (2015). The design of smart product-service systems (PSSs): An exploration of design characteristics. *International Journal of Design*, 9(1).
- van Ooijen, I. and Vrabec, H. U. (2019). Does the gdpr enhance consumers’ control over personal data? an analysis from a behavioural perspective. *Journal of consumer policy*, 42(1):91–107.
- Wang, Z., Chen, C.-H., Zheng, P., Li, X., and Khoo, L. P. (2019). A novel data-driven graph-based requirement elicitation framework in the smart product-service system context. *Advanced engineering informatics*, 42:100983. Publisher: Elsevier.
- Wang, Z., Chen, C.-H., Zheng, P., Li, X., and Khoo, L. P. (2021). A graph-based context-aware requirement elicitation approach in smart product-service systems. *International Journal of Production Research*, 59(2):635–651.
- Wetter-Edman, K., Sangiorgi, D., Edvardsson, B., Holmlid, S., Grönroos, C., and Mattelmäki, T. (2014). Design for value co-creation: Exploring synergies between design for service and service logic. *Service Science*, 6(2):106–121. Publisher: INFORMS.
- Yang, Q. (2018). A novel recommendation system based on semantics and context awareness. *Computing*, 100(8):809–823.
- Zheng, P., Liu, Y., Tao, F., Wang, Z., and Chen, C.-H. (2019a). Smart product-service systems solution design via hybrid crowd sensing approach. *IEEE Access*, 7:128463–128473.
- Zheng, P., Wang, Z., Chen, C.-H., and Khoo, L. P. (2019b). A survey of smart product-service systems: Key aspects, challenges and future perspectives. *Advanced engineering informatics*, 42:100973. Publisher: Elsevier.

A Music Tangible User Interface for the Cognitive and Motor Rehabilitation of Elderly People

Adriano Baratè^a, Helene Korsten and Luca A. Ludovico^b

Laboratory of Music Informatics, Department of Computer Science, University of Milan, via G. Celoria 18, Milan, Italy
{adriano.barate, luca.ludovico}@unimi.it, helene.korsten@studenti.unimi.it

Keywords: Music, Tangible User Interface, Web, Kibo.

Abstract: This paper proposes the combined use of tangible user interfaces, digital technologies, and musical expression in the context of cognitive and motor rehabilitation of elderly people. After analyzing the state of the art about common age impairments and tangible user interfaces in rehabilitation, we will introduce the *Kibo*, a MIDI controller based on the concept of fiducials and able to communicate with other MIDI devices via Bluetooth. The peculiar characteristics of the *Kibo* will be exploited in a Web framework aiming to foster the development or recovery of cognitive and motor abilities through 3 specially designed games. The paper will also report remarks from domain experts (music therapists and psychiatrists) and the consequent redesign.

1 INTRODUCTION

According to a 2019 report by the United Nations, there were 703 million persons aged 65 years or over in the world in 2019 and the number of older persons is projected to double to 1.5 billion in 2050. These changes for individuals are mirrored in society since older persons represent a growing demographic group in society. One of the goals for sustainable development should be investing in education, health, and well-being for all, including the elder.

The target audience of the project documented in this paper is represented by elderly people who need rehabilitation sessions or are dissatisfied or inactive in their social interactions. Starting from an analysis of the most common impairments and changes in cognitive and motor abilities due to age, we designed and developed a prototype applicable to music-therapy rehabilitation sessions. Interventions can address a variety of healthcare and educational goals (e.g., promoting wellness, managing stress, alleviating pain, expressing feelings, enhancing memory, improving communication, promoting physical rehabilitation, etc.).

The result is a Web-based prototype that offers intuitive interaction with musical parameters via a specific tangible user interface, called the *Kibo*, so as to take advantage of the haptic interaction skills with the

environment. In particular, three browser games have been developed in order to train rhythm, spatial, and recognition skills.

2 STATE OF THE ART

In this section, we will review the scientific literature that is most relevant for our purposes. Due to the vastness of the subject, we cannot be exhaustive. We will limit ourselves to a few references of particular importance for our work, addressing: i) common age impairments, both cognitive and physical; ii) tangible user interfaces and their applicability to the music field; and iii) experiences of tangible user interfaces in rehabilitation.

Starting from **age impairments**, due to an increasing life expectancy, the number of elderly people has increased significantly worldwide. Thus, society will have to offer more and more products and services that meet the specific needs and desires of a geriatric age group that makes up a great percentage of the population. These people are seeking solutions to help them cope with daily life, give them the opportunity to interact socially, and find alternative ways of entertainment and learning. Moreover, health systems are dealing with the ever-increasing burden of finding solutions and cures for age-related and degenerative conditions (e.g., dementia and Alzheimer's disease) and impaired-movement pathologies (e.g.,

^a  <https://orcid.org/0000-0001-8435-8373>

^b  <https://orcid.org/0000-0002-8251-2231>

Parkinson's).

Everyone experiences changes as an inevitable part of the natural degeneration associated with aging (Iversen, 2015). Concerning *changes in cognitive abilities*, aging causes a decline in spatial cognition, which is the ability to represent spatial relationships among objects. The results of a study conducted in 2008 by Iachini *et al.* show that some spatial abilities, such as the ability to mentally rotate visual images and to retrieve spatio-temporal sequences, decline with age (Iachini *et al.*, 2008). Elderly people also struggle more with multitasking, especially if the tasks are complex (Wecker *et al.*, 2005). Another common age impairment is the decline in fluid intelligence, which refers to the processing and reasoning components of intelligence and the natural ability to learn something new (Czaja and Lee, 2007). Because of reduced processing efficiency, the working memory, which is the ability to keep information active while processing or using it, declines with age (Salt-house, 1991). Similarly, prospective memory, which is the ability to remember to do something in the future, also declines with age (Maylor, 1995). Another issue emerging with age is the ability to select information in the environment, e.g., to attend to information on a Web page. Attention is the ability to focus on a specific task or an object in the environment while ignoring other things. This ability changes with age and older people are slower to move their attention from one thing to another (Czaja and Lee, 2007).

Another category of impairments due to aging is that of *changes in physical abilities*. Response time and accuracy of movement decline with age. Older persons' movements, reflexes, and reactions to stimuli are typically slower than younger persons' (Farage *et al.*, 2012). Differently from the above-mentioned normal age impairments, a disease that limits movement is Parkinson's, a neurodegenerative brain disorder that progresses slowly and worsens with age. Symptoms are involuntary shaking of the hands, arms, legs, jaw, chin, and lips. Other main symptoms include slowness of movement, stiffness of arms and legs, and trouble with balance (Jankovic, 2008).

Even if not an impairment, another phenomenon relevant for our work is anxiety towards technology. In a study by Czaja and Sharit dating back to 1998, older people saw themselves as having less control over computers than younger people (Czaja and Sharit, 1998). They had significantly less efficacy in completing computer tasks, but, surprisingly, they also perceived computers as being more useful compared to younger people. Another study revealed that older subjects reported higher levels of computer anxiety than younger persons and that the anxiety level

was related to the decision time on the computer when performing a test (Laguna and Babcock, 1997). The elderly may have a harder time using new technology, but it helps with practice (Chu, 2010). These considerations pushed us to develop a solution based on a computing system but relying on a tangible interface to ease user interaction.

Focusing on **tangible user interfaces (TUIs)**, they are intended to replace the graphical user interfaces (GUIs), more common in computing systems, with real physical objects the user can interact with. The key idea is to give digital information a physical form and let these physical forms serve as a representation and control for digital information. A TUI lets users manipulate digital information with their hands and perceive it with their senses. One of the pioneers in tangible user interfaces is Hiroshi Ishii, a professor at MIT who heads the *Tangible Media Group* at the MIT Media Lab. His particular vision for TUIs, called *Tangible Bits*, is to give physical form to digital information, making bits directly manipulable and perceptible (Ishii, 2008b). *Tangible Bits* pursues the seamless coupling between physical objects and virtual data. "*TUIs will augment the real physical world by coupling digital information to everyday physical objects and environments*" (Ishii and Ullmer, 1997). All physical objects can potentially be a part of a digital user interface (Ishii, 2008a). For example, if an object, which is a part of a TUI, is moved or put in a specific position, a digital signal will be sent from either the tangible object itself or from another device that senses the object. Currently, there are different research areas and applications related to TUIs. For instance, tangible augmented reality implies that virtual objects are "attached" to physically manipulated objects; in tangible tabletop interaction, physical objects are moved upon a multi-touch surface; moreover, physical objects can be used as ambient displays or integrated inside embodied user interfaces.

TUIs have been used considerably in musical performances and music-therapy treatments. A tangible interface – implying something "real", "concrete" – offers a physical way to interact with music and sound parameters. Most traditional musical instruments are played through this kind of interaction, but the advent of digital technologies paved the way for innovative and original approaches. Concerning technology-enhanced TUIs for music, the scientific literature describes a number of theoretical approaches, prototypes, and available products. Music TUIs can play a number of roles: for example, synthesizers to generate sound, sequencers that perform audio samples and mix them together, remote controllers for music and sound parameters, or interfaces for music-related

games. To cite but a few references, Paradiso *et al.* review TUIs based on magnetic tags (Paradiso *et al.*, 2001), Newton-Dunn *et al.* describe a way to control a dynamic polyrhythmic sequencer using physical artifacts (Newton-Dunn *et al.*, 2003), and Schiettecatte and Vanderdonckt present a distributed cube interface based on interaction range for sound design (Schiettecatte and Vanderdonckt, 2008). A noticeable example of commercially available music TUI is the *reacTable* (Jordà *et al.*, 2007), used by renowned artists such as Björk in their live performances. Our solution adopts a low-cost device called the *Kibo*, whose features will be described in detail in Section 3.

Finally, let us address the applicability and role of **tangible user interfaces in rehabilitation**. Motivation is one of the main problems evidenced in traditional therapy sessions, often hampered by the repetitive nature of exercises. Most studies show that an effective rehabilitation must be early, intensive and repetitive (Rego *et al.*, 2014; Burke *et al.*, 2010). As such, these approaches are often considered repetitive and boring by the patients, resulting in difficulties in maintaining their interest and in assuring that they complete the treatment program (Rego *et al.*, 2014). On the other hand, due to their nature, games can motivate and engage the patients' attention and distract them from their rehabilitation condition. On one side, they require some motor and cognitive activity, but, on the other, they have a story and can offer feedback and levels of challenge and difficulty that can be adapted to the patients' skills.

Serious games are an option that provides learning combined with entertainment. The locution "serious games" refers to playful activities that provide training and physical or mental exercise in a fun and enjoyable way (Dörner *et al.*, 2016). These games can be not only a way to prevent the feeling of loneliness (De Carvalho *et al.*, 2012), but they can also enable social interaction (Fonseca *et al.*, 2021). During the last decades, digital games have become a popular leisure activity. Ijsselsteijn *et al.* claim that digital games can be considered a promise to improve the lives of seniors. To this end, it is important to develop interesting and accessible games which could provide an option to spend quality time with clear benefits (Ijsselsteijn *et al.*, 2007). Pearce reports that the studies of digital games with an emphasis on older people are still rare because this need is relatively new (Pearce, 2008). This perception may be justified by the fact that the current elderly population did not have much access to technology at their earlier age.

Many rehabilitation games based on TUIs are already available. For example, *Handly* is an integrated upper-limb rehabilitation system for persons with a

neurological disorder (Vandermaesen *et al.*, 2016). *Handly* consists of tangibles for training four-hand tasks with specific functional handgrips and a motivational game. The system consists of four tangible training boxes, which each present one essential grip and associated hand task: push-pull, squeezing, knob turning, and key turning. *Handly* combines tangibles specifically designed for repetitive task-oriented motor skill training of typical daily activities with serious gaming, thus offering a comprehensive approach.

Segara is an integrated hand rehabilitation system for patients with rheumatoid arthritis (Zhao *et al.*, 2021). *Segara* consists of tangibles for training six tasks with interactive functional handgrips and a motivational serious game. It shows that a system combining games and tangibles to enhance hand rehabilitation is feasible and highly appreciated by patients.

Resonance is an interactive tabletop artwork that targets upper-limb movement rehabilitation for patients with an acquired brain injury (Duckworth *et al.*, 2015). The artwork consists of several interactive game environments, which enable artistic expression, exploration, and play. *Resonance* provides uni-manual and bi-manual game-like tasks and exploratory creative environments of varying complexity geared toward reaching, grasping, lifting, moving, and placing tangible user interfaces on a tabletop display. Each environment aims to encourage collaborative, cooperative, and competitive modes of interaction for small groups of co-located participants.

NikVision is a tangible tabletop based on a user-centered design approach for the cognitive stimulation of older people with cognitive impairments and dementia problems in nursing homes (Cerezo *et al.*, 2020). The general experiences of the users when working with the tangible tabletop were assessed and applied to the design of new cognitive and physical stimulation activities. Game activities are specially designed for the elderly, including daily tasks such as getting dressed, cognitive assignments, and upper-half motor skills training. Activities present different levels of difficulty and audio feedback.

3 EMPLOYED TECHNOLOGIES

In this paper, we want to take advantage of real physical objects and the benefits that TUI technologies offer in order to provide the elder with a playful way to interact in rehabilitation therapies. As mentioned before, one of the advantages of a TUI is the user experience, because a physical interaction occurs between the user and the interface itself. Another advantage is usability since the user intuitively knows how to ma-

nipulate the interface by knowing the function of the physical object; consequently, the user does not need to learn the functionality. That is why TUIs are often used to make technology more accessible for elderly people.

In this section, we will address the key technologies employed in the project: a music TUI called the *Kibo* (Amico and Ludovico, 2020), the way it is connected to a Web application, and the Web languages and formats used to implement the browser games described later.

The *Kibo* is a wooden board produced by Kodaly that presents eight distinct, easy-to-recognize tangibles.¹ These geometric shapes can be inserted into and removed from the corresponding slots, thus triggering events encoded in the form of MIDI messages. The device is also sensitive to pressure variations on single tangibles and individual dynamic responses can be communicated via MIDI messages. As an additional controller, there is a knob that can be rotated clockwise and counterclockwise. Finally, the device is equipped with a gyroscope.

The *Kibo* can be connected via Bluetooth or USB to iOS and macOS devices running a proprietary app that acts as both a synthesizer and a configuration center. Windows and Android operating systems are also supported via third-party drivers. The communication between the controller and the app occurs by exchanging standard MIDI 1.0 messages. The MIDI engine integrated into the app supports up to 7 *Kibo* units simultaneously, without perceivable latency. This aspect is particularly interesting for collaborative experiences (Baratè et al., 2021). Being a fully compatible controller, *Kibo* can also be integrated into any MIDI setup without the intervention of the app as a mediator, as we will do in our proposal.

The control over music parameters is mainly based on the 8 tangibles shown in Figure 1. Each object has a different shape fitting in a single slot and presents symmetry properties so that it can be rotated and flipped before being inserted. Tangibles have a magnetic core; they can be stacked one on top of the other and interact through magnetic fields. The body of the *Kibo* contains a multi-point pressure sensor that allows for the detection of the insertion and removal of tangibles. The characteristics of the sensor make the instrument extremely sensitive and, simultaneously, very resistant.

Even if our proposal bypasses the native *Kibo*'s app, it is interesting to analyze the three original operating modes and shed some light on rehabilitative and therapeutic scenarios:



Figure 1: The *Kibo*'s body and tangibles.

1. *Musical Instrument Mode* — In this scenario, *Kibo*'s tangibles are mapped onto pitches. Associations between shapes and notes can be customized, even triggering multiple notes through a single tangible. The device is able to detect *after-touch* effects, namely the possibility to track pressure variations over tangibles after note attacks;
2. *Beat Mode* — In this scenario, tangibles are mapped onto single percussive instruments. The pressure sensor, presenting a high level of resistance to strong mechanical stresses but also a noticeable sensitivity, allows effects ranging from hard mallet beats to delicate brush rubbing;
3. *Song Mode* — In this scenario, *Kibo* is employed as a controller to trigger already available music loops. Tangibles are associated with mutually synchronized but independent tracks, like in a multi-track environment. When tangibles are inserted, the corresponding tracks are activated; when they are removed, tracks are muted.

The configurability of *Kibo*, coupled with the adoption of standard communication protocols, enables numerous and heterogeneous scenarios. Multiple *Kibo* units in an ensemble can be configured to cover distinct note ranges and timbres, or even to work in different operating modes, thus providing the therapist with great flexibility. Moreover, the standard MIDI output of the *Kibo* allows the implementation of additional operating modes where other meanings, even extra-musical ones, can be assigned to user gestures. Our proposal, described in detail in Section 4, explores this possibility.

Even if originally conceived as a general-purpose tangible MIDI controller, when used in a suitable scenario the *Kibo* can also be considered an *assistive* technology falling in the category of communication boards. Under this perspective, it has a *therapeutic* function since it encourages upper-limb movements and challenges cognitive skills. Moreover, it is *compensatory* both from a motor point of view, being able to translate even small movements into sound, and from a cognitive point of view, enabling intuitive musical expression by lowering the barriers of a traditional instrument.

¹<https://www.kodaly.app/>

In this project, the *Kibo* has to be directly connected via Bluetooth to the computing system that hosts the browser. A specific protocol – called MIDI over Bluetooth or, simply, Bluetooth MIDI – is specially conceived to exchange MIDI messages over Bluetooth connections (Bartolomeu et al., 2005). The *Kibo* adopts Bluetooth Low Energy (BLE), a wireless personal-area network technology that, compared to the original Bluetooth protocol, is intended to provide considerably reduced power consumption and cost while maintaining a similar communication range. A BLE MIDI device will transparently operate with a MIDI-compatible application on most mobile and desktop platforms with no additions and provides “out of the box” support for most use cases compared with any wired or wireless alternative.

The visual feedback of the system is implemented in the form of a Web page by adopting W3C standard languages and formats. As a result, the browser games can be experienced using any HTML5 browser. The project was realized using HTML, CSS, and JavaScript.

Please note that all the mentioned technologies are client-side, so the platform could also be enjoyed locally on the user’s client, with no need to connect to a server. Nevertheless, we have publicly released it over the Web in order to distribute it easily at no cost for the user and to keep it up-to-date in the case of a new release.

4 KIBO WEB GAMES

The goal of linking music therapy, rehabilitation exercises, and technology through the *Kibo* brought to the design and implementation of a Web platform that proposes three serious games. The name of the platform is *Kibo Web Games*. All game activities allow working the association between physical elements (geometric tangibles) and the concepts they represent (e.g., notes, tracks, hit buttons). The platform is available at <https://kibogames.lim.di.unimi.it/> and its original Web interface is shown in Figure 2. *Kibo Web Games* need a *Kibo* device connected via BLE.

A number of game parameters have been introduced to allow flexibility in usage and adaptation to gradual improvements without causing frustration in the player. The user or therapist can set different types of activities and levels of difficulty according to the following parameters: i) the **game** to play, that is basically a rhythm-based activity where given shapes are proposed to the player and must be timely touched, inserted, removed, or continuously pressed depending on the game mode, as explained in the following;

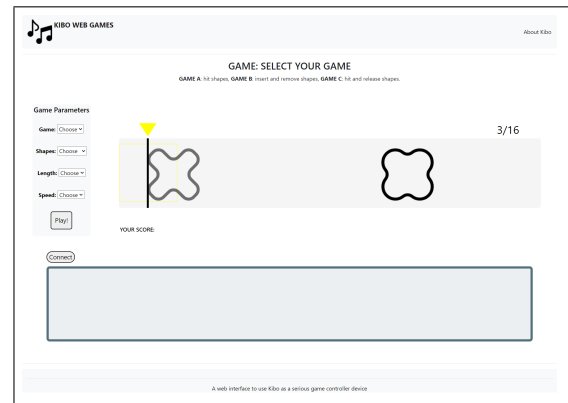


Figure 2: The interface of *Kibo Web Games*.

ii) the **number of shapes** (1 to 8 out of the 8 available) that can randomly be involved during the game experience; iii) the **total length** of the game sequence, i.e. the number of actions a user should carry out in a game session; and iv) the **speed** of the shape generation and scroll animation.

The central part of the interface is taken by the *Game field*, which displays the moving shapes during a game session. Shapes move from right to left, and the perfect timing for user’s actions is when they hit the black vertical line. Another relevant part of the interface in Figure 2 is the *MIDI message console*, namely the lower rectangle that displays the MIDI messages received from the *Kibo*. Finally, the *Play* button starts the selected game with the parameters set by the user.

The second game expects the user to insert or remove the shapes from their slots in synchronization to the game’s graphic scenario is able to train precision in performing movements. The third game asks the user to simultaneously press and release the recognized shapes. The scores, levels of difficulty, and feedback given to the player depend on her/his temporal precision in performing the movement.

Concerning the gameplay, *Kibo Web Games* focus on the interactions occurring between the player and the *Kibo*, i.e. simple motor movements like hitting, tapping, grabbing, holding, releasing, placing, and removing the geometrical tangibles. Furthermore, recognition and listening skills are trained to help restore or keep active cognitive functions. The *Kibo* gives tactile and musical feedback to every action that is performed and allows the user to proceed in small steps. It also allows users to manipulate objects giving the clearer image of the connection between physical interaction and the response that it triggers.

All games share the same game field and functional concepts. The system generates shapes scrolling across the screen from right to left and the

goal is to timely recognize the shape by hitting, pressing, inserting, or removing the corresponding tangible from the *Kibo* physical body when it reaches the vertical line in the delimiter box (Figure 2). Throughout the duration of a game session, the arrow, square, and vertical line in the *Game field* animate to facilitate the game experience. Arrow and square flash when the user is supposed to perform a task. The vertical line turns either red or green according to the accuracy of the performance. The score field keeps track of the points gained while playing and remains displayed until the next game session is started.

Perfect hits are those performed in a very small range around the computed timing. Good hits occur in a slightly wider timing range. The exact values depend on the game speed set by the user, but they are in the order of tenths of a second. Currently, score penalties for wrong shape recognition or missed hits are not implemented to avoid frustration in players, but this feature could be easily integrated.

Please note that the *Kibo Web Games* interface allows users to play with the *Kibo* instrument as a simple musical controller, regardless a game session is active or not, after connecting it to the Web page. This provides sound feedback (C-major scale) and console messages only, but it might help to gain confidence and increase motivation. As another remark, the three games help gain rhythm awareness and present musical feedback, but they never imply a music-education background nor make use of any kind of music-related concept.

Game A – Tap the Shapes is based on the *Kibo's Beat Mode* (see Section 3). Tangibles are mapped onto single notes of a C-major scale, one grade per shape. Pressing, hitting, or tapping a shape sends a Note On message immediately followed by a Note Off. MIDI messages are interpreted by the interface and played back thanks to an embedded audio synthesizer. Difficulty varies according to speed, number of shapes involved, game length, and score precision. The rhythm game expects the user to simply activate a midi ON message by pressing the expected tangible in the right timing range, namely when the arrow and delimiter box turn yellow. In order to play this game, all tangibles must be inserted.

Game B – Insert and Remove Shapes recalls the *Kibo's Song Mode* (see Section 3). Tangibles are associated with independent synchronized tracks within a multi-track environment realized with Pro Tools² for this purpose. The multi-track starts playing when the game is initialized but has no volume. Each track is associated with a shape and is unmuted when the corresponding tangible is inserted into the device's

wooden base. Conversely, removing the shape will mute the track. All tangibles must be removed and placed in front of the player or at a reachable distance before starting the game. For a better user experience, long game sequences must be set. The aim of Game B is to score points by inserting or extracting the *Kibo's* tangibles once the randomly generated shapes touch the hit-line or perfectly fit the delimiter box before the animation ends and the shape disappears. The random algorithm should assure that, in the case of right user performance, at the end of the session no shapes are left inside *Kibo's* body. This game mode solicits the development or recovery of some cognitive abilities, including memory (shapes have to be quickly recognized, found in the space around the *Kibo*, and inserted into the right slot) and creative reasoning (e.g., finding the best initial layout for the pieces around the *Kibo's* body).

Game C – Hold and Release Shapes recalls the *Kibo's Musical Instrument Mode* and takes benefits from the polyphonic aftertouch feature. Tangibles are mapped once again onto the grades of a C-major scale. Somehow similar to Game A, when a shape touches the vertical line or enters the delimiter box, the user is expected not only to tap the corresponding tangible but also to keep it pressed until the same shape appears again. Pressing a tangible sends a MIDI Note On message, and releasing it sends the corresponding Note Off message. This game mode is probably the most challenging one, from both a motor and a cognitive point of view. For example, some combinations of shapes require not only the ability to have them selected simultaneously but also a good strategy to have a hand free for the next insertion.

5 EARLY EXPERIMENTATION

Due to the restrictions imposed on society during the COVID-19 pandemic and the impact it had on the elderly and fragile populations, it has not been possible to test this prototype in a real-life setting so far. Nonetheless, the games were proposed to elder family members, colleagues, and domain experts to get objective feedback on their benefits and usefulness.

Overall, *Kibo Web Games* were appreciated and described as an enjoyable form of entertainment, yet challenging enough to possess rehabilitative properties. Differentiating between similarly shaped tangibles (e.g., star and flower) was initially perceived as demanding by many but overcome with practice. Another challenging aspect was maintaining the level of performance with the speed increment.

An early prototype was presented to music thera-

²<https://www.avid.com/pro-tools>

pists and physiatrists of *Fondazione Don Carlo Gnocchi*, Milan. Many useful remarks emerged about the games' structure and the user interface. Game A was particularly appreciated due to its simplicity, but the interface was considered too complex for users with impairments, above all cognitive ones. A first suggestion was to move the parameter configuration and the MIDI console away from the game screen, thus introducing a sort of back-office area to set game sessions; in fact, in a standard scenario, the parameter configuration on one side and the gameplay on the other involve two distinct actors, and the presence of side controls and messages can be a source of distraction. Similarly, the use of colors should be limited as much as possible, just to differentiate the actions required (e.g., either insert or remove a tangible). The numeric score, too, should be turned into a more direct and comprehensible representation, such as a given number of stars, a growing bar, or something similar. Games B and C involve persistent actions (i.e. shapes to keep inserted or pressed for a given time). In this case, the experts' objections focused on the graphical representation itself of the game field. In fact, in their opinion, having 8 scrolling areas corresponding to the 8 tangibles could help, above all in the case of cognitive impairment. Actually, a colored multi-track representation for Game C would recall the piano-roll visualization typically adopted by MIDI sequencers. Moreover, experts agreed that the last two games are more challenging from any point of view and, for this reason, they suggested a step-by-step process before playing the final version (e.g., first locate given shapes with no time constraint, then test the physical actions by moving shapes into their slots, etc.). Early experimentation of the revised interface should start at *Fondazione Don Carlo Gnocchi* in October 2022. Thanks to the positive feedback received, it is likely that *Kibo Web Games* have the requisites to be utilized among the elderly population. Future testing on this specific population will be necessary to fine-tune the development of a truly suitable device that can be used in the rehabilitative space. Connecting performance results to a database or exporting them into a CSV file could provide a more accurate track of progress.

The approach of *Kibo Web Games* can be easily generalized and adapted to other scenarios. For example, the platform can be profitably used to develop music parameters and geometric concepts awareness in young students, even in preschool age, thanks to TUI-based playful activities. Moreover, *Kibo Web Games* can help recover cognitive and motor abilities not only in elderly people but also in younger users; some experiences in this sense are reported in (Baratè et al., 2021).

6 CONCLUSIONS

In this project, we investigated the suitability of the *Kibo* to compensate for age impairments. The device proved to be a good solution from a technical and a physical point of view due to user-friendliness and robustness. Unfortunately, it is not an affordable product. Thanks to the *Kibo*'s 8-voice polyphony, multiple participants have the opportunity to access the interface simultaneously. Collaboration could be advantageous for seriously disabled individuals needing assistance from caregivers. However, the described games are not designed to explore this scenario in full.

The possibility to choose parameters and set difficulty levels allows users to progress at their own pace without confronting insurmountable obstacles. A therapist or a caregiver can suitably configure game sessions based on the players' attitudes and abilities. Moreover, no game penalties were implemented with the distinct purpose of not discouraging elderly users who are often not technology savvy and likely to be afraid of making mistakes or getting stuck.

The inspiration for future work mainly comes from the observations and remarks by experts reported in Section 5. Short-term future improvements include the creation of a different game field for each specific game, keeping the interface's look as simple as possible to avoid any distractions during the gameplay, and the implementation of two different windows (one for the user/patient who can only play the game and one for the specialist who can set the game parameters). In addition, tracking parameters like pressure variations could test and improve tactile and fine motor skills in action. After on-the-field experimentation, gaming speed will probably need to be adjusted to accommodate slower apprehension in aged people. Finally, an aspect to further explore is the use of proper musical feedback to be applied in music therapy.

Our expectation is that *Kibo Web Games* are able to stimulate declining cognitive abilities and reignite strained motor skills in elderly people in order to keep them active and fully integrated into society through a set of engaging activities.

REFERENCES

- Amico, M. D. and Ludovico, L. A. (2020). Kibo: A MIDI controller with a tangible user interface for music education. In *Int. Conf. on Computer Supported Education*, pages 613–619. SCITEPRESS.
- Baratè, A., Ludovico, L. A., and Oriolo, E. (2021). An ensemble of tangible user interfaces to foster music awareness and interaction in vulnerable learn-

- ers. In *Proc. 5th Int. Conf. on Computer-Human Interaction Research and Applications*, pages 48–57. SCITEPRESS.
- Bartolomeu, P., Fonseca, J., Duarte, P., Rodrigues, P., and Girao, L. (2005). Midi over bluetooth. In *2005 IEEE Conference on Emerging Technologies and Factory Automation*, volume 1, pages 8 pp.–102.
- Burke, J., McNeill, M., Charles, D., Morrow, P., Crosbie, J., and McDonough, S. (2010). Augmented reality games for upper-limb stroke rehabilitation. In *2010 Second International Conference on Games and Virtual Worlds for Serious Applications*, pages 75–78.
- Cerezo, E., Bonillo, C., and Baldassarri, S. (2020). Therapeutic activities for elderly people based on tangible interaction. In *ICT4AWE*, pages 281–290.
- Chu, R. J.-c. (2010). How family support and internet self-efficacy influence the effects of e-learning among higher aged adults—analyses of gender and age differences. *Computers & Education*, 55(1):255–264.
- Czaja, S. J. and Lee, C. C. (2007). The impact of aging on access to technology. *Universal access in the information society*, 5(4):341–349.
- Czaja, S. J. and Sharit, J. (1998). Age differences in attitudes toward computers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 53(5):P329–P340.
- De Carvalho, R. N. S., Ishitani, L., Nogueira Sales De Carvalho, R., et al. (2012). Motivational factors for mobile serious games for elderly users. *Proceedings of XI SB games*, pages 2–4.
- Dörner, R., Göbel, S., Effelsberg, W., and Wiemeyer, J. (2016). *Serious games*. Springer.
- Duckworth, J., Mumford, N., Caeyenberghs, K., Eldridge, R., Mayson, S., Thomas, P. R., Shum, D., Williams, G., and Wilson, P. H. (2015). Resonance: an interactive tabletop artwork for co-located group rehabilitation and play. In *Int. Conf. on Universal Access in Human-Computer Interaction*, pages 420–431. Springer.
- Farage, M. A., Miller, K. W., Ajayi, F., and Hutchins, D. (2012). Design principles to accommodate older adults. *Global journal of health science*, 4(2):2.
- Fonseca, X., Slingerland, G., Lukosch, S., and Brazier, F. (2021). Designing for meaningful social interaction in digital serious games. *Entertainment Computing*, 36:100385.
- Iachini, T., Borghi, A. M., and Senese, V. P. (2008). Categorization and sensorimotor interaction with objects. *Brain and Cognition*, 67(1):31–43.
- Ijsselstein, W., Nap, H. H., de Kort, Y., and Poels, K. (2007). Digital game design for elderly users. In *Proceedings of the 2007 Conference on Future Play, Future Play '07*, page 17–22, New York, NY, USA. Association for Computing Machinery.
- Ishii, H. (2008a). Tangible bits: Beyond pixels. In *Proc. 2nd Int. Conf. on Tangible and Embedded Interaction, TEI '08*, page xv–xxv. Association for Computing Machinery.
- Ishii, H. (2008b). The tangible user interface and its evolution. *Commun. ACM*, 51(6):32–36.
- Ishii, H. and Ullmer, B. (1997). Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 234–241.
- Iversen, T. R. (2015). Exploring tangible interaction: Alternative interfaces for assisting elderly users. Master’s thesis, Department of Informatics, University of Oslo.
- Jankovic, J. (2008). Parkinson’s disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, 79(4):368–376.
- Jordà, S., Geiger, G., Alonso, M., and Kaltenbrunner, M. (2007). The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proc. 1st Int. Conf. on Tangible and Embedded Interaction*, pages 139–146.
- Laguna, K. and Babcock, R. L. (1997). Computer anxiety in young and older adults: Implications for human-computer interactions in older populations. *Computers in human behavior*, 13(3):317–326.
- Maylor, E. A. (1995). Prospective memory in normal ageing and dementia. *Neurocase*, 1(3):285–289.
- Newton-Dunn, H., Nakano, H., and Gibson, J. (2003). Block jam: A tangible interface for interactive music. *Journal of New Music Research*, 32(4):383–393.
- Paradiso, J. A., Hsiao, K.-y., and Benbasat, A. (2001). Tangible music interfaces using passive magnetic tags. In *Proceedings of the 2001 Conference on New Interfaces for Musical Expression, NIME '01*, page 1–4, SGP. National University of Singapore.
- Pearce, C. (2008). The truth about baby boomer gamers: A study of over-forty computer game players. *Games and Culture*, 3(2):142–174.
- Rego, P. A., Moreira, P. M., and Reis, L. P. (2014). Architecture for serious games in health rehabilitation. In *New Perspectives in Information Systems and Technologies*, volume 2, pages 307–317, Cham. Springer.
- Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*, 2(3):179–183.
- Schietecatte, B. and Vanderdonckt, J. (2008). Audiocubes: A distributed cube tangible interface based on interaction range for sound design. In *Proc. 2nd Int. Conf. on Tangible and Embedded Interaction, TEI '08*, page 3–10, New York, NY, USA. Association for Computing Machinery.
- Vandermaesen, M., De Weyer, T., Feys, P., Luyten, K., and Coninx, K. (2016). Integrating serious games and tangible objects for functional handgrip training: A user study of handily in persons with multiple sclerosis. In *Proc. of the 2016 ACM Conf. on Designing Interactive Systems*, pages 924–935.
- Wecker, N. S., Kramer, J. H., Hallam, B. J., and Delis, D. C. (2005). Mental flexibility: age effects on switching. *Neuropsychology*, 19(3):345.
- Zhao, X., Deng, Z., Jin, M., Li, X., Zhang, W., Jin, M., Ying, F., Wang, Q., and Wang, G. (2021). Segara: Integrating serious games and handgrip for hand rehabilitation in rheumatoid arthritis patients. In *The 9th Int. Symp. of Chinese CHI*, pages 101–104.

Context-aware Personalized Decision Support based on User Digital Life Model

Alexander Smirnov^a and Tatiana Levashova^b

*St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
{smir, tatiana.levashova}@iias.spb.su*

Keywords: Decision Support, Personalized Recommendations, Personality, User Ontology, Digital Traces, User Digital Life Model.

Abstract: Digital traces is a source of information about the users and their actions while the online activities. A structured part of this source, which represents information related to the decision-making process of a user, is proposed to be formalized in the form of user digital life model. The decision support system addresses this model for information to recognize user types and recommend personalized decisions. A user type is characterized with common personality traits of the users as decision makers and common decision-making behaviours of these users as consumers. A user ontology represents a priori knowledge on the user types and supports the user classification into them. The paper considers kinds of factors influencing decision-making styles and consequently personality traits of decision makers as well as behaviour variables determining decision-making behaviour. The user digital life model provides information to score these factors and instantiate the variables. A decision support scenario is described and its application to a search problem is demonstrated.

1 INTRODUCTION

Personalized decision support based on information from user digital traces has recently gained popularity as a result of the success of the efforts on personality prediction from these traces (Stachl, Pargent, et al., 2020). Digital traces are beneficial for obtaining users' personality traits without burdensome questionnaires. Decision support systems (DSSs) and recommendation systems add the personality information in the processes of decision support and recommendation.

One of the problems of using information from digital traces is their weakly structured and poorly curated content that complicates its analysis in a context-aware meaningful way (Breiter & Hepp, 2018). Information from the digital traces represented in a well-structured way would simplify its analysis to predict personality.

Context-aware integration of information describing personality traits and online behaviour that the user manifest while decision-making to predict or

recognize personality would increase the efficiency of DSSs.

The paper proposes a conceptual framework of personalized decision support based on user digital life model, in which such a model systematizes the content of digital traces and represents information related to the decision-making process of a user. This model is used as an information source to recognize user types and recommend personalized decisions. A user type groups users with common personality traits of them as decision makers and common decision-making behaviours as consumers. A user ontology represents a priori knowledge on the user types and supports the user classification into these types. The types are context-sensitive, i.e. the same user in different contexts can be classified into different user types. A DSS that implements the conceptual framework infers the user type and recommends a decision based on the knowledge about the kinds of decisions customary for the users of this type.

The rest of paper is as follows. Section 2 outlines related research. Section 3 introduces a conceptual framework of personalized decision support based on

^a  <http://orcid.org/0000-0001-8364-073X>

^b  <http://orcid.org/0000-0002-1962-7044>

user digital life model. Section 4 presents the user ontology and provides ideas of user classification. Section 5 describes a scenario of personalized decision support according to the conceptual framework. Section 6 demonstrates the scenario applied to a book search problem. Section VII discusses main concluding remarks.

2 RELATED RESEARCH

Many studies have concluded that personalities influence human decision making process and interests (Rentfrow & Gosling, 2003).

The most commonly, research on personality prediction from user digital traces analyses traces from social media platforms. Such approaches include integrating information from Twitter on self-language usage, avatar, emoticon, and responsive patterns (Wei et al., 2017) or on profile attributes and language (Sumner et al., 2012); integrating text, image, and users' meta features from Twitter and Instagram (Skowron et al., 2016); integrating demographic features, social network activities, and language extracted from Facebook, Twitter, and YouTube (Farnadi et al., 2016); analysing text and pictures contained in traces left in multiple public and private social media platforms (Azucar et al., 2018).

Researchers agree that personality traits and individual behaviour patterns are strongly related (Augstein et al., 2019). In this direction, approaches integrate information on networks-related behaviour and personal traits (Lee & Kim, 2017; Zhao & Zhu, 2019); personal traits and emotions as a behaviour regulating factor (Lerner et al., 2015; Tkalcic, 2020); personal traits and behaviour activities of smartphone users (Stachl, Au, et al., 2020), and others.

Various personalized DSSs and recommendation systems exploit digital traces as a resource of personality information (e.g., (Courtin & Tomasena, 2016; Narducci et al., 2019)). Comprehensive reviews of personality-based recommendation systems that use information on personality traits extracted from digital traces and combine it with user behaviour can be found in (Augstein et al., 2019; Suhaim & Berri, 2021).

Multiple personalized and adaptive systems rely upon user types (or stereotypes) characterized with common personal features. Researches that support prediction of user types through classification include a classification of social media users into types that reflect the level of their engagement in the media usage (Lee & Kim, 2017), personality classification based on Twitter text (Pratama & Sarno, 2015), and

others. Ontologies, which provide classification service, support a user categorization based on personality traits, facets, culture, and age in relation to specific tasks (El Bolock et al., 2020); personality traits and facets recognized in different tests (Garcia-Velez et al., 2018); textual data contained in the digital traces (Alamsyah et al., 2021); combinations of text and social behavioural aspects of user on multiple social media (Sewwandi et al., 2017), and many others.

The present research relies upon a user digital life model as a collection of user-specific information. This model is a structured representation of a part of the content of user digital traces. Personality is recognized through ontology-based classification. A user ontology represents a priori knowledge on user types. It infers a user type based on the type definitions and information from the user digital life model. The information used concerns the user traits and online behaviour that this user manifest while decision-making.

3 CONCEPTUAL FRAMEWORK

The conceptual framework of personalized decision support based on user digital life model (Figure 1) is intended to recommend decisions that the user would made in the current situation (context). The main components of this framework are user digital traces, user profile, user digital life model, user segment, user ontology, and context (Smirnov & Levashova, 2020). The user digital life model is the main source of information for the rest of the components.

User profile is a set user characteristics that can be used to create a descriptive portrait of an individual and to identify one. *User digital traces* is a set of records fixing information on the user activities including decision-making. *User digital life model* is a structured representation of a part of the content of user digital traces, which carries information related to the decision-making process of the user. *User segment* is a group of users with common needs and behavioural reactions when making decisions. *User ontology* is a user model, which formalizes knowledge to classify a user into a user type, i.e. into a category of users distinguished by common personality traits of these users as decision makers and common decision-making behaviour. *Context* is any information that characterizes the situation of the user in the decision-making process. In the conceptual framework, context comprises the user identifying information and the information on the user preferences, the user type, the problem requiring

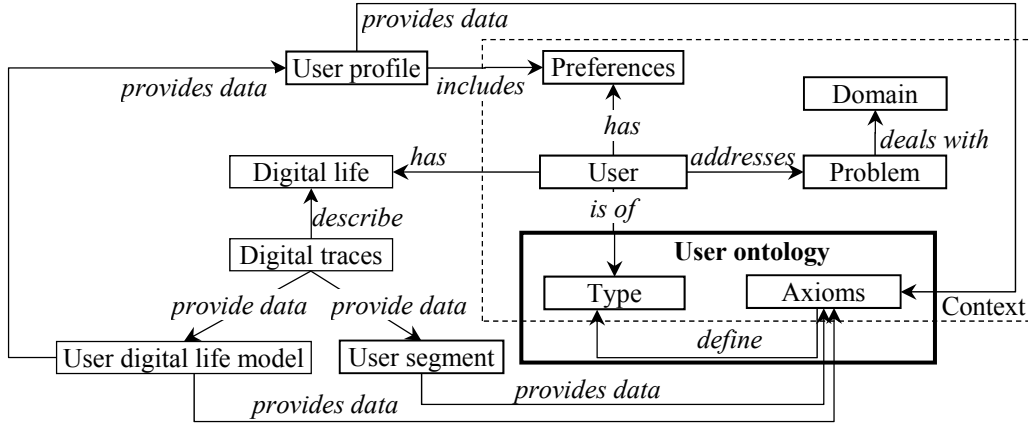


Figure 1: Conceptual framework for personalized decision support based on user digital life model.

a decision, and the knowledge domain that this problem deals with.

When the user requests the DSS with a problem requiring a decision, or when the DSS finds out that the user needs a recommendation, the system infers the user type and recommends a decision based on the knowledge about the kinds of decisions customary for the users of this type.

The framework components are modelled using the set-theoretic approach.

User profile (UP):

$$UP = (User_ID, P_out, P_in(C)),$$

$$P_in(C) = DM_Type(C) \cup P_c(C) \cup Pr(C),$$

where *User_ID* is the unique user identifier, *P_out* is the set of context-independent user characteristics; *P_in(C)* is the set of context-sensitive user characteristics in the context $C(T)$; *DM_Type(C)* is the user type in the context $C(T)$; *Pr(C)* is the set of user preferences in the context $C(T)$; *P_c(C)* is the set of context-sensitive user characteristics other than the user type and the user preferences (e.g., the user location, local time, etc.); *T* is the period of existence of the context *C*.

User digital life model (DL):

$$DL = (User_ID, Problem(t_0, t_n), Domain(t_0, t_n),$$

$$\{Action(t_a^-, t_a^+)\}, Decision(t_n), R_1, R_2, R_3),$$

$$R_1 \in Problem \times Domain, R_2 \in Problem \times$$

$$Decision, R_3 \in Action(t_a^-, t_a^+) \times Problem,$$

where *Problem* is the kind of the decision-making problem (*Problem*(t_0, t_n) means that the user addresses the problem *Problem* in the interval (t_0, t_n)); t_0 is the time instant when the user starts decision-making; t_n is the time instant when the user has made a decision; *Domain* is the knowledge domain that the problem *Problem* deals with (*Domain*(t_0, t_n) means that the domain knowledge is dealt with in the interval (t_0, t_n)); *Action*(t_a^-, t_a^+)

is the action carried out in the interval (t_a^-, t_a^+) ($t_0 \leq t_a^-, t_a^- < t_a^+, t_a^+ < t_n$); *Decision*(t_n) is the decision made at the time instant t_n .

User segment (S):

$$S = (Domain, Consumer_Type, Var, R_4, R_5),$$

$$R_4 \in Domain \times Var,$$

$$R_5 \in Var \times$$

$$Consumer_Type,$$

where *Consumer_Type* is the type of users sharing a common behavioural pattern when choosing a decision, *Var* is the set of behavioural variables providing data to the behavioural pattern.

User ontology (O_U):

$$O_U = (Cl, Rel, A),$$

$$Cl = Cl_o \cup Type, A = A_o \cup A_{DM_Type},$$

where *Cl* is the set of ontology classes, *Rel* is the set of class relationships ($Rel \rightarrow Cl \times Cl$), *Type* is the class that represents the user types, $Cl_o = Cl \setminus Type$, *A* is the set of ontology axioms, A_{DM_Type} is the set of axioms that define the membership of the class *Type* by a user, $A_o = A \setminus A_{DM_Type}$.

Context (C):

$$C(T) = (user_ID, user_type(T), domain(T), \\ problem(T), Pr_u(T), R_6), \quad (1)$$

$$user_ID \subset User_ID, \quad user_type(T) = (dm, consumer), \quad dm \subset DM_Type, \quad consumer \subset Consumer_Type, \quad domain(T) \subset Domain, \quad problem(T) \subset Problem, \quad Pr_u(T) \subseteq Pr, \quad R_6 \in domain(T) \times Pr,$$

where *user_ID* is the unique user identifier; *user_type(T)* is the user type in the context $C(T)$; *problem(T)* is the problem for that the user is making a decision in the context $C(T)$; *domain(T)* is the knowledge domain that the problem *problem(T)* deals with; *Pr_u(T)* is the set of user preferences in the context $C(T)$; $T = (t_0, t_n)$.

4 USER ONTOLOGY

The user ontology represents a priori knowledge on user types and supports the user classification.

4.1 User Types

A user type is compound. It combines the type of users as decision makers and the type of users as online consumers. Two direct subclasses of the class *Type*, that are *DM_Type* and *Consumer_Type*, represent these subtypes, respectively. The user types are context-sensitive.

The class *DM_Type* represents types of the users depending on their decision-making styles and class axioms that specify user characteristics influencing these styles. A decision-making style affects personality traits of a decision maker (El Othman et al., 2020), preferences for the selection of an alternative (Sharma & Pillai, 1996), and eventually the decision. Kinds of the decision-making styles are adopted from the management domain. According to them, decision makers can be spontaneous, rational, inert, risky, and cautious.

Detailed descriptions of the decision-making styles (Allen, 2017; Bovol'ár & Orosová, 2015; Sharma & Pillai, 1996) allowed us to find out factors that influence decision making and that can be scored based on information from digital traces (Table 1).

Number of decision makers expresses a preference for an individual or collective decision-making.

Decision making time expresses the thoroughness of the analysis and evaluation of the alternatives (this time includes the time of searching for information). It is scored as low, medium, or high.

Confidence degree scores the confidence of the user in his/her knowledge and assessments in the scale of low, medium, and high.

Complexity of decision-making procedure is the complexity of the process thought that the user has to go to reach the final decision (searching for information, analysing and evaluating the

alternatives, consulting, and decision coordinating). This procedure is proposed to be assessed as simple, medium, or complex.

Criterion is the preference criterion (latent or explicit) that the user applies to evaluate the alternatives.

The class *Consumer_Type* classifies the users as consumers of the recommended decisions based on the behavioural variables (*Var*). The classification used has come from the customer behaviour segmentation, which focuses on the division of the users into groups based on common online behaviour patterns. In the paper, it is supposed that the users play the role of Internet service consumers, the DSS is an online service that recommends decisions, and the users' behavioural patterns when they are thinking of either accept a decision or do not reflect the online behavioural patterns of these users as consumers. Progressives, consolidators, always-hurrying, traditionalists, and security-concerned consumers are represented in the class *Consumer_Type*.

Based on detailed descriptions of the consumer types, behavioural variables are identified, values for that can be found in digital traces (

Table 2). The meanings of these variables are intuitively clear; therefore, they are not described unlike the factors.

4.2 User Digital Life Model

Actions and decisions specified in the model of user's digital life are analysed to identify the information that can be used to score the factors (Table 3) and variable values (Table 4).

Actions on information search (information search requests) and communicating actions provide information to score the number of decision makers (f1) and to score the user confidence degree (f3) (a decision maker that relies only on own knowledge, is characterized by an extreme (high) degree of confidence and opposite, a decision maker that looks through large volumes of irrelevant information and

Table 1: Factors influencing decision making allocated to decision maker types.

Factor	Decision Maker Type				
	Spontaneous	Rational	Inert	Risky	Cautious
f1. Number of decision makers	one-two	group	one-two	one/group	group
f2. Decision-making time	low	medium	high	medium	high
f3. Confidence degree	high	medium	low	high	low
f4. Complexity of decision-making procedure	simple	medium	complex	medium	complex
f5. Criterion	maximizing rapidity of getting benefit	maximizing effectiveness of problem resolving	maximizing effectiveness of problem resolving	maximizing benefits	minimizing losses

Table 2: Behavioural variables influencing consumer type.

Variable	Consumer Type				
	Progressives	Consolidators	Always-hurrying	Traditionalists	Security-concerned
v1. Time spent in the Internet	much	moderate	moderate	little	little
v2. Degree of involvement in social networks	medium	high	do not use, usually	low	low
v3. Degree of Internet services consuming	high	high	medium	low	low
v4. Degree of interest to innovations	high	medium	low	low	low
v5. Loyalty level	low	medium	high	high	medium
v6. Preferable communication means	no preferences	no preferences	written	voice	voice
v7. Criterion	maximizing own benefits	maximizing benefits of other users	maximizing own benefits, minimizing time	utility maximization	minimizing losses

contacts other individuals for help is characterized by a low degree of confidence). Time values fixed for the actions on information search and on the interactions of the user with the DSS are used to calculate how much time the user gathers and analyses information, and evaluates the recommendation in order to either accept or decline it (f2). All kinds of actions used to score factors f1–f3 are used to score complexity of the decision-making procedure (f4) (if the decision is made quickly (the decision-making time is low) then the procedure is assessed as simple; if the decision-making time is high then the procedure is evaluated as complex). The result of an analysis of the user decisions is the criterion (f5), which expresses explicit or latent preferences of the user.

Table 3: Digital life model as source for factor scores.

Factor	Information from digital life model
f1	Number of individual recipients or groups to which the user sent requests
f2	Time taken to analyze the request results, and to evaluate the recommendation
f3	Kinds, number, and relevance of the knowledge sources used
f4	All the above
f5	Decision

Table 4 provides information from the user digital model that can be analysed to instantiate the behaviour variables. Durations calculated based on the time values fixed for the corresponding actions instantiate the time behaviour variables (the first three rows in the table). The degree of the Internet-services consuming is calculated as the ratio of the number of the kinds of the services used to the number of hits normalized to low, medium, or high. The loyalty level is determined by the information showing the user

interest to similar services offered by different organizations. Communication activities provide information about phone calls or written messages of a user, on the basis of which the frequency of both is calculated and the preferred means of communication is determined. The criterion is determined based on the information about the decisions that the user digital life model specifies.

Table 4: Digital life model as source for variable values.

Variable	Information from digital life model
v1	Time fixed for the on-line activities
v2	Time fixed for the social networks activities
v4	Time fixed for service hints comparing with time of the service releases
v3	Kinds of services used, service hits
v5	- Activities on searching for the services that are the same as offered by a specific site - Activities on the usage of services that are the same as offered by a specific site services on other sites - Activities on the usage of services offered by competitors
v6	Communication activities
v7	Decision

5 PERSONALIZED DECISION SUPPORT SCENARIO

The scenario of personalized decision support based on user digital life model implements the general idea of the conceptual framework and leaves aside cases when the user does not accept the recommendation. Such cases create a reason to a refinement of user types and system learning, but they are out of research

scope so far. The scenario considers situations that the DSS addresses on the interval from the time instant when the user starts decision-making until a recommendation has been provided (Figure 2).

A problem that the user faces (the decision-making problem) initiates the scenario. The information on the unique user identifier $user_ID$, the time t when the user performs the actions causing the system reaction, these actions $action(t^-, t^+)$, $action(t^-, t^+) \subseteq Action(t_a^-, t_a^+)$ (the very first action in time is $action(t_0, t^+)$), the decision-making problem $problem(t)$, and the domain $domain(t)$ is identified in the user digital traces and becomes represented in the user digital life model.

A formal problem model is built. The DSS uses this model to solve the problem $problem(t)$ as a decision support problem. The function $M: problem(t) \rightarrow PM$ assigns the problem its formal model.

Scores for the factors determining decision-making styles and values for the behaviour variables are assessed based on the user characteristics

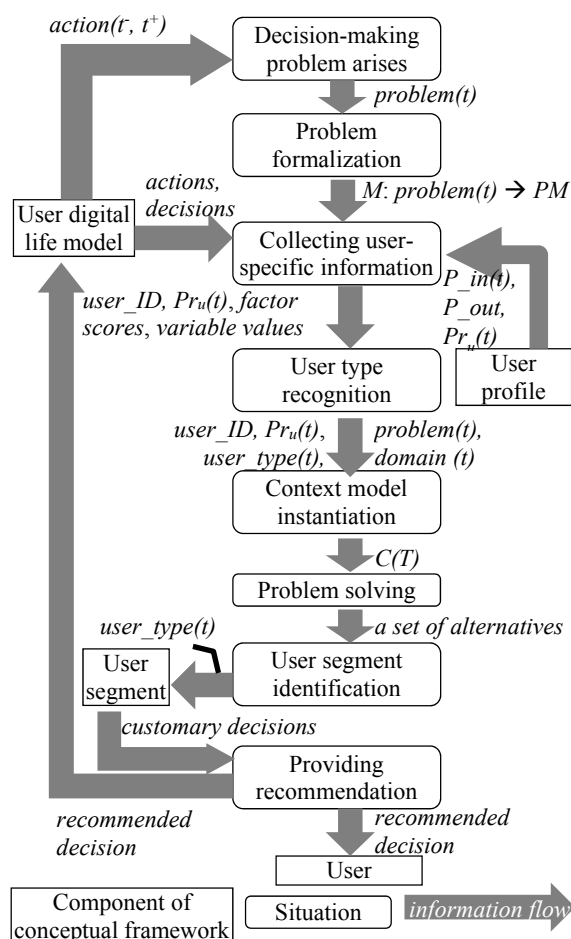


Figure 2: Scenario of personalized decision support.

contained in the user profile and the information from the user digital life model. These scores and values fully or partly instantiate the ontology axioms and rules that define the user types. The fully instantiated axioms and rules become assertions (axioms and rules describing individuals (Glimm et al., 2012)).

Based on the set of the assertions, the ontology solves the classification problem and derives the user type (*user_type(t)*) made up of the user type as decision maker and the user type as consumer.

The information on the user identifier ($user_ID$), the user type ($user_type(t)$), the kind of the problem that the user addresses ($problem(t)$), the knowledge domain ($domain(t)$), and the set of user preferences ($Pr_u(t)$) instantiates the context model $C(T)$ (1).

Using the model (*PM*), the DSS solves the problem *problem(t)* as a decision support problem. The result of problem solving is a set of alternatives.

In accordance with the user type as consumer, the segment of the users of this type is distinguished and the kind of decisions customary to the users of this segment is identified. Based on this kind of decisions a recommended decision from the set of alternatives is selected. The recommended decision is delivered to the user, appears in the user digital traces, and saved in the user digital life model:

$$DL(t_n) = (User_ID, problem(T), domain(T), \{action(t_n)\}, decision(t_n)).$$

6 USE CASE

The scenario above is demonstrated by an example of decision support for the user that searches a book on programming in Java in a library.

The user name is Alex; the part of the digital life model built for Alex based on his digital traces for the problem in question is as follows:

$$\begin{aligned} DL_{Alex}(t_0, t^+) = & (Alex, Search(t_0), Library(t_0), \{action(t_0, t^+)\}) \\ t_0 = 2020-11-19\ 19:55:16.057, \{action(t_0, t^+)\} = & \\ = \left\{ \begin{array}{l} Press_Menuitem, Search(2020-11-19\ 19:55:17.0648, 2020-11-19\ 19:55:17.926) \\ Chose_Option, Title(2020-22-19\ 19:55:16.057, 2020-11-19\ 19:55:18.873) \\ Enter_Title, Programming\ in\ Java(2020-11-19\ 19:55:19.203, 2020-11-19\ 19:55:19.936) \end{array} \right\} \end{aligned}$$

The model above specifies that the user identified as Alex ($user_ID = \text{Alex}$) addresses the problem of searching ($problem(t) = Search(t_0)$) a book entitled “Programming in Java” in the library ($domain(t) = Library(t_0)$).

The preferences of Alex coming from his profile for the problem of book searching declare that the preferable language for Alex is English. Based on the scores for the factors influencing decision-making styles and values for the behaviour variables, the user

ontology classifies Alex as a spontaneous decision maker and as an always-hurrying consumer. The scores are obtained as results of an analysis of the actions and decisions represented in Alex's digital life model. All the actions and decisions that concern the problem of books searching are analysed, not just those for the given time (for the time interval $[t_0, t^+]$ this model does not contain a decision).

Below, the instantiated context model (1) is given.

$$C(T_c) = (Alex, SA(T_c), Library(T_c), Search(T_c), language:English(T_c)),$$

$SA = (\text{spontaneous, always-hurrying})$, $T_c = [t_0, t_c]$, t_c – the moment of the context model instantiation.

A set of available in the library books in English devoted to programming in Java is the result of problem solving. The DSS uses the information that the always-hurrying users prefer quickly implementable decisions to select a recommendation from this set.

Among various options that the library suggests are benefits for its subscribers. The DSS checks in the Alex's profile that he is a library subscriber and recommends the book: "Herbert Schildt, Java: The Complete Reference, Ninth Edition, McGraw-Hill Education – Europe, 2014, 1312 p. (English, Paperback), ISBN: 9780071808552." This book is digitalized and the library can provide access to a digital copy of the book immediately after the request of a subscriber. Alex accepts the recommendation by ordering the book from the library.

Alex's digital life model is updated with the recommended decision:

$$DL_{Alex} = (Alex, Search(t_0, t_n), Library(t_0, t_n), \{action(t_0, t_n)\}, "9780071808552"), \\ (decision(t_n) = 9780071808552).$$

7 CONCLUSIONS

Personalized decision support needs user-specific information. Digital traces is a valuable resource of such information. The paper proposes a conceptual framework of personalized decision support in which the core source of information about user specifics is user digital life model. This model systematizes and contextualizes the content of digital traces that can be used to recognize personal user characteristics. Based on these characteristics the user ontology infers a user type. The DSS implementing the proposed framework refers to the user type to recommend a decision customary for the users of this type.

The novelties of the presented research are that it proposes a user digital life model, which offers a new

means for organizing weakly structured content of digital traces; suggests a way to recognise context-sensitive user types from their digital traces, and gives idea of a system dealing with context-sensitive user types to provide personalized recommendation.

Future research can address the development of decision support scenarios for the cases when the user does not accept the recommendation.

ACKNOWLEDGEMENTS

The analysis of related research and the conceptual framework for personalized decision support based on user digital life model (Sections 2, 3) are due to the grant from RFBR no. 20-07-00455, the user ontology (Section 4) and the scenario implementing the framework (Section 5) are due to the grant from RFBR no. 20-07-00490, the user digital life model derived from digital traces (Section 3) is due to State Research (project no. FFZF-2022-0005).

REFERENCES

- Alamsyah, A., Dudija, N. & Widiyanesti, S. (2021). New Approach of Measuring Human Personality Traits Using Ontology-Based Model from Social Media Data. *Information*, 12(10), 413. <https://doi.org/10.3390/info12100413>.
- Allen, M. (Ed.). (2017). Spontaneous Decision Making. In *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc. <https://doi.org/10.4135/9781483381411.n583>.
- Augstein, M., Herder, E. & Wörndl, W. (2019). *Personalized Human-Computer Interaction*. De Gruyter. <https://doi.org/10.1515/9783110552485>.
- Azucar, D., Marengo, D. & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, 124, 150–159. <https://doi.org/10.1016/j.paid.2017.12.018>.
- Bavol'ár, J. & Orosová, O. (2015). Decision-making styles and their associations with decision-making competencies and mental health. *Judgment and Decision Making*, 10(1), 115–122. <http://journal.sjdm.org/14/141103/jdm141103.pdf>.
- Breiter, A. & Hepp, A. (2018). The Complexity of Datafication: Putting Digital Traces in Context. In *Communicative Figurations* (pp. 387–405). Springer International Publishing. https://doi.org/10.1007/978-3-319-65584-0_16.
- Courtin, C. & Tomasena, M. (2016). A benchmarking platform for analyzing corpora of traces. *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, 178–186. <https://doi.org/10.1145/3006299.3006303>.

- El Bolock, A., Herbert, C. & Abdennadher, S. (2020). CCOnto: Towards an Ontology-Based Model for Character Computing. In F. Dalpiaz, J. Zdravkovic & P. Loucopoulos (Eds.), *Research Challenges in Information Science. RCIS 2020* (Vol. 385, pp. 529–535). Springer. https://doi.org/10.1007/978-3-030-50316-1_34.
- El Othman, R., El Othman, R., Hallit, R., Obeid, S. & Hallit, S. (2020). Personality traits, emotional intelligence and decision-making styles in Lebanese universities medical students. *BMC Psychology*, 8(1), 46. <https://doi.org/10.1186/s40359-020-00406-4>.
- Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F. & De Cock, M. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, 26(2–3), 109–142. <https://doi.org/10.1007/s11257-016-9171-0>.
- Garcia-Velez, R., Galan-Mena, J., Lopez-Nores, M. & Robles-Bykbaev, V. (2018). Creating an Ontological Networks to Support the Inference of Personality Traits and Facets. *2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 1–4. <https://doi.org/10.1109/INTERCON.2018.8526453>.
- Glimm, B., Horrocks, I., Motik, B., Shearer, R. & Stoilos, G. (2012). A novel approach to ontology classification. *Journal of Web Semantics*, 14, 84–101. <https://doi.org/10.1016/j.websem.2011.12.007>.
- Lee, J. & Kim, G. (2017). Social media user classification. *Proceedings of the International Conference on Electronic Commerce*, 1–3. <https://doi.org/10.1145/3154943.3154963>.
- Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, 66(1), 799–823. <https://doi.org/10.1146/annurev-psych-010213-115043>.
- Narducci, F., Basile, P., Iovine, A., De Gemmis, M., Lops, P. & Semeraro, G. (2019). A Framework for building Chat-based Recommender Systems. In A. Cuzzocrea, F. Bonchi & D. Gunopulos (Eds.), *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)* (Vol. 2482). <http://ceur-ws.org/Vol-2482/paper47.pdf>.
- Pratama, B. Y. & Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. *2015 International Conference on Data and Software Engineering (ICoDSE)*, 170–174. <https://doi.org/10.1109/ICoDSE.2015.7436992>.
- Rentfrow, P. J. & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6), 1236–1256. <https://doi.org/10.1037/0022-3514.84.6.1236>.
- Sewwandi, D., Perera, K., Sandaruwan, S., Lakchani, O., Nugaliyadde, A. & Thelijjagoda, S. (2017). Linguistic features based personality recognition using social media data. *2017 6th National Conference on Technology and Management (NCTM)*, 63–68. <https://doi.org/10.1109/NCTM.2017.7872829>.
- Sharma, A. & Pillai, R. (1996). Customers' Decision-Making Styles and their Preference for Sales Strategies: Conceptual Examination and an Empirical Study. *Journal of Personal Selling & Sales Management*, 16(1), 21–33. <https://doi.org/http://dx.doi.org/10.1080/08853134.1996.10754042>.
- Skowron, M., Tkalcic, M., Ferwerda, B. & Schedl, M. (2016). Fusing Social Media Cues. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 107–108. <https://doi.org/10.1145/2872518.2889368>.
- Smirnov, A. & Levashova, T. (2020). Intelligent Decision Support Based on User Digital Life Model: Principles and Conceptual Framework. *2020 27th Conference of Open Innovations Association (FRUCT)*, 247–252. <https://doi.org/10.23919/FRUCT49677.2020.9211055>.
- Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B. & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30), 17680–17687. <https://doi.org/10.1073/pnas.1920484117>.
- Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D. & Bühner, M. (2020). Personality Research and Assessment in the Era of Machine Learning. *European Journal of Personality*, 34(5), 613–631. <https://doi.org/10.1002/per.2257>.
- Suhaim, A. Bin & Berri, J. (2021). Context-Aware Recommender Systems for Social Networks: Review, Challenges and Opportunities. *IEEE Access*, 9, 57440–57463. <https://doi.org/10.1109/ACCESS.2021.3072165>.
- Sumner, C., Byers, A., Boochever, R. & Park, G. J. (2012). Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. *2012 11th International Conference on Machine Learning and Applications*, 386–393. <https://doi.org/10.1109/ICMLA.2012.218>.
- Tkalcic, M. (2020). Complementing Behavioural Modeling with Cognitive Modeling for Better Recommendations. In D. Helic, G. Leitner, M. Stettinger, A. Felfernig & Z. W. Raš (Eds.), *Foundations of Intelligent Systems. ISMIS 2020* (Vol. 12117, pp. 3–8). Springer. https://doi.org/10.1007/978-3-030-59491-6_1.
- Wei, H., Zhang, F., Yuan, N. J., Cao, C., Fu, H., Xie, X., Rui, Y. & Ma, W.-Y. (2017). Beyond the words: predicting user personality from heterogeneous information. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 305–314. <https://doi.org/10.1145/3018661.3018717>.
- Zhao, N. & Zhu, T. (2019). Psychological Profiles Prediction Using Online Social Network Behavior Data. In Z. Yan (Ed.), *Analyzing Human Behavior in Cyberspace* (pp. 75–87). IGI Global. <https://doi.org/10.4018/978-1-5225-7128-5.ch005>.

Providing Personalised Recommendations of Critical Incident Narratives in a Cross-platform Mobile Application

Tim Wenzel¹, Doris Fetscher², Wolfgang Golubski¹, Susanne Klein² and Rainer Wasinger¹ ^a

¹*Faculty of Physical Engineering/Computer Sciences, University of Applied Sciences, Zwickau, Germany*

²*Faculty of Applied Languages and Intercultural Communication, University of Applied Sciences, Zwickau, Germany*
{tim.wenzel.gai, doris.fetscher, wolfgang.golubski, susanne.klein.fjq, rainer.wasinger}@fh-zwickau.de

Keywords: Personalisation, User Modelling, Mobile Applications, HCI, Prototyping, Critical Incident Narratives.

Abstract: This work describes the design and implementation of a cross-platform mobile application that has been created to provide users with personalised recommendations of Critical Incident (CI) narratives. CIs provide brief descriptions of situations in which misunderstandings arise as a result of the cultural differences of the interacting parties. They are useful for increasing intercultural awareness. This paper describes the design and implementation of the mobile application called 'Nils2Go'. The main focus is the personalisation strategy that is employed in the recommendation of CIs, and the identification and use of user demographic characteristics that can be utilised to further personalise a retrieved list of CI narratives.

1 INTRODUCTION

Various studies show that the use of mobile applications often drops sharply within the first few weeks and months after installation. For example, (Yan and Chen, 2011) show how the average usage drops by about 50% over a period of three months. Although also dependant on the type of application (e.g. communication applications are usually used for significantly longer than games that are used exclusively for entertainment (Li et al., 2020)), 'user experience' has been shown to influence long-term application usage (Kujala et al., 2011).

This paper focuses on the question "How can CI recommendations in a mobile application be personalised?". This is important because personalisation has been shown to lead to an improvement in user experience in the past. To achieve this in the context of a mobile application, a user model is created, and personalisation in the form of a recommender system is carried out to present users with a list of relevant CIs.

This paper starts with a definition of Critical Incidents. Following this, Section 2 outlines the background concepts of personalisation, user modelling, and recommender systems, and also provides an outline of the Network Intercultural Learning and Sensitivity (NILS) website. In Section 3, the low-fidelity prototype and the implemented mobile application

called Nils2Go¹ are presented, as too the personalisation strategy and technical implementation of the application. Finally, in Section 4, our conclusions and directions for future work are provided.

1.1 Critical Incidents

Critical Incidents (CIs) are brief descriptions of situations in which a misunderstanding, problem, or conflict arises as a result of the cultural differences of the interacting parties (Apedaile and Schill, 2008). CIs form an important tool for increasing our awareness and understanding of human attitudes, expectations, behaviours, and interactions. Some examples of CIs are provided in Figures 1, 2B, and 3.

In addition to the narrative, different metadata about the CI is also stored. For example, it can be relevant where the information came from, who recorded it, and who the actors are. A reflection (i.e. a retrospective view of the past situation) by the author from today's perspective, is also often included in a CI.

The mobile application described in Section 3 incorporates both the CI narrative and its associated metadata into the presentation of CIs to the user. This metadata is of decisive importance in the personalisation of the CIs that are returned to the user and is described in more detail in Section 2.4.1.

¹Nils2Go App: Available at <https://play.google.com> and <https://apps.apple.com>.

^a  <https://orcid.org/0000-0003-1069-5814>

2 BACKGROUND

In this section, the concepts of personalisation, user modelling, and recommender systems are outlined, as too the NILS website. These concepts form the basis on which this work is then built upon.

2.1 Personalisation

There are various approaches to increasing the user experience in the area of mobile applications. One of these approaches is the personalisation of the software used. In this context, personalisation involves assigning relevant characteristics to individual users. Based on these characteristics, different users will have different experiences when using the software. The experiences are adapted to the preferences and needs of the user. The work in this paper focuses on how personalisation can be used in the presentation of Critical Incident narratives.

A sub-area of the personalisation of software is represented by so-called recommendation systems. These systems suggest specific information to the user based on matches to his or her personal interests. In general, personalisation deals with the discrepancy between general-purpose applications, which achieve the highest possible benefit in a broad field with the least possible development effort, and specialised applications, which should fulfil an individual-specific benefit (Kuo, 2013).

2.1.1 Advantages and Disadvantages of Personalisation

The development of personalised applications has a number of advantages. One clear advantage, especially for software development companies, is driven by finance. In (Gavril and Ionescu, 2017), the authors demonstrate that personalised applications generate 16.5% to 24% higher revenues. Another advantage that concerns the end user is the increase in user satisfaction. In (Liang et al., 2006), the authors show that personalisation, especially in the form of recommendation systems, has a positive effect on user satisfaction. The authors in (Tong et al., 2012) also demonstrate the positive influence that personalisation has on user satisfaction in software systems. Another positive aspect is that personalised applications often stand out from non-personalised applications (Arora et al., 2008).

In addition to the advantages mentioned, there are also various disadvantages. The disadvantages include the higher development costs and the increased complexity of the development of an application, but

these are often offset by higher revenues (Bleichschmidt et al., 2005). Another possible disadvantage, especially for the end user, is data privacy, as personal data is often collected and used for user modelling (Bleichschmidt et al., 2005). It is however often the case that the positive aspects outweigh the negative ones. In the Nils2Go application described in this paper, user data that is collected never leaves the device.

2.2 User Modelling

Whereas personalisation is the process of providing individually relevant and interesting information for individual users of an application, user modelling refers to the construction of a User Model (UM) that incorporates user-specific data. This model is generated and managed by means of corresponding software components. The collected and managed data can refer to the demographic characteristics of the user being modelled. Furthermore, contextual data of the user can also be relevant. The resulting model reflects assumptions of the system regarding the modelled user and can often be retrieved in the form of a user profile. Depending on the application, such a model can also be designed for reuse in other systems (Kay, 1998).

Examples of user model data relevant to the Nils2Go application that is described in this work include demographic data (e.g. age, gender, countries of interest, spoken languages, and country of residence), general pre-defined topics of interest that the user can select from (e.g. education, food, family, holidays, and travel), and specific interests provided in the form of user-defined keywords. This data is shown in Figure 5D.

In (Vu and Proctor, 2011; Hothi and Hall, 1998), the following types of User Model are outlined:

- **Static UMs:** This is the simplest type of UM. Once the necessary data and information has been provided, the UM remains unchanged for entirety of the application's remaining use. Changes in user interests and preferences are not registered or updated in the model.
- **Dynamic UMs:** Dynamic user models can accommodate for changes in a user's interests and preferences. In this way, there is always an up-to-date model of the user. Furthermore, a user's interactions with the application can also flow into the model. Depending on the use-case, it is also possible to manually adjust the model.
- **Stereotype-based UMs:** Stereotype-based UMs enable a much more anonymous way of generating and maintaining user models. Here, the interests and characteristics of the individual user are

not part of the modelling. Rather, users are classified into common stereotypes. General information about these groups is then used to generate preferences for the user. The basis for these assumptions are primarily driven by statistics that have been compiled for the respective groups of people.

- **Highly Adaptive UMs:** Highly adaptive UMs realise the counterpart to stereotype-based modelling. Here, great amounts of information about a specific user are collected and generated, such that almost all anonymity of the user is lost. One advantage of this UM is that user profiles are extremely specific and accurate. As a result, corresponding applications can be particularly well adapted to the current user and the user receives the best possible user experience.
- **Hybrid UMs:** Hybrid UMs attempt to combine the advantages of both static and dynamic UMs, such that a balance is found between good performance with low memory requirements and good personalisation with detailed user models. Often, as is the case in (Billsus and Pazzani, 2000), the UM consists of sub-models that take both the current point in time as well as a longer period of time into account. This allows information about a user to be collected in order to first generate a static user model, while also allowing for dynamic changes in the interests over a longer period of time.

As described later in Section 3.4 the mobile application in this work incorporates a hybrid UM approach.

2.3 Recommender Systems

In (Melville and Sindhvani, 2010), a Recommender System is defined as a software system that provides relevant and meaningful recommendations or suggestions of different products or information items to a number of users who may be interested in them. Recommender systems usually make use of either collaborative filtering, contextual filtering, or a combination of both. For the purpose of brevity, the reader is directed to (Aggarwal, 2016) for further details on recommender systems. The mobile application described in this work incorporates a content-based filtering approach as is described in Section 3.4.

2.4 NILS: Network Intercultural Learning and Sensitivity

Network Intercultural Learning and Sensitivity (NILS) is a research project of the University of Applied Sciences Zwickau. The work focuses on a scientifically innovative preparation of critical incident narratives (Fetscher and Klein, 2020). In addition to gaining new insights into intercultural communication, the aim of the project is to sensitise users by helping them to question cultural stereotypes. It is assumed that this goal can be achieved through the interactive browsing of CIs (Fetscher and Klein, 2020).

The NILS website² has been primarily designed for research, teaching, and learning purposes. It allows researchers (as well as other registered users) to enter CI narratives into the system. It also allows users of the website to retrieve CIs and filter the CIs based on keyword searches. There are currently around 200 CIs available through the website.

Data entry and retrieval of CI narratives is accessible via the project's website, with the work described in this paper providing the missing link to further view CIs in a personalised manner via a cross-platform mobile application that has been developed for Android and iOS smartphones and published on the platforms' respective App store fronts. An example CI narrative from the website is shown below in Figure 1, while Figures 2B and 3 in Section 3 show further example narratives available through the mobile application.

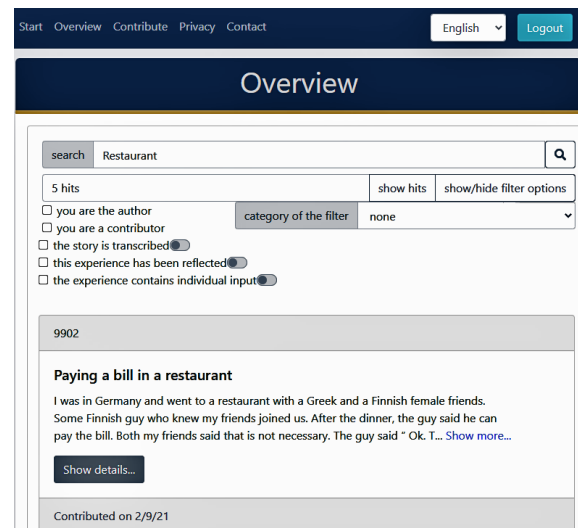


Figure 1: NILS website illustrating a search for relevant CIs based on the keyword “Restaurant”.

²NILS Website: <https://nils.fh-zwickau.de/>

2.4.1 Critical Incident Metadata

Each CI has up to 105 attributes associated with it. These attributes can be grouped into the categories: textStory, origin, medium, media, hotspots, contact domains, communication domains, reflection, author, actors, and location. Each of these categories has multiple attributes associated with it, e.g. a textStory includes the narrative (labelled as textStory.story in the dataset) as well as attributes depicting whether it has been transcribed, the language it is written in, and details about the author.

Some of the more important attributes that are collected for each CI include an *ID*, the *title* and the actual *story text*, a *timestamp* of when the data was created, the *origin* (i.e. whether the story is a personal experience, an observation, retelling, or hearsay), the *language* in which the content is recorded, the *kind* of data (i.e. whether it is primary data or secondary/edited data), the narrative *perspective* (i.e. first or third person), the *location* of the event, the *actors* to whom the narrative refers, and personal data about the *author*. Authors also have the ability to reflect on their experience by including a *reflection* at a later time. Authors can also provide *hotspots* in the form of free-text keywords to describe features of the CI (e.g. humorous, historical).

In general, the CIs can be written in any language. In addition to German, the current set of CIs contain some narratives in languages such as English, French, Spanish, and Russian. The actors in the set of CIs come from various European and non-European countries and belong to different age groups.

3 Nils2Go MOBILE APPLICATION

3.1 Purpose and Functionality

The mobile application described in this paper has two main purposes. Firstly, it should allow for the clear presentation of CI narratives included in the NILS dataset. Both the actual CI story as well as relevant metadata (see Section 2.4.1) should be made available to the user. Secondly, the mobile application should offer the possibility to personalise all existing CI narratives within the given framework. In this way, relevant elements classified as interesting for the user are selected from the total amount of data (though still allowing the user to access other CIs via a simple toggle button as shown in Figure 5A).

In contrast to the website, the mobile application has been kept deliberately simple and lightweight. It

does not provide the ability to add new CIs, and neither does it allow users to create an account from the mobile device (which would be required to add new CIs). Instead, it has been designed to present CIs quickly and efficiently, and in a manner that is personalised to a given user's profile.

3.2 Low-fidelity Prototype

The mobile application was developed using a User-Centred Design (UCD) approach (Norman, 2013). This process was combined with a semester-long course at the University of Applied Sciences Zwickau³ that focused on the creation of mobile application prototypes for the search and retrieval of CIs. In particular, the course focused on HCI principles in which 13 groups of students designed low-fidelity prototypes in Balsamiq⁴ and iteratively tested their designs, first in a formative and later in a summative manner with 3-5 participants each time. The evaluation technique used for this process was the think-aloud technique (Shneiderman, 2016). These low-fidelity prototypes were used as the starting point for the design of the current mobile application, which then additionally incorporated the concept of personalisation as shown in Figure 2, with the ability to select thematic interests (A) and to browse a list of personalised CIs (B).

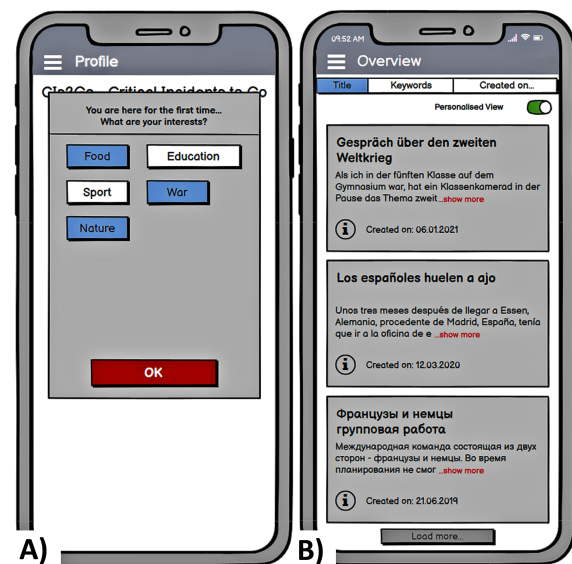


Figure 2: Low-fidelity prototype of the mobile application showing a user's thematic interests (A) and a personalised list of CIs (B).

³WHZ, URL: <https://www.fh-zwickau.de/>

⁴Balsamiq, URL: <https://balsamiq.com/wireframes/>

3.3 High-fidelity Mobile Application

The mobile application has two main functions. The first is to provide access to the list of CIs. These CIs are packaged locally with the mobile application on compilation. This means that there is no reliance on the Internet or the NILS website to access the CIs. The set of CIs included with the mobile application are however only a subset of the total available to registered users of the NILS website. Updates to the included list of CIs is accomplished via application updates to the respective App store fronts. Figures 3A, 3B, and 3C show the *overview* list in which CIs are arranged vertically in an abbreviated form. This means that for each CI, the title and the first three lines of the actual story are visible. Furthermore, the creation date for each narrative, as well as an information-button that shows the related metadata is provided. To see the complete story, the user simply selects the 'show more' button. Feedback is also provided (when in the personalised mode) to indicate if the match is due to 'Interest Theme', 'Keyword', or 'Demographic characteristic' (Figures 3A, 3B, and 3C).

In addition to the CI narrative, the user also has access to the *metadata* associated with each CI (Figure 3D), which the user can swipe through via a right-to-left swipe gesture (in this case to see details on the CI's Hotspots - Figure 3D top-left) and via a left-to-right swipe gesture (to see details on the CI's Author - Figure 3D top-right). In this view, all of the CI's additional metadata is shown. This information is presented separately to the main story so that it does not impact on the readability of the CI narratives in the overview page.

Also visible in Figure 4 is the ability to view favoured CIs and to sort CIs, both alphabetically, and via the manner in which CIs were selected as being relevant, i.e. based on the 'Field of interest', 'Demographic Features', or 'Keywords'.

The second function is related to the personalisation of the retrieved list of CIs. The information contained in the *user profile* is decisive for the personalisation of the application and contains important data collected about the user.

Figure 5B shows a multi-choice set of topics that is provided to the user during their first use of the application. This allows the application to determine the user's thematic interests for the purpose of personalisation. The user also has the ability to skip this process and configure their profile at a later time via the application's side-menu. Also visible in Figure 5A is the 'Personalised View' toggle-switch that allows a user to see the CIs that are most relevant to them.

The user's profile page is accessed via the side-

menu and is shown in Figure 5D. This provides a representation of the user model, which is statically generated at the beginning and - if necessary - adapted over the application's useful life (see Section 2.2). The sections on the user profile page are divided into demographic characteristics and interests. The demographic characteristics list personal information about the user including their age, gender, countries that they affiliate with, spoken languages, and their current location of residence. The ability to enter this data is mainly provided to the user via pre-defined drop-down lists that limit the amount of free-form text that the application needs to interpret. The section on interests provides predefined topics that the user selected on first use of the application, as well as user-defined keywords that can be provided in free-form. At the bottom of the screen is also a button that allows the user to delete all of the information that has been gathered on him or her. If the user selects this button, their intention to delete their profile must be confirmed and the user model is subsequently and irrevocably removed from the application.

3.4 Personalisation in the Application

Section 2.2 outlined static, dynamic, stereotype-based, highly customisable, and hybrid approaches to user modelling. For the Nils2Go mobile application described in this paper, a hybrid UM approach was used. The user model is generated during the first use of the application, similar to a static UM approach. The user can then 'manually' adjust their interests and demographic characteristics, which makes it possible to update the UM without complex dynamic adjustment functionalities. Users can voluntarily further customise their UM with demographic data like age, gender, and spoken languages.

Users can also decide not to provide such demographic data, but the list of returned CIs will then be much less precise as the personalisation strategy will not incorporate CI metadata based on the actors and authors with similar demographic attributes to the user. If the user does initially decline to provide demographic data, they can still decide to add it later via the application's side-menu.

The recommender system used in this application uses a content-based filtering approach. In particular, recommendations are implemented based on rules specified in the software. These rules use the Levenshtein string distance algorithm (Navarro, 2001) to determine similarities between the user's profile data (Table 1, left) and the CIs (Table 1, right).

Using the data in Table 1 row 1 to illustrate, one such rule finds similarities between a user's thematic

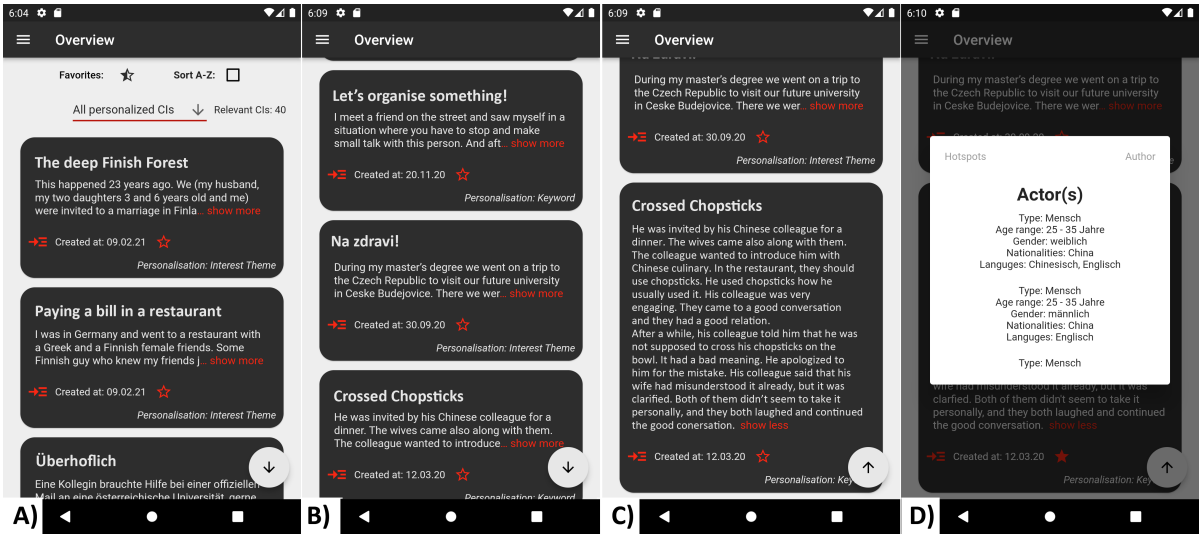


Figure 3: The mobile application for Android, showing the user’s personalised list of CIs (A, B, C), and some of the CI metadata (D).

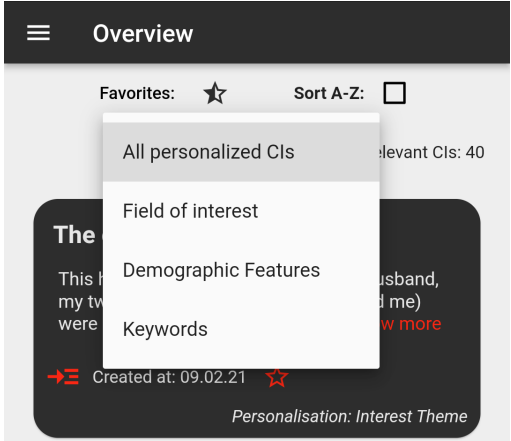


Figure 4: Methods of sorting CIs, including both alphabetic (A-Z) and personalised options.

topics of interest and the CI text story and reflection text. Similarly, if a user provides details on the language(s) that they speak, this will be used to help determine relevant CIs based on the language spoken by the actors, the author, and the language that the CI has been written in (row 5 in the table).

From a technical perspective, the recommendation of relevant CIs is based on the Levenshtein string distance algorithm. This is well suited to applications in which the objective is to find matches for short strings (e.g. our pre-defined themes or user-defined keywords) in longer texts (i.e. our dataset of Critical Incidents). In this mobile application, Levenshtein distance thresholds are used to adjust the relevance of the returned CIs, i.e. if words are less than or equal to the Levenshtein distance, the CI is more likely to be deemed relevant to the user. From a code perspective,

Table 1: Recommender system data. Demographic data is marked with an asterisk (*).

User Profile	Critical Incidents (CI) + Metadata
Thematic Interests	Text story, Reflection
Individual Keywords	Title, Text story, Reflection
Age*	Age (Actors, Author)
Gender*	Gender (Actors, Author)
Languages*	Language (Actors, Author), Language (CI)
Countries*	Nationality (Actors, Author), Country (CI)
Place / Region*	Place (CI), Region (CI)

this is represented by the following variable:

```
const MAX_LEVENSHTEIN_DISTANCE_TO_MATCH = 3;
```

In addition to the Levenshtein distance threshold (which can also be adjusted by the user as shown in Figure 5C), CIs are also determined to be relevant based on the user’s demographic data (when available). As outlined in (Beel et al., 2013), the use of demographic data can have a significant impact on the success of the recommender system. In our mobile application, it is a combination of several matching demographic characteristics that is required for a CI to be marked as being relevant based on the user’s demographic data (Figure 5C).

3.5 Technical Implementation

In this section, we outline the framework and technologies that were used to build the cross-platform mobile application.

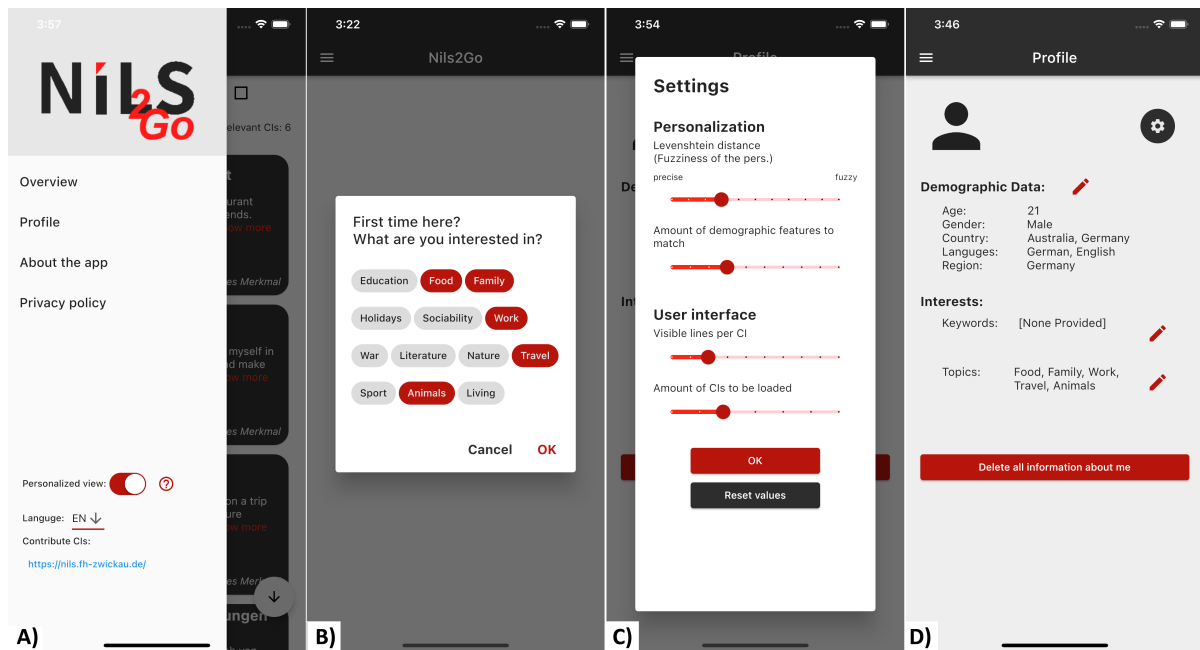


Figure 5: The mobile application for Apple iOS, showing the side-menu (A), the user’s thematic interests (B), personalisation settings (C), and the user profile (D).

For the technical implementation of the mobile application, the open source framework Flutter⁵ was used in conjunction with the programming language Dart⁶. Flutter is a portable UI toolkit developed by Google for the development of natively-compiled applications for mobile platforms (iOS, Android) as well as desktop (MacOS, Windows, Linux) and web (Web Apps) (Flutter, 2022). Most importantly, this means that the framework offers the ability to develop applications once and then make them available natively on different platforms.

Cross-platform development technologies like Flutter and Dart allow developers to reuse the same code across multiple platforms. They also provide access to native features of the smartphone, including local storage, which is relevant for this mobile application.

The Nils2Go application is based on Flutter v3, which was released in May 2022. According to a 2022 survey by Statista of over 31,743 software developers (Vailshery, 2022), the use of the Flutter framework by software developers increased to 42% between 2019 and 2021 (an increase of 12%). Flutter is thus now already far ahead of competing frameworks such as React Native (used by 38% of software developers in the survey), Cordova (used by 16%), Ionic (used by 16%) and Xamarin (used by 11%).

⁵Flutter, URL: <https://flutter.dev/>

⁶Dart, URL: <https://dart.dev/>

In comparison to web-based cross-development platforms like React Native, Flutter uses its own high-performance rendering engine rather than web technology. The Flutter engine is written in C/C++ and applications written in the Dart programming language are compiled into native code Ahead-of-Time (AoT) for both iOS and Android⁷. This focus on performance means that Flutter applications can run faster than the competing cross-platform technologies.

A comparison of Flutter applications written in Dart compared to native Android applications written in Kotlin and iOS applications written in Swift shows that there is still a performance penalty when using Flutter (Olsson, 2020), but the advantage of having a cross-platform application outweighs the disadvantages for the purposes of this work. In (Olsson, 2020), it is further stated that from their survey of 39 people from the IT industry, 74% of end users could not detect the difference in look and feel between a Flutter and a native mobile application, which provides even more incentive to use Flutter/Dart.

4 CONCLUSIONS

This work described the design and implementation of the Nils2Go cross-platform mobile application that

⁷Flutter FAQ, URL: <https://flutter.dev/docs/resources/faq>

has been created to provide users with personalised recommendations of CI narratives. The focus was on the personalisation strategy which was demonstrated using a hybrid user model and a content-based recommender system that incorporated the Levenshtein distance as well as demographic user data (when available). Future work will now be to conduct a user-study with the mobile application to see if the employed personalisation strategy does in fact lead to an increase in user experience for its users as has been shown by other past research.

ACKNOWLEDGEMENTS

A special thank you goes to the PTI06970 HCI students (Alec Sichenender, Jonas Langner, Justin Nt-zold, Konrad Platz) and PTI06640 Mobile Applications students (Kenny Musterer, Felix Schober, Dominik Strer, Tommy-Lee Bannert) at the Westschische Hochschule Zwickau, who worked on earlier prototypes of this application. Special thanks also go to the NILS development team (Tobias Haubold, Wolfgang Grs, Peter Huster, and Dariya Kapinus) and to our graphic designer (David Mantilla), who provided the colour schemes, graphics, and icons.

REFERENCES

- Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.
- Apedaile, S. and Schill, L. (2008). Critical Incidents for Intercultural Communication. Technical report, NorQuest College Intercultural Education Programs.
- Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., Joshi, Y., Kumar, V., Lurie, N., Neslin, S., et al. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, 19(3):305–321.
- Beel, J., Langer, S., Nürnberger, A., and Genzmehr, M. (2013). The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems. In *International Conference on Theory and Practice of Digital Libraries*, pages 396–400. Springer.
- Billsus, D. and Pazzani, M. J. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10(2):147–180.
- Blechschimidt, T., Wieland, T., Kuhmunch, C., and Mehrmann, L. (2005). Personalization of End User Software on Mobile Devices. In *Second IEEE International Workshop on Mobile Commerce and Services*, pages 130–137. IEEE.
- Fetscher, D. and Klein, S. (2020). Wissensplattform für Critical Incidents und interkulturelle Erfahrungen [Knowledge Platform for Critical Incidents and Intercultural Experiences]. In *Enjeux et défis du numérique pour l'enseignement universitaire*, pages 79–92. Peter Lang, Berlin.
- Flutter (2022). Flutter: Design beautiful apps.
- Gavril, L. and Ionescu, S. (2017). Financial Advantages of Software Personalization. *Journal of Information Systems & Operations Management*, 11(1):207–217.
- Hothi, J. and Hall, W. (1998). An Evaluation of Adapted Hypermedia Techniques using Static User Modelling. In *Proceedings of the Second Workshop on Adaptive Hypertext and Hypermedia*, pages 45–50.
- Kay, J. (1998). *A Scrutable User Modelling Shell for User-Adapted Interaction*. PhD thesis, Basser Department of Computer Science, Faculty of Science, University of Sydney.
- Kujala, S., Roto, V., Vninen, K., Karapanos, E., and Sin-nel, A. (2011). UX Curve: A method for evaluating long-term user experience. *Interacting with Computers*, 23:473–483.
- Kuo, T. C. (2013). Mass customization and personalization software development: a case study eco-design product service system. *Journal of Intelligent Manufacturing*, 24(5):1019–1031.
- Li, T., Zhang, M., Cao, H., Li, Y., Tarkoma, S., and Hui, P. (2020). "What Apps Did You Use?": Understanding the Long-term Evolution of Mobile App Usage. In *Proceedings of the Web Conference 2020*, pages 66–76.
- Liang, T.-P., Lai, H.-J., and Ku, Y.-C. (2006). Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings. *Journal of Management Information Systems*, 23(3):45–70.
- Melville, P. and Sindhvani, V. (2010). Recommender Systems. *Encyclopedia of Machine Learning*, 1:829–838.
- Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM Computing Surveys*, 33(1):3188.
- Norman, D. (2013). *The Design of Everyday Things*. Basic Books.
- Olsson, M. (2020). A Comparison of Performance and Looks between Flutter and Native Applications. Bachelor's thesis, Faculty of Computing, Blekinge Institute of Technology, Sweden.
- Shneiderman, B. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson.
- Tong, C., Wong, S. K.-S., and Lui, K. P.-H. (2012). The Influences of Service Personalization, Customer Satisfaction and Switching Costs on E-Loyalty. *Int. Journal of Economics and Finance*, 4(3):105–114.
- Vailshery, L. (2022). Cross-platform mobile frameworks used by software developers worldwide in 2019 to 2021. Technical report, Statista.
- Vu, K.-P. L. and Proctor, R. W. (2011). *Handbook of Human Factors in Web design*. CRC Press.
- Yan, B. and Chen, G. (2011). AppJoy: personalized mobile application discovery. In *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, pages 113–126.

A Study of Search User Interface Design based on Hofstede's Six Cultural Dimensions

Karen Chessum¹^a, Haiming Liu²^b and Ingo Frommholz³^c

¹*School of Computer Science and Technology, University of Bedfordshire, Luton, U.K.*

²*School of Electronics and Computer Science, University of Southampton, U.K.*

³*School of Engineering, Computing and Mathematical Sciences, University of Wolverhampton, Wolverhampton, U.K.*
karen.chessum@beds.ac.uk, h.liu@soton.ac.uk, ifrommholz@acm.org

Keywords: Cross-cultural Information Retrieval, Cross-cultural Theory, Website Design, Human-Computer Information Retrieval (HCIR), Hofstede's Cultural Dimensions, Human-Computer Interaction (HCI).

Abstract: An information seeker's cultural background could influence their preference for search user interface (UI) design. To study cultural influences Geert Hofstede's cultural dimensions have been applied to website design for a number of years. In this paper, we examine if Hofstede's six cultural dimension can be applied to inform the design of search engine user interfaces. The culturally designed search user interfaces have been evaluated in a study with 148 participants of different cultural backgrounds. The results have been analysed to determine if Hofstede's cultural dimensions are appropriate for understanding users' preferences on search user interface design. Whilst the key findings from the study suggest Hofstede cross-cultural dimensions can be used to model users' preferences on search interface design, further work is still needed for particular cultural dimensions to reinforce the conclusions.

1 INTRODUCTION

A user's cultural orientation could influence their preferences on user interface (UI) design, as noted by (Reinecke et al., 2010). Contemporary search engines such as Google, Bing, and Yahoo!, do not differentiate between different user types or the cultural backgrounds of their users. Research conducted by Slone (2002, p 1166) states, "Both motivation and experience, elements of goals and mental models, played equally strong roles in this result. In fact, goals and mental models work in tandem to determine overall searching behaviour". A user's cultural background influences their mental model formation.

One means of defining culture is mental programming. Mental programming can be thought of as patterns of thinking, feeling and actions, based upon what we have learned throughout our lifetime. Hofstede (1991, p 4) writes, "Much of it has been acquired in early childhood". Thinking patterns, feelings and actions which have been learned over

one's lifetime, once learned have to be unlearned before new patterns can be absorbed. Hofstede, (1991, p 5) also noted, "Culture is learned not inherited". Culture is different from human nature on one hand and from personality on the other hand. Culture as described by Hofstede (1991, p 5) is "the collective programming of the mind which distinguishes the members of one group or category of people from another".

In our work, we examine cross-cultural aspects in search UI design. The overall aim of our research is to study the potential differences and different preferences between cultures when it comes to search UI. Due to the reported importance of Hofstede's work in international communication, (Wardrobe, 2005) international management, (Bing, 2004) international marketing, (Mooij and Hofstede, 2010) and use in website design, (Marcus and Gould, 2000: Liu, 2021), we base our considerations on Hofstede's model.

Our contribution is as follows: firstly, we discuss how we have used Hofstede's six cultural dimensions

^a <https://orcid.org/0000-0002-2980-8746>

^b <https://orcid.org/0000-0002-0390-3657>

^c <https://orcid.org/0000-0002-5622-5132>

to inform our prototype search UI design. Secondly, in our study we use the prototype UIs we designed based on Hofstede's dimensions to determine if Hofstede's cultural dimensions and index scores for different countries match the user preference of the search UI design. For example, if Hofstede's cultural dimensions Index scores indicate a country's culture is towards the high end of the Masculinity dimension, do users from this cultural background actually prefer a search UI designed with high Masculinity in mind?

To accomplish our aims, the remainder of the paper is structured as follows: to justify our choice of applying Hofstede's dimensions, we briefly review different cultural models in the next section. Subsequently, we discuss by example how Hofstede's dimensions informed the design of our prototype UIs. Next, we present our study and its results to answer the question of whether Hofstede's index scores can be used to indicate the search UI preferences of users from different cultural backgrounds. Finally, we offer our conclusion.

2 CULTURAL MODELS

Several cultural models have been critically reviewed for the suitability of this research, which is to effectively model different cultures and be able to inform the design of cross-cultural search UIs. Below are the details on what they are and why we decide to continue our investigation with Hofstede's model.

2.1 Hall

Edward Hall, an anthropologist, was a pioneer in cross-cultural business communication. Hall (1976) defined culture as using 'high context' (HC) and 'low context' (LC). A high context communication, as noted by Smith et al., (2004) is where, "little has to be said or written because most of the information is either in the physical environment or within the person, while very little is in the coded, explicit part of the message". Liu (2021) notes 'people from high-context cultures prefer face-to-face communication' and continues by saying high-context cultures 'look for both less-direct verbal and subtler nonverbal cues during the communication'.

This high-low context for cultures refers to how information is stored and flows. Whereas in a 'low context' culture the information contained in the message is explicit, little is hidden.

Hall identified the Primary Message Systems (PMS). These systems are non-lingual ways in which humans communicate with one another. Hall

identified 10 PMS each relating to a facet of human activity, (Hall, 1990). However, it is Hall's 'high-context' and 'low-context' work that is most cited within a Human-Computer Interaction (HCI) perspective.

2.2 Trompenaars & Hampden-Turner

Trompenaars and Hampden-Turner also looked at culture from a dimension level and defined seven cultural dimensions. They took Parson's five relational orientations, Parsons (1951) as a starting point.

These dimensions are units that can be used to make comparisons and are as follows: Universalism vs Particularism, Individualism vs Communitarianism, Specific vs Diffuse, Neutral vs Emotional, Achievement vs Ascription, Sequential time vs Synchronous time and Internal direction vs Outer direction.

2.3 Nisbett

Nisbett, a social psychologist, examines the differences between Eastern and Western cultures. As noted by Oshlyansky (2007), Nisbett looks at the "processes of thought, perception, attention, organisation of knowledge, understanding" and other mental processes. He uses 'holistic' and 'analytic' thought patterns or mental processes to distinguish between Eastern and Western cultures, with the West on the analytics side and East/Asian on the holistic side. Nisbett and Miyamoto, (2005) says, "the evidence indicates that people in Western cultures focus on salient objects and use rules and categorization for purposes of organizing the environment. By contrast, people in East Asian cultures focus more holistically on relationships and similarities among objects when organizing the environment."

2.4 Hofstede

Geert Hofstede, a Dutch anthropologist, carried out in-depth interviews with hundreds of IBM employees in 53 countries. He identified four cultural dimensions and published his research at the end of the 1970s. The fifth dimension Long-term Time Orientation (LTO) was added in 1991 from work carried out by Michael Harris Bond, supported by Hofstede. An additional sixth dimension, Indulgence versus Restraint (IND) added in 2010 is described in Hofstede et al., (2010) following the analysis of the World Values Survey data (from the World Values

Survey Organisation) from 93 countries by Michael Minkov. Hofstede et al., (2010) contains details of all six dimensions.

2.5 Why Hofstede?

Geert Hofstede's cultural research is probably one of the most well-known cultural models and as noted by Ghemawat and Reiche, (2011) the most widely used. Hofstede created six dimensions by which cultures can be compared, (Reid, 2015). Hofstede's work has been used in a number of disciplines, examples of which are global branding and advertising', (Mooij and Hofstede, 2010), consumer behaviour, (Milner et al., 1993), management control systems, (Chatterjee, 2014), cross-cultural psychology, (Hofstede, 2011) and cross-cultural HCI research (Smith et al., 2004). As noted by Oshlyansky (2007) when discussing cultural models used in HCI states, "By far the most popular of these models is Hofstede's". Smith et al., (2004) also state, "Hofstede's (1991) dimensions of culture that are the most often quoted theories in relation to cross-cultural usability". Mooij and Hofstede (2010) assert "People perform information-seeking tasks faster when using web content created by designers from their own culture". Mooij and Hofstede (2010) go on to say, culturally adapted websites are more usable, and users are more likely to have a more positive outlook toward them.

3 HOFSTEDE'S CROSS CULTURAL THEORY

Examples of how Hofstede's work has been used in HCI, are shown by Marcus and Gould (2000), who used Hofstede's five-dimensional model to develop a set of website design guidelines and Smith et al., (2004) who incorporated Hofstede's dimensions in their process model. Research undertaken by Chessum et al., (2014) looked at Hofstede's five dimensions to group users for cross-cultural information retrieval. However, since then, Hofstede has added a sixth dimension, called 'Indulgence v Restraint' (IND), to his cultural model.

This work examines whether Hofstede's dimensions can be used in the design of search UIs. The attributes of Hofstede's six dimensions have been researched extensively in relation to HCI by the authors of this paper. We considered the following interpretations of the six dimensions (user interface

designs UI 1 to UI 12¹) applicable to Human-Computer Interaction (HCI) design.

There has been generally less research conducted regarding Hofstede's most recent, sixth dimension, Indulgence v Restraint (IND). Many works conducted did not include this last dimension. Examples are (Marcus and Gould, 2000), (Smith et al., 2004), (Chessum et al., 2014) and (Karreman et al., 2016). Hofstede's additional sixth dimension is included in this research and the descriptions and references given below (Sections 3.1 to 3.6).

3.1 Power Distance (PD)

Power Distance is the amount of unequal power within a culture that members of that culture are prepared to accept or expect. User interface designs for UI 1 and UI 2 are given below:

3.1.1 UI 1 High (PD)

- Images of Experts, official buildings, official logos, prominence given to security and restrictions (Marcus and Gould 2000).
- Structured website design (Burgmann et al. 2006).
- "Older people are both respected and feared" use images of older people for wisdom and credibility (Hofstede 2011).

3.1.2 UI 2 Low (PD)

- Status is displayed to leaders rather than the population, staff or consumers. Information hierarchy is shallow (Marcus and Gould, 2000).
- Use a looser structure to allow users to explore your site for themselves (Nahai, 2013), (Marcus and Gould, 2000).
- Use earned evaluations e.g. ratings, testimonials, likes, to promote your goods or services (Nahai, 2013).
- "Older people are neither respected nor feared" Show images of younger or youthful people (Hofstede, 2011).

3.2 Individualism (IDV)

Individualism within a culture is where the individual is expected only to take care of themselves and their immediate family. There is no expectation for them to take care of anyone else, as opposed to a collectivist society, where members take care of extended

¹ Examples of UI 1 to UI 12 can be seen at <https://github.com/ifromm/cross-cultural-ui-designs/>

families and other group members. User interface designs for UI 3 and UI 4 are given below:

3.2.1 UI 3 High (IDV)

- “High text-to-image ratio”. Avoid cluttered graphics. Show positive images of goal achievement (Gould et al., 2000).
- “Create competitions and challenges to engage your customers”. “Give visitors a sense of personal achievement to motivate actions”. Have content that has ‘novelty’ and ‘difference’ in order to ‘attract attention’ (Nahai, 2013).
- Have their own personal goals. Follow their likes and dislikes (Sinha, 2014).
- “Speaking one's mind is healthy” (Hofstede 2011).

3.2.2 UI 4 Low (IDV)

- “High image-to-text ratio” (Gould et al., 2000).
- Transparency, give users full disclosure, for example how their data would be used. “Show that you respect privacy and security of personal info”. “Engage the community – ‘we’ not ‘me’” (Nahai, 2013).
- Emphasis on social and organisational goals.
- An individual's goals are less important (Gould et al., 2000).
- Members of a collective society, aspire to achieve their in-groups' goals (Sinha, 2014).
- “Harmony should always be maintained” (Hofstede, 2011).

3.3 Masculinity (MAS)

With this dimension, Hofstede refers to gender roles rather than to physical gender. Masculine roles consist of assertiveness, toughness and competition. Masculine work objectives incorporate “earnings, recognition, advancements and challenge”, as noted by Marcus (2002). While feminine roles are traditionally ones with an emphasis on caring for the home, family/children, people and tenderness are considered prevalent. User interface designs for UI 5 and UI 6 are given below:

3.3.1 UI 5 High (MAS)

- User attention obtained by games and competitions. Work tasks, roles, and skills, quick results obtained for limited actions. Navigation focused on exploring but also on control (Marcus and Gould, 2000).

- Masculine societies are competitive. Motivated by achievement, heroism, assertiveness, and materialism (Idler, 2013).
- “Admiration for the strong” (Hofstede 2011).
- Bright contrasting colours (Voehringer-Kuhnt, 2002), (Dormann and Chisalita, 2002).

3.3.2 UI 6 Low (MAS)

- User attention is obtained by the use of poetry, aesthetics, and appealing to uniting values (Marcus and Gould, 2000).
- Provide contact information and be prepared for feedback and questions. “This group is very cooperative and if they want to give feedback, they don't hesitate to get in contact with you”.
- “Feminine societies are consensus-oriented”. With a preference for values, corresponding to cooperation, modesty, care for the weak, and quality of life (Idler, 2013).
- “Sympathy for the weak” (Hofstede, 2011).
- Pastel colours, low saturation (Voehringer-Kuhnt, 2002, (Dormann and Chisalita, 2002).

3.4 Uncertainty Avoidance (UA)

Uncertainty avoidance relates to the extent a culture is either comfortable or uncomfortable with uncertainty or unknown situations. User interface designs for UI 7 and UI 8 are given below:

3.4.1 UI 7 High (UA)

- Tries to show/predict the results or effects of actions before the user acts. Navigation structures are designed to help prevent users from becoming lost. Any ambiguity can be decreased by the use of “Redundant cues”, e.g. design, sound visual aids (Burgmann et al., 2006).
- Simplicity, with clear metaphors, restricted options, and limited volume of data (Marcus and Gould, 2000).
- “The uncertainty inherent in life is felt as a continuous threat that must be fought”. “Need for clarity and structure” (Hofstede, 2011).

3.4.2 UI 8 Low (UA)

- Information is maximised by the use of colour coding, typeface, font, and sound. Use multiple links but not redundant cueing. Limited control over navigation e.g. Links could open content in new windows that lead away from the original webpage(s). Complexity with maximum content

and options. Acceptance of exploring and risk (can even be encouraged), with a stigma on “over- protection.” (Marcus and Gould, 2000).

- “The uncertainty inherent in life is accepted and each day is taken as it comes”, “Comfortable with ambiguity and chaos” (Hofstede, 2011).

3.5 Long-term Time Orientation (LTO)

Also known as, Long-term Orientation versus Short-term Normative Orientation (LTO). This dimension was identified later by Hofstede and Bond (1984), where Bond had a questionnaire re-designed, with a Chinese culture bias. This he called the Chinese value survey, (CVS). LTO is a Confucian philosophy, where members value long-term gain over short-term gain. User interface designs for UI 9 and UI 10 are given below:

3.5.1 UI 9 High (LTO)

- Offer ways for the user to save browsing history, e.g., wish lists. Together with means of sharing on social media. Persons with long-term orientation decisions are comprehensive and grounded “for the future” (Idler, 2013).
- Patience shown in attaining results and reaching goals. “Relationships as a source of information and credibility” (Marcus and Gould, 2000).
- “Perseverance in achieving results” (Makkonen, 2012).
- “Thrift and perseverance are important goals”. “Large savings quota, funds available for Investment” (Hofstede, 2011).

3.5.2 UI 10 Low (LTO)

- Users require quick results that are consistent with known values and traditions. Persons with a short-term orientation would appear “to live more in the past and in the present than in the future” (Idler, 2013).
- Persons from a very short-term oriented culture e.g. Spain have a tendency “to live in the moment” (Nahai, 2013).
- A wish for instant results and achieving goals. “Rules as a source of information and credibility” (Marcus and Gould, 2000).
- “Focus on achieving quick results” (Makkonen, 2012).
- “Service to others is an important goal”. “Social spending and consumption” (Hofstede, 2011).

3.6 Indulgence vs Restraint (IND)

The additional sixth dimension, relates to happiness, freedom of expression and feeling in control of your own life, (Hofstede, 2011). User interface designs for UI 11 and UI 12 are given below:

3.6.1 UI 11 High (IND)

- Use and encourage user-generated content. “Make interactions fun”. “Reflect loose gender roles by using a range of models” (Nahai, 2013).
- People from an Indulgent culture have a tendency to put an emphasis on individual happiness and wellbeing. Their leisure time is more significant and people experience more freedom and “personal control” (MacLachlan, 2013).
- Maintaining order in the nation is not given a high priority. A perception of personal life control. Freedom of speech is seen as important (Hofstede, 2011).

3.6.2 UI 12 Low (IND)

- Frugal, show how they can save money. “Emphasise how you serve the community”. “Strict, cultured gender roles” (Nahai, 2013).
- People from a restrained culture do not display positive emotions as easily, with freedom, happiness and leisure time not assigned the same significance (MacLachlan, 2013).
- Higher number of police officers per 100,000 population. A perception of helplessness: what happens to me is not my own doing. Freedom of speech is not a primary concern (Hofstede, 2011).

4 SEARCH INTERFACE DESIGN

As discussed above, the characteristics of Hofstede's six cultural dimensions have been used to create twelve prototype UIs. These consist of two UIs created for each dimension, with one interface set for the lower end and one for the higher end of each dimension. (e.g., “high masculinity”, “low masculinity”).

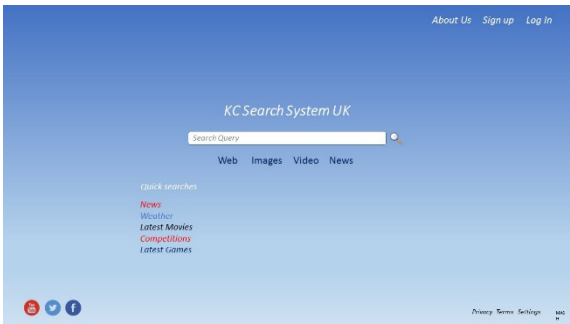


Figure 1: Prototype High Masculinity (MAS) UI 5.

Here we offer two examples of how Hofstede’s Masculinity dimension, both, high MAS shown in Figure 1 and low MAS shown in Figure 2, have been applied to the search user UI design². The prototype UI design was constructed using the design features described in 3.3 above. How the design features are implemented are shown below in tables 1 and 2 respectively.

Table 1: High Masculinity.

High (MAS) HCI Design Feature	HCI Design Feature Implementation UI 5
User attention obtained by games and competitions. (Marcus and Gould, 2000)	This has been achieved by, showing text links for ‘Competitions’ and ‘Latest Games’
Work tasks, roles, and skills, quick results obtained for limited actions. (Marcus and Gould, 2000)	‘Quick’ search textual links are provided, offering quick results for popular searches.
Navigation focused on exploring but also control. (Marcus and Gould, 2000)	‘Quick’ searches and search links to ‘Web’, ‘Images’, ‘Video’ and ‘News’
Masculine societies are competitive. Motivated by achievement, heroism, assertiveness, and materialism. (Idler, 2013)	A non-cluttered interface with textual links to News, Weather and Latest movies. Also, textual links for ‘Competitions’ and ‘Latest Games’. Only graphical image icons are for ‘YouTube’, ‘Twitter’ and ‘Facebook’, allowing faster access to these social media organisations.
Admiration for the strong” (Hofstede, 2011)	A general masculine ‘look and feel’.
Bright contrasting colours. (Voehringer-Kuhnt, 2002),	Bold colours such as red, blue, dark blue and black

(Dormann and Chisalita, 2002)	have been used for ‘Quick’ search textual links. With a contrasting white for ‘About Us’, ‘Sign up’ and ‘Log In’. With black textual links for ‘Privacy’, ‘Terms’ and ‘Settings’.
-------------------------------	---

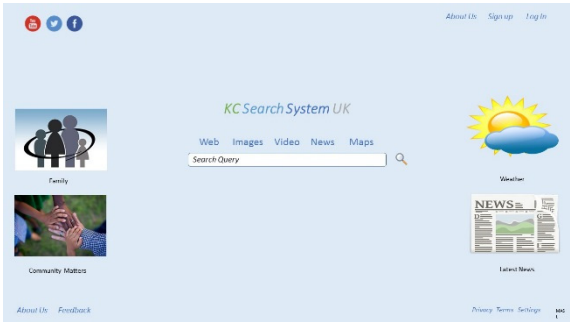


Figure 2: Prototype Low Masculinity (MAS) UI 6.

Table 2: Low Masculinity.

Low (MAS) HCI Design Feature	HCI Design Feature Implementation UI 6
User attention is obtained by the use of poetry, aesthetics, and appealing to uniting values. (Marcus and Gould. 2000)	A general ‘softer’ appearance with more and larger images and icons.
Provide contact information and be prepared for feedback and questions. (Idler, 2013)	Two ‘About Us’ links and a ‘Feedback’ link.
“Feminine societies are consensus-oriented”. With a preference for values, corresponding to cooperation, modesty, care of the weak, and quality of life. (Idler, 2013)	A ‘Community Matters’ link and imagery showing multicultural inclusion. ‘Family’ link and imagery.
“Sympathy for the weak” (Hofstede, 2011)	A general softer, less masculine aesthetic ‘look and feel’
Pastel colours, low saturation. (Voehringer-Kuhnt, 2002), (Dormann and Chisalita, 2002)	The use of pastel colours for background, ‘Weather’ icon, ‘News’ icon, Logo, and textual system links, such as ‘About Us’, ‘Sign up’, ‘Log in’, ‘Feedback’, ‘Privacy’, ‘Terms’ and ‘Settings’

² Examples of UI 1 to UI 12 can be seen at <https://github.com/ifromm/cross-cultural-ui-designs/>

This paper examines the data obtained from the research of the twelve prototype search UIs, designed using Hofstede's six cultural dimensions.

5 USER EVALUATION

Participants of this survey have been taken mostly from the staff, the current student base and the alumni of the University of Bedfordshire, England. However, this survey was also promoted internationally via social media by a number of University staff. The survey was completely anonymous. The data for this study was collected via an online survey. Our survey was constructed using the following two sections.

5.1 Survey Section One

This section consists of potentially seven questions depending upon the user's responses. This collected general background information about the participant and consisted of closed questions relating to gender, age, occupation, and culture most identified with. In addition, there were several questions relating to languages spoken and place of residence.

5.2 Survey Section Two

In this section of the survey, the participant/user is exposed to the twelve prototype UIs, these consisting of two for each of the six dimensions and are sequentially numbered User Interface (UI) 1 to User Interface (UI) 12.

The users were asked to pick their preferences from the twelve UIs designed using the above design features (Sections 3.1 to 3.6). The search UI'S were paired for each of the six dimensions, with one UI being the low-end design and the other UI being the high-end design for all six dimensions, i.e. two UIs per dimension making twelve in total. The user could only pick one for each pair, as the responses are mutually exclusive. This paper is based upon the findings from this section together with the participant's nationality and the index scores for Hofstede's six dimensions.

5.3 Hypothesis

Our six hypotheses have been developed from both the low and high aspects of Hofstede dimensions and are detailed below:

H1: Higher PD Countries will show a preference for UI 1 and Lower PD Countries for UI 2.

H2: Higher IDV Countries will show a preference for UI 3 and Lower IDV Countries for UI 4.

H3: Higher MAS Countries will show a preference for UI 5 and Lower MAS Countries for UI 6.

H4: Higher UA Countries will show a preference for UI 7 and Lower UA Countries for UI 8.

H5: Higher LTO Countries will show a preference for UI 9 and Lower LTO Countries for UI 10.

H6: Higher IND Countries will show a preference for UI 11 and Lower IND Countries UI 12.

5.4 Data Analysis

The data was collected via our survey as described in 5.2 where the user was asked to select a preference for one of the UI pairs. The participant's nationality is collected in section one of the survey as described in 5.1 above. Hofstede's country index scores have been applied to participants who identify with the matching nationality. These index scores are applied for each nationality for all six dimensions.

Once the data had been collected, a quantitative data analysis tool has been to analyse the data.

6 EVALUATION RESULTS

We had 148 participants who completed our survey, made up of 101 participants who at present are residing in the UK and 47 who are residing overseas. We had 97 male and 51 female participants.

The 148 participants are from 33 countries. Unfortunately, many of the countries only had 1 to 2 respondents and as such we have not included their results. Likewise, several countries not having an Index score calculated by Hofstede, have also been excluded. The countries and cultures results analysed are as follows, the number of participants given in brackets. U.K. (Great Britain) (51) Germany (21), Poland (3), Pakistan (10), Nigeria (5), Bangladesh (3), Ethiopia (3), China (6), Nepal (5), Sri Lanka (3) and India (7), with 117 participants in total.

6.1 Analysis of Preferences for Each User's Culture

6.1.1 Hypothesis 1 Power Distance

Both Germany and the UK are considered to be low PD countries (both Hofstede's country's Index scores 35), which means a culture that supports the concept that inequalities in their society be kept to a minimum. Therefore, the expectation for H1 (Section 5.3) would be that such countries prefer search UI 2.

This was not found in our data, with only 12% (UK) resp. 14% (Germany) prefer UI 2.

Pakistan would be considered close to the central point for a PD country with an Index score of 55. Here, H1 (Section 5.3) has been partially supported with 80% of users showing a preference for UI 1 and 20% for UI 2.

From the countries with a high PD according to their Index score, our data confirmed the expected preference for UI 1 for Poland, Nigeria, Ethiopia, China, Nepal, Sri Lanka and India.

Against their expected preference, participants from Bangladesh seem to not prefer UI 1. However, we have to note that we only had a few participants from these countries.

6.1.2 Hypothesis 2 Individualism

UK, Germany, and Poland are considered countries with a high IDV score, which, according to our hypothesis H2 in 5.3, means UI 3 would be preferred.

While this is supported in the case of Poland, we observe this is not the case for the UK and Germany.

India would be considered close to the central point for an IDV country with an index score of 48. H2 has not been supported with 86% of users showing a preference for UI 4 and 14% for UI 3.

Pakistan, Nigeria, Bangladesh, Ethiopia, China, Nepal, and Sri Lanka are low IDV countries according to Hofstede's Index scores, accordingly we would expect a preference for UI 4. This is confirmed, except for Bangladesh and Ethiopia.

6.1.3 Hypothesis 3 Masculinity

Hypothesis H3 (Section 5.3), suggests that higher MAS Countries will show a preference for UI 5 and lower MAS Countries will show a preference for UI 6.

According to Hofstede's Index score, the UK, Germany, Poland, Nigeria, Ethiopia and China are considered high masculine countries – we would expect a preference for search UI 5 for these countries. This is confirmed for the UK, Germany and Ethiopia, whereas preferences are mixed for China and India and not confirmed for Nigeria.

Pakistan, Bangladesh and India are considered close to the central point for a MAS country. Indeed, no clear preference for either of the two search UIs could be determined by our data as would be expected.

Nepal and Sri Lanka are countries with a low MAS index score. While the data does not allow for determining a clear preference, there is a slight tendency towards the (expected) search UI 6.

6.1.4 Hypothesis 4 Uncertainty Avoidance

Hypothesis H4 in 5.3, says that higher UA Countries will show a preference for UI 7 while lower UA Countries will show a preference for UI 8.

Poland, Pakistan, Germany and Bangladesh are considered countries with high uncertainty avoidance. Participants from these countries indeed exhibited a preference for UI 7, as expected.

Sri Lanka, Nigeria and Ethiopia are countries considered to be in the middle range when it comes to the UA score. We would not expect any clear preference for any of the two UIs. However, this has not been confirmed in the case of Nigeria whose participants exhibit a preference for UI 8. However, Ethiopia with an index score of 55, did show a preference for UI 7.

The UK, China, Nepal and India are countries with a low UA score. Except in the case of Nepal, our data disagrees with the expected preference for UI 8.

6.1.5 Hypothesis 5 Long-term Time Orientation

Long-term time orientation cultures value virtuous behaviour, perseverance and patience for achieving goals and results. Hypothesis H5 suggests that higher LTO Countries will show a preference for UI 9 and lower LTO countries will exhibit a preference for UI design 10.

Germany and China are considered countries with a high long-term orientation, according to Hofstede's Index score. However, our data do not confirm the expected preference for search UI 9.

Countries in the middle range when it comes to LTO are the UK, Pakistan, Bangladesh, India and Sri Lanka. We can argue the expectation of no clear preference for either UIs is confirmed (though Pakistan shows a slight tendency for UI 9).

Countries with a low LTO score are Poland and Nigeria, both exhibiting the expected preference for UI 10.

There is no LTO index score reported by Hofstede for Ethiopia and Nepal.

6.1.6 Hypothesis 6 Indulgence

For the Indulgence (IND) dimension, our hypothesis H6 states that higher IND countries will show a preference for UI 11 while lower IND countries will show a preference for search UI 12.

Nigeria and the UK are both considered high indulgence countries. Based on our data, we regard H6 to be confirmed due to the preference for UI 11.

Germany, Bangladesh, India and China have a low IND score according to Hofstede's Index. Preferences for these countries are rather mixed, so H6 is not confirmed in these cases. However, it has been confirmed for Poland, also considered a low IND country. There is no IND index score reported by Hofstede for Ethiopia, Nepal and Sri Lanka.

7 DISCUSSION AND CONCLUSION

Overall, we have found 33 out of the possible 61 results that support Hofstede's Index scores. A further 12 results were difficult to fully categorise but partially support the hypothesis. 16 results do not support the respective hypothesis.

Firstly, Hofstede's Index scores do not address search UIs but had a different focus originally. One of our contributions is exactly to verify to what degree these scores can be applied to search UIs. It seems, from our data, that Hofstede's Index scores show the potential to inform search UIs, but they also show further research needs to be carried out to shed some light on the reasons why we get at times inconclusive results and how we can better inform culturally aware search UI design. We consider our study as an important contribution to triggering this discussion.

Secondly, we had a limited number of participants from some countries, although our study attracted many participants in particular from the UK (51), Germany (21) and Pakistan (10). However, looking at the data, it would seem the most supported dimension across the countries reported in this study is Long-term Time Orientation, with 7 from 9 countries (2 countries did not have Hofstede Index scores for this dimension) being confirmed. With Power Distance having 6 confirmations and 2 partially confirmed and Masculinity having 5 confirmations with 4 partially confirmed. It indicates there is a stronger link between specific cultural dimensions and search UI design. This aspect requires further research, possibly using the 3 most popular dimensions and more participants.

ACKNOWLEDGMENTS

We would like to thank all the participants for contributing to this user study.

REFERENCES

- Bing, J.W. (2004) Hofstede's consequences: The impact of his work on consulting and business practices, An Executive Commentary by John W. Bing, Academy of Management Executive, February 2004, Vol. 18, No. 1.
- Burgmann, I., Kitchen, P., and Williams, R. (2006) Does Culture Matter on the Web? Marketing Intelligence & Planning (24:1), 62-73.
- Chatterjee, D. (2014) Management Control Systems and Hofstede's Cultural Dimensions: An Empirical Study of Innovators and Low Innovators. Global Business Review. 15(3) 565-582.
- Chessum, K., Liu, H., and Frommholz, I. (2014) Applying Cross-cultural theory to understand users' preferences on interactive information retrieval platform design. In Proceedings euroHCIR.
- Dormann, C., and Chisalita, C. (2002) Cultural Values in Web Site Design, in Proceedings of the 11th European Conference on Cognitive Ergonomics, Catania, Italy, September 8-11.
- Ghemawat, P., and Reiche, S. (2011) National Cultural Differences and Multinational Business, Globalization Note Series.
- Gould, E., Zakaria, N., and Yusof, S. (2000) Applying Culture to Website Design: A Comparison of Malaysian and US Websites. Paper Presented at the Proceedings of IEEE Professional Communication Society International Professional Communication Conference and Proceedings of the 18th Annual ACM International Conference on Computer Documentation. Technology and Teamwork 162-171.
- Hall, E.T. (1976) Beyond Culture. Doubleday, Garden City, New York.
- Hall, E.T., (1990) The Silent Language. New York: Anchor Books.
- Hofstede, G. (1991) Cultures and Organization: Software of the Mind, Paperback Edition published 1994, HarperCollinsBusiness, An imprint of HarperCollinsPublishers, 77 - 85 Fulham Palace Road, Hammersmith, London W6 8JB.
- Hofstede, G. (2011) Dimensionalizing Cultures: The Hofstede Model in Context. Online Readings in Psychology and Culture, 2(1). <https://doi.org/10.9707/2307-0919.1014>
- Hofstede, G., and Bond, M. (1984) The Need for synergy among Cross-cultural Studies. Journal of Cross-Cultural Psychology 15. (4), 417-433 December 1984.
- Hofstede, G., and Minkov, M. (2013) Value Survey Module 2013 Manual. <https://geerthofstede.com/wp-content/uploads/2016/07/Manual-VSM-2013.pdf>
- Hofstede, G., Hofstede, G.J., and Minkov, M. (2010) Cultures and Organizations: Software of the mind: intercultural cooperation and its importance for survival. The McGraw Hill Companies. USA.
- Idler, S. (2013) How to Design for a Cross-Cultural User Experience (part 2/2) 23 April 2013. <https://usabilla.com/blog/designing-for-a-cross-cultural-user-experience-part2/>

- Ingwersen, P., and Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.
- Karremen, J., Romeo, P., Li, Q. (2016) Cross-Cultural HCI and UX Design: A Comparison of Chinese and Western User Interfaces *working paper 2016*.
- Liu, F. (2021) Modify Your Design for Global Audiences: Crosscultural UX Design. Nielson Norman Group. <https://www.nngroup.com/articles/crosscultural-design/>
- MacLachlan, M. (2013) Indulgence vs. Restraint – the 6th Dimension. <https://www.communicaid.com/cross-cultural-training/blog/indulgence-vs-restraint-6th-dimension/>
- Makkonen, E. (2012) Cultural differences and localization in user interfaces. Thesis for Bachelor of Engineering (B.E.), Media technology. Central Ostrobothnia University of Applied Sciences.
- Marcus, A., and Gould, E. (2000) Crosscurrents Cultural Dimensions and Global Web User-Interface Design. *Interactions*, (7) 4, 32-46. DOI= <http://dl.acm.org/citation.cfm?doid=345190.345238>.
- Milner, L., Fodness, D., and Speece, M. (1993) Hofstede's Research on Cross-Cultural Work-Related Values: Implications For Consumer Behavior. *European Advances in Consumer Research*, 1, 70-76. <http://acrweb.org/volumes/11610/volumes/e01/E-01>
- Mooij, M., and Hofstede, G. (2010) The Hofstede model Applications to global branding and advertising strategy and research. *International journal of advertising* 29(1), 85-110.
- Nahai, N. (2013) How to Sell Online to Individualist vs Collectivist Cultures. *Psychology Today*. <https://www.psychologytoday.com/gb/blog/webs-influence/201307/how-sell-online-individualist-vs-collectivist-cultures>
- Nahai, N. (2013) How to Sell Online to Indulgent vs Restrained Cultures. *Psychology Today*. <https://www.psychologytoday.com/us/blog/webs-influence/201308/how-sell-online-indulgent-vs-restrained-cultures>
- Nahai, N. (2013) How to Sell Online to Short vs Long-Term Cultures. *Psychology Today*. <https://www.psychologytoday.com/gb/blog/webs-influence/201308/how-sell-online-short-vs-long-term-cultures>
- Nahai, N. (2013), How to Sell Online to Different Cultures: Power Distance. *Psychology Today*. <https://www.psychologytoday.com/gb/blog/webs-influence/201306/how-sell-online-different-cultures-power-distance>
- Nisbett, R.E., and Miyamoto, Y. (2005) The influence of culture: holistic versus analytic perception. Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1109 *Trends in Cognitive sciences* Vol 9 No 10. October 2005.
- Oshlyansky, L. (2007) Cultural Models in HCI: Hofstede, Affordance and Technology Acceptance. PhD Thesis, Swansea University.
- Parsons, T. (1951) *The Social System*. New York: The Free Press, 1951.
- Pixabay (2021). Stunning free images & royalty free stock. Available at: <https://pixabay.com/>
- Reid, L. (2015) The Importance of Hofstede's Dimensions of Culture. <https://sites.psu.edu/global/2015/04/25/the-importance-of-hofstedes-dimensions-of-culture/>
- Reinecke, K., Schenkel, S. and Bernstein, A. (2010). Modeling a User's Culture. *The Handbook of Research in Culturally-Aware Information Technology: Perspectives and Models*.
- Sinha, J.B.P. (2014) *Psycho-Social Analysis of the Indian Mindset*, 2, Springer India 2014 27. DOI: 10.1007/978-81-322-1804-3.
- Slone, D. J. (2002). The influence of mental models and goals on search patterns during web interaction. *Journal of the American Society for Information Science and Technology*, 53(13), pp. 1152–1169.
- Smith, A., French, T., Dunkley, L., Minocha, S., and Chang, Y. (2004). A Process Model for developing Usable cross-cultural websites, *Interacting with Computers: Special Edition – Global human-computer systems cultural determinants of usability*. 16, 1 (Feb 2004), 63-91.
- Trompenaars, F., and Hampden-Turner, C. (1998) *Riding the Waves of Culture. Understanding Cultural Diversity in Business*. (2nd.Ed.) Nicholas Brealey Publishing Limited, 36 John Street, London. WC1N 2AT.
- Voehringer-Kuhnt, T. (2002) *Kulturelle Einflüsse auf die Gestaltung von Mensch-Maschine Systemen*, Munich: GRIN Verlag.
- Wardrobe, W. (2005). *Beyond Hofstede: Cultural applications for communication with Latin American*, Association for Business Communication Annual Convention.

Subjective Assessment of Commercially Common Input and Display Modalities in a Driving Simulator

Kian Motahari^a and Margarita Vinnikov^b

*Department of Informatics, New Jersey Institute of Technology, Newark, U.S.A.
km755@njit.edu, margarita.vinnikov@njit.edu*

Keywords: Virtual Reality, Driving Simulator, Task Load, Usability, Simulator Sickness, Steering.

Abstract: Driving simulators could be valuable tools for better understanding human behavior while driving. The latter has the potential to guide the design of roads and vehicles. In addition, studies of driving habits can help gain insight into the relationship between human attention and locomotion. A critical aspect of a driving simulator involves the right choice of stimulus presentation and interaction methodologies. This paper presents a driving simulator study that evaluates a series of commonly used display and steering devices in terms of usability, physical and cognitive task load, and simulator sickness indices. Specifically, conventional large display, commercially available head-mounted display, and various steering devices were compared. Our study used two steering devices (stationary and wireless) and a gamepad controller to control the simulated vehicle. We hypothesize that using a Head-Mounted Display (HMD), although suitable for inducing immersion, might come with the cost of possible simulator sickness and a decrease in driving performance. We analyzed the individual and combined impact of display and steering input device types on our subjective metrics measurements. Our results proved our hypothesis on higher perceived immersion for HMD based driving simulator. Also, the paper highlights the trade-offs between big monitor setup and Virtual Reality (VR) in terms of workload and fatigue.

1 INTRODUCTION

VR is a class of graphical user interfaces that generates an immersive, spatially realistic, and interactive environment that could be both experienced and interacted with using head-mounted stereo goggles, multi-projected installation, headphones, peripheral suits, and gloves (Brey and Søraker, 2009). The use of VR has been prevalent in both empirical research and industry for various operational tasks such as driving, surgical procedures, and work with machinery. Historically, multi-projected installations such as CAVETM systems were the most commonly used in VR (Cruz-Neira et al., 1992). Such installations require sizeable physical space, a separate tracking system, and integration of input devices. Nowadays, HMD are highly preferred since they still provide binocular stereo vision, with head and hand tracking already incorporated into a head-worn system. The advantage of using VR for training and operating purposes is the ability to simulate and replicate

hazardous and realistic situations in a controlled manner, where the possibility of endangering the users is either diminished or eliminated. For example, driving simulators place the driver in an artificial environment to recreate a real driving experience that is conducted safely and in a controlled manner. The driving simulator could be used not only for training purposes but also to provide insight into driving performance, road design, and assessment of human behavior under different circumstances, such as the influence of substances and severe weather and lighting conditions (Chang, 2016). Therefore, to account for human behavior during driving, we need to understand what display types and interaction modalities are most appropriate when designing driving simulators for an average user. Consequently, this study examines the impact of utilizing VR based displays (HMD and large TV screen) and VR controllers (wireless and stationary steering wheels and conventional game controllers) on subjective metrics related to driving performance. We

^a  <https://orcid.org/0000-0002-5398-5881>

^b  <https://orcid.org/0000-0001-9043-7166>

specifically analyze the effect of these interaction and display methods (combined and individually) on human behavior and self-reported, driving-related metrics.

The rest of the paper is structured as follows: First, we review the literature related to driving in terms of concepts of driving simulators, VR simulations, task-dependent human behavior, evaluation of steering and proposed methods, the fidelity of interaction and stimulus, and bodily/cognitive task load.

2 RELATED WORKS

Driving performance metrics might include the time to interact (move/rotate) with the device or the higher-level measurements like negotiating bends or the ability to adjust the steering after perturbations. For example, Fitts' law is a predictive model proposed to design interfaces to generalize involved factors. It states that the time required to move a pointer to a target area is dependent on the distance to the target divided by the area of the target (Fitts, 1954). Accot et al. (Accot and Zhai, 1999) developed a trajectory-based testing paradigm as a study scenario in parallel with Fitts' positioning task. Their proposed steering law can predict completion time relative to tunnel (path) parameters (such as length and width), which could be used as a tool for assessing trajectory-based tasks. Yet, control of a vehicle involves the ability to steer to negotiate bends at high speeds and take the right trajectory along the curved road (Kountouriotis et al., 2015). Van Leeuwen et al. (van Leeuwen et al., 2015b) investigated how driving performance and gaze behavior change throughout the time a novice driver's driving skills progress. A driving simulator was used to evaluate performance and gaze activity as participants practiced driving. They showed that improvement in the driving performance can be tracked and is correlated with the gaze activity and its changes throughout practicing using the driving simulator. They also showed that horizontal gaze variance could be used to predict an increase in speed. To quantify the trajectory of steering, multiple processes have been proposed (Hedegaard et al., 2019, Sportillo et al., 2017). Hedegaard et al. (Hedegaard et al., 2019) used angular deviation, which is the difference between a steering input signal and an average of all input signals to evaluate the steering performance across different steering modalities. They also quantified the smoothness of steering by using the frequency of significant peaks of the steering input signal. On the other hand, Sportillo et al. (Sportillo et al., 2017) used the number of steering turns as an

indicator of steering stability after regaining control of a simulated vehicle. In our works, we will assess the steering performance to analyze the impact of display and input fidelity on the stability of a vehicle's trajectories while negotiating bends in two urban and rural environments and also to acquire gaze-steering coordination, which is a good indicator of driving performance and level of engagement with the control of a vehicle.

The degree of realism in a simulator is referred to as "physical fidelity" (van Leeuwen et al., 2015a). According to mentioned definition, physical fidelity could be defined for multiple aspects of simulation that include stimuli presentation (display, audio, field of view, texture quality, etc.), input methods, and interaction with the simulated content. To elicit realistic responses from participants while performing tasks like driving, we need a certain level of physical fidelity of the stimuli. However, according to (van Leeuwen et al., 2015a), there are several drawbacks to high fidelity simulation, which include: higher fidelity, the experiment will face a large number of variables to be controlled during the conduction of the experiment; Simulator discomfort, which can cause participants' withdrawal and reduced data quality; and, realism and high fidelity of certain types of visual information might be unnecessary or even distracting to evoke realistic performance from the participants. As a result, we need to address both the impact of display or interaction fidelity solely or combined. Van Leeuwen et al. (van Leeuwen et al., 2015a) in their study, implemented three levels of visual fidelity for a driving task experiment to evaluate the lane-keeping and gaze strategies adopted by participants. They demonstrated that the highest fidelity scenario resulted in higher steering activity, driving speed, and horizontal gaze variance compared to medium and low fidelity scenarios. Also, their driving simulator study addressed the impact of visual fidelity on curve negotiation, gaze behavior, and self-reported discomfort. They used three levels of visual fidelity (high fidelity with texture, mid-range fidelity without textures, and low-fidelity with only lane markers). They showed higher steering activity on the straight road, higher speed, and a higher degree of gaze variance for the high-fidelity scenario compared to the other two scenarios. It is not enough to address the impact of fidelity of only display or interaction method on driving; it is also important to analyze the complex effect of fidelity of display and input modality combined. McMahan et al. (McMahan et al., 2012) addressed the independent and compound effect of display and interaction fidelity on participants' performance in a VR first-person shooter game using a six-sided CAVE

system. They showed that the compound effect of low fidelity and high fidelity for both display and interaction could significantly improve the in-game performances (completion time, damage taken, etc.). This effect could be associated with familiarity with those combinations (low or high fidelity for display and interaction), which resembles the standard desktop FPS games and real-life experiences. In the current study, we examine the compound impact of physical fidelity in terms of field of view, which is presented in the form of HMD based and display-based stimuli presentation (display fidelity), and input methods, presented here in the form of controller-based methods and realistic commercial steering devices (input fidelity) on driving-related performances, and also on perceived levels of immersion, sense of agency, cognitive and physical task load, and simulator discomfort.

An important aspect of a simulation is the mental and physical effort the users need to engage with the environment and interaction modalities. *Cognitive task load* is the total amount of user's mental demands required to use a specific system (Luxton et al., 2016) and *workload*, according to (Hart and Staveland, 1988), is defined as the perceived relationship between the number of available resources and the required amount by a given task. Various surveys can be used to quantify the cognitive and bodily task load. For example, the most commonly used survey is the NASA Task Load Index (NASA TLX), which is a self-reported measure of workload (Hart and Staveland, 1988). It evaluates the task load along six axes: mental demand, physical demand, temporal demand, effort, frustration, and self-performance. Although NASA-TLX was initially designed for pilots of space flights, it has been adapted for other environments and tasks such as surgery and driving activity. For example, Kountouriotis et al. (Kountouriotis et al., 2015) examined the impact of cognitive load on steering performance. Their study evaluated the effect of attention on steering performance in a simulated driving task. They used gaze fixation points (as a contributor to the cognitive load) that their position was either changing relative to the vehicle or was fixed relative to the vehicle. Their result indicated that the gaze fixations that were not analogous to the future path cause steering biases. Furthermore, Harris et al. (Harris et al., 2020) discussed the additional factors that are needed to be incorporated into the current workload measures for VR tasks. These factors include a degree of immersion, perceptual challenge, and methods to control the environment. In this work, we use NASA-TLX, simulation sickness, and a usability questionnaire (that includes items related to

immersion, overall satisfaction and excitement, and fatigue in involved body parts) tailored for our specific VR task to evaluate users' experience of driving in VR.

3 METHOD

The present work aims to evaluate the impact of display and steering input methods on driving-related fidelity for subjective and objective measures. The study also explored the relationship between eye-hand coordination in driving, control of the vehicle's trajectory, and perceived sense of immersion, fatigue, and cognitive and bodily effort. The present study follows a within-subject design with two independent variables (display type and controller type) that formed six blocks. These blocks included all possible combinations of two display options (large display and HMD) and three steering devices (stationary steering device, wireless steering device, and manual controller). The main task included driving through a series of designated paths that involved driving straight and turning right and left. In all three possible instructions, participants were asked to control the steering and velocity of the virtual vehicle through an intersection and then either continue straight, turn right, or turn left. Each trial ends when a participant reaches a designated endpoint. Trials took place in two distinct virtual scenes (urban and rural characteristics) with an equal amount of turns in each scene. Consequently, each block consisted of twelve trials that had two variables of scene type (urban and rural), driving instruction type (straight, left, right), and two repeats. To test the impact of each scenario block, we collected the participants' subjective assessment of simulation discomfort, immersion, and task cognitive and physical workload by using questionnaires after each block. Gaze movements and steering activities were recorded during driving tasks, and we discuss them in a separate report. We hypothesized that: A higher degree of immersion will be experienced in the HMD scenarios, regardless of the input method. And less physical discomfort and simulation sickness would be experienced using the steering wheel (stationary and wireless) scenarios.

3.1 Participants

The present study includes 18 participants (5 female and 13 male), all over 18 years old ($M = 20.6 \pm 3.5$). With all of them having previous driving experience and 8 participants who had prior experience with VR. Written informed consent was obtained from all par-

participants following a protocol approved by the Institutional Review Board at NJIT. Participants were asked to report any proneness to cybersickness or prior experience of dizziness with VR. One participant was excluded from the experiment and performing the tasks.

3.2 Stimuli

This experiment has used costume-made software developed with Unity 3D software (Haas, 2014). At the start, participants' ID and input device, and display type are set, and before the onset of the main scene, the calibration scene is presented. We also had a neutral scene for training, which we let participants try before the start of each to familiarize them with the control of the vehicle using the block's input device. If participants had already trained with a given input device in a subsequent block, we skipped the training session (a total of four times of running the training scene). The calibration scene was the standard calibration process provided by the Pupil Labs eye tracker add-on for unity. The driving tasks of the experiment took place in two scene types - urban and rural scenes. The urban scene included intersections in a city model (Manufacture), and the rural scene features a rural area with a roundabout in the center of the virtual scene. Both scenes included active pedestrians, but only in the urban scene, participants could cross the intersection. In the HMD-based blocks, instructions were shown on the speedometer and car dash- board monitor. Instructions were presented on the speedometer for the monitor-based blocks due to the fixed camera position. Also, both scenes included an intersection (a roundabout for the rural scene) where participants were instructed to perform the driving task (right and left turns and driving straight) on them. Both scenes include pedestrians, but only in the urban scene do they cross the road, depending on the traffic light.

The platform used for the experiment was a PC (six cores Core i7 3.7 GHz, Nvidia GTX 1070). Two steering wheel devices were Logitech G920 (for PC) (Logitech, 2015) as the stationary steering wheel setup, and Hyperkin S Wheel Wireless Racing Controller (Hyperkin, 2019) as the wireless steering wheel setup. Both devices feature hand pedals that we used to control of the vehicle. For the manual control as the input method, we used 6-DoF HTC VIVE controllers (Corporation, 2018) and Xbox one controller (Microsoft) respectively paired with HMD and large monitor scenarios. A 42 inches 4k TV, and HTC VIVE Pro VR headset (Corporation, 2018) were used for the large display scenarios, and the HMD-

based scenarios, respectively.

3.3 Procedure

Upon arrival, each participant was given a consent form and a short study description. Participants were strongly encouraged to read the provided documents, ask questions, and express any concerns before signing the consent document. This was followed by a short demographic questionnaire and a simulation sickness questionnaire. Then participants were exposed to six experimental blocks in random order. Before the start of each block, participants were provided with a short trial scene to familiarize themselves with the input/display modalities of the block. After participants were comfortable with steering with the provided controller for the trial, an experimental block would start. At the beginning of each block, users' gaze was calibrated based on Pupil Labs' mobile eye-tracker and VR add-on. The calibration was a non-invasive process that required users to stare at five fixation targets for a predefined amount of time. After all points were collected, an experimenter was presented with an accuracy index. If the index was below 0.9, the calibration procedure was repeated. If the calibration procedure had failed after three trials, the experiment for that participant was terminated. During the block, the participant was presented with 12 trials. Participants were instructed to finish a driving task for each trial within the block. We terminated the trial if participants crashed the vehicle to the sides of the road, or in the case of the urban scene, also crashing with the virtual pedestrians caused the trial termination. After finishing each block, participants were provided with usability, simulator sickness, and NASA task load questionnaires (TLX). Each experiment was followed by a debriefing session.

3.4 Data Analysis

NASA Task Load Index (NASA-TLX) included six questions on a designated seven-point Likert scale. R programming language was used to import the cumulative data for statistical analysis. We used raw scores from each section to determine the significance of differences between the six experimental blocks. The usability questionnaire (table 1) used in this study included 30 questions that cover physical comfort, aspects of game-play like accuracy and control of the task, immersion, sense of presence, and sense of control over the task. All answers were available seven-point Likert scale responses. Scores were ordinal data without the assumption of normality. Simulator Sick-

ness Questionnaire (SSQ) used in this study included the assessment of 15 symptoms that range from general discomfort to specific symptoms like nausea and sweating. Each symptom had a four-level response that included none, slight, moderate, and severe. For the analysis of subjective data scores, with regards to block specifications, we used R and (Kuznetsova et al., 2017) package to perform a mixed model analysis of the relationship between task load, simulator sickness, and usability scores, separately, and input and display modalities of the driving block. As fixed effects, input and display types with interaction terms in the model were used. As random effects, we had intercepts for participants. *P*-values were obtained by likelihood ratio tests of the full model with effects in question against the model without the effect in question.

4 RESULTS

The usability questionnaire questions were split into three sections: questions related to the overall satisfaction of the experiment, questions related to the sense of immersion and presence, and those associated with physical comfort. This discrimination is due to the contextual similarities of the questions and similar relationships with the display and input scenarios. Results of the mixed model analyses for the usability questionnaire items revealed that the choice of input device significantly impacted 19 out of 30 questions, and 18 out of 30 questions were predicted by the display type (for concrete results, see table 1). However, no significant effect of interaction between display type and input device type on any of the usability questions has been observed. Among five usability questions related to the overall satisfaction of the experience and enjoyment (Q18 to Q22 table 1), satisfaction or enjoyment of the experience, sense of accomplishment, and suitability of task to gameplay increased by the two steering devices (stationary and wireless). However, excitement with experience increased by HMD and steering devices. The score of all questions related to the sense of immersion, presence, and engagement with the virtual environment (Q1 to Q12 from table 1) increased for the scenarios with HMD display and steering wheels in comparison to a large monitor and gamepad controller. The only exception is the perceived picture quality (Q12, table 1), which showed the highest scores for conditions with monitor display. For the fatigue-related and physical comfort items (Q13 to Q17 of table 1), fatigue in the wrist and fatigue in neck decreased for the monitor display blocks, fatigue in fingers decreased

for the steering devices, and overall physical comfort enhanced for the monitor-based scenarios, and comfort with the viewing position increased for the steering devices. For the NASA task load scores, the interaction effect of display and input impacted the physical score ($\chi^2(2) = 6.31, p = 0.04$), the interaction of monitor display and stationary wheel input increasing it by 1.16 ± 0.47 (Likert scale score), and monitor display and wireless wheel input increasing it by 0.82 ± 0.47 (figure 1b). Also, the interaction effect of input and display types affected performance score ($\chi^2(2) = 11.11, p = 0.004$), monitor display and stationary wheel lowering it by 1.69 ± 0.55 . Lastly, input type impacted the frustration level ($\chi^2(2) = 8.19, p = 0.01$), stationary wheel input lowering it by 1.07 ± 0.39 (figure 1a). From simulator sickness related scores, we observed a significant effect of input device type on the general discomfort level ($\chi^2(2) = 6.53, p = 0.038$). Level of the fatigue was impacted by the input device type ($\chi^2(2) = 7.02, p = 0.03$). Self reported headache induced by the trial blocks was significantly affected by the type of display ($\chi^2(1) = 6.26, p = 0.01$), and also by the input device type ($\chi^2(2) = 6.98, p = 0.03$), but no interaction effect was observed.

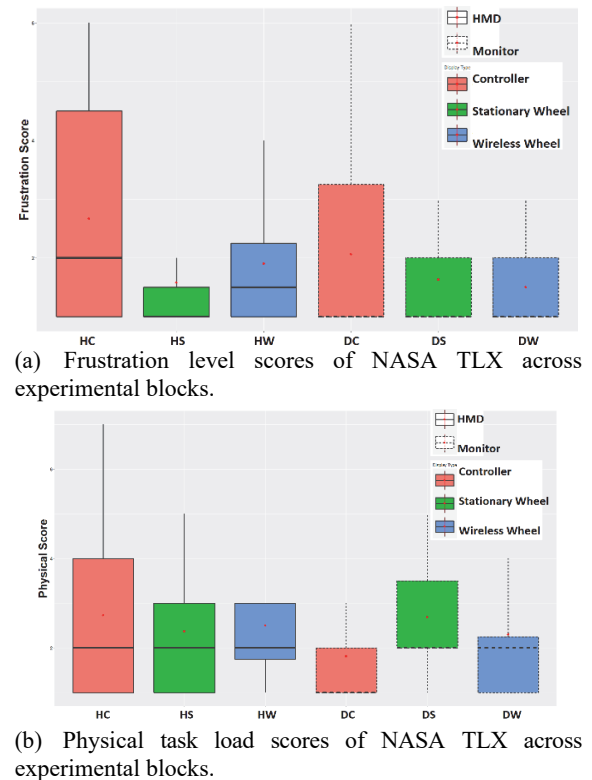


Figure 1: Compound effect of the input device and display type on frustration level and physical load of NASA TLX questionnaire based on experimental blocks.

Table 1: Results of the mixed model analysis of the effect of the input and display modalities on the usability items. Impact of display and/or input modality on each score is indicated according to the likelihood ratio (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

	Usability Question	Display			Input		
		p -value	χ^2	Df	p -value	χ^2	Df
Q1	Rate your sense of immersion	***	20.56	1	***	24.37	2
Q2	Rate your sense of being in the virtual environment	***	17.42	1	**	10.73	2
Q3	How much did it seem as if you could reach out and touch the objects or people you saw	***	17.42	1	**	10.73	2
Q4	How often when an object seemed to be headed toward you did you want to move to get out of its way	***	17.42	1	**	10.73	2
Q5	To what extent did you experience a sense of 'being there' inside the environment you saw	***	16.90	1	*	6.28	2
Q6	How often did you want to try to touch something you saw	***	16.9	1	*	6.28	2
Q7	How involved was the experience	***	14.28	1	***	16.93	2
Q8	How completely were your senses engaged	***	18.44	1	***	14.03	2
Q9	To what extent did you experience a sensation of reality	***	20.95	1	**	10.62	2
Q10	How engaging was the experience	***	17.81	1	***	22.27	2
Q11	Overall, how much did touching the things and people in the environment you saw/heard feel like it would if you had experienced them directly	***	17.81	1	***	22.27	2
Q12	How was the picture quality during the experience	***	13.30	1	2.93	0.23	2
Q13	Rate your physical comfort level after your experience	*	5.07	1	0.08	4.91	2
Q14	Rate your soreness or fatigue of the wrist	**	9.28	1	0.87	0.27	2
Q15	Rate your soreness or fatigue in fingers	0.06	3.28	1	***	20.65	2
Q16	Rate your soreness or fatigue in neck or shoulder	*	4.77	1	0.33	2.20	2
Q17	How comfortable were you with your viewing position	0.20	1.62	1	*	7.66	2
Q18	How exciting was the experience	*	5.16	1	*	8.39	2
Q19	Overall, how satisfying or enjoyable was the experience you just had	0.97	0.0008	1	***	23.8	2
Q20	Rate suitability of the task to gameplay	0.85	0.03	1	***	24.25	2
Q21	Rate your sense of enjoyment	0.12	2.40	1	***	36.53	2
Q22	Rate your sense of accomplishment	0.41	0.65	1	***	16.68	2
Q23	Rate your soreness or fatigue in arms	0.30	1.07	1	0.81	0.40	2
Q24	Rate your precision/accuracy	0.81	0.05	1	0.12	4.22	2
Q25	Rate your control of the task	0.79	0.07	1	0.31	2.31	2
Q26	Rate your sense of challenge	0.75	0.10	1	0.09	4.70	2
Q27	Overall, how well do you feel you have been driving	0.38	0.76	1	0.17	3.46	2
Q28	Overall, how easily do you feel you have been driving	0.88	0.02	1	0.37	1.93	2

Eye strain was predicted by the display type ($\chi^2(1) = 13.80$, $p = 0.0002$). A significant impact of the input device type on the sense of nausea was observed ($\chi^2(2) = 9.07$, $p = 0.01$). For the sense of dizziness, we saw display type significantly impacting it ($\chi^2(1) = 4.26$, $p = 0.03$). Finally, We observed that the sense of "fullness of head" was impacted by the display modality ($\chi^2(1) = 4.66$, $p = 0.03$).

5 DISCUSSION

This article explored how different combinations of

steering input devices and display types for our driving simulator could impact participants' assessment of task load, simulator sickness, and usability. It was observed that both steering input devices lowered the general discomfort from SSQ, while the gamepad controller increased it. This observation agrees with the higher level of frustration from the usability data for the conventional controller. Wireless and stationary steering wheel devices also caused lower fatigue levels (form SSQ). The observations from the result section provide support for the claim that HMD and the realistic steering devices improve the sense of immersion and engagement with the driving experience.

Also, realistic steering devices helped participants to have a more satisfying driving experience and report less fatigue in their fingers. However, the monitor display was superior to the HMD in terms of reporting less fatigue in the neck and wrist and higher perceived picture quality and physical comfort. Relatively lower reported physical comfort and higher fatigue in the neck for the HMD scenarios were expected due to the additional weight of the headset. However, higher fatigue in the wrist for the HMD scenarios could be because, throughout the experiment, we observed that participants tend to match the physical position of the wireless steering device with its in-game position when using the HMD (that could cause more physical load on the wrists), while for the monitor-based scenarios they tend to hold the wireless steering device in a more comfortable position (usually on their feet) rather than trying to match it with the on-display position. Also, a higher immersion and engagement with an environment were reported for the HMD scenarios that made participants mimic and reproduce the real-life driving experience. Also, since the output resolution of the display and HMD were the same, higher perceived picture quality for the monitor scenarios could be related to the results from the simulator sickness data that participants reported a lower level of eye strain for the monitor scenarios, and therefore it impacted their perception of picture quality. In addition, according to (Vinnikov and Allison, 2014), which explored the impact of depth of field (DOF) for stereoscopic and non-stereoscopic displays on perceived image quality and viewing comfort, DOF rendering in HMD scenarios could be the reason for lower perceived image quality over monitor-based scenarios. Due to the no interaction effect of input and display modalities on usability questions, display, and input modalities aren't interdependent. Results indicated a strong interaction effect of input and display types on physical task load and self-evaluation of performance from the NASA task load questionnaire. Also, from the NASA task load, it was observed that the realistic input devices (i.e., wireless and stationary wheels) lowered the level of reported frustration compared to the gamepad controllers. Although the stationary steering wheel caused the lowest amount of frustration regardless of the display type, it decreased the self-evaluated performance when paired with the monitor display. This observation implies that the realistic input device results in the best performance assessment paired with a more immersive display method. In line with these results, (McMahan et al., 2012) also showed that the compound effect of input and display fidelity, when both are high or low fidelity, im-

proves the performance in their game scenario discussed in the related works section. Steering input devices and large display scenarios lowered the self-reported ratings for the headache symptom relative to other conditions. Higher levels of eye strain (from SSQ) have been recorded for the HMD in comparison to the monitor display scenarios. Interestingly, regarding the nausea symptom (from SSQ), wireless stationary steering input devices recorded lower levels, regardless of the display setup. This is significant because no interaction effect has been shown on nausea. Also, HMD-based scenarios indicated a higher degree of reported dizziness than monitor-based ones. Finally, from the SSQ, similar to the dizziness symptom, a higher degree of fullness of the head has been shown for the monitor display scenarios. These results indicate that when a realistic controlling scheme is paired with an immersive display type, symptoms associated with simulator sickness tend to decrease. However, although in general monitor-based scenarios caused fewer simulator sickness symptoms, when paired with realistic input devices caused more symptoms in comparison to when paired with a controller. Also, according to the results, lower discomfort for the monitor display scenarios from the SSQ data confirms the observation from the usability questionnaire. In addition, the usage of realistic input devices (stationary and wireless steering devices) lowered the discomfort, which could be because the control scheme for those devices was intuitive to the driving task (higher suitability of task to gameplay for the steering wheels from the usability results) in comparison to the scenarios with a gamepad controller. Despite these observations, results from NASA TLX questionnaire indicated that a stationary steering device, when paired with the monitor display, causes a higher level of physical task load. This observation suggests that although the more intuitive control scheme causes lower physical discomfort when paired with immersive stimulus presentation, it also mitigates the physical task load overhead of realistic interactions and force feedback.

6 CONCLUSION

This study focused on evaluating different types of steering devices and display methods in a driving simulator in relation to participants' subjective assessment of the driving experience. Six stimulus presentation and interaction blocks were designed to comprise all possible pairs of two steering devices and controllers and two display methods. The results indicate that in terms of immersion, sense of presence,

sense of agency, comfort, and achievement, two steering wheels (wireless and stationary), especially when paired with VR headset, are superior to conventional VR and non-VR controllers. Further analysis of gaze-steering correlation, distribution of fixations across the Area of Interest (AOI)s, and the relationship between subjective assessment of task load and calculated steering effort, based on made trajectories, is in perspective. One shortcoming of the current works was the lack of the ability of natural steering (actual rotation) for the conventional controllers. Also, due to the increasing number of variables, we did not account for participants' positional and rotational information in the blocks with large displays. Finally, this work had a limited number of road types and driving scenarios. Nonetheless, more studies are needed to address the exact variables associated with steering devices and displays related to observed impacts on subjective assessments. Possible improvements for controller-based scenarios should be examined. Due to the variability of optical flow in different scenes, when driving, further investigation into the relationship between changes in optical flow and its impact on gaze-steering correlation is essential. Then, we can address how changes in the visual information (in urban vs. rural scenes) that alter perceived motion and are associated with simultaneous steering actions could constitute the levels of bodily discomfort and task load (physical, cognitive, temporal).

REFERENCES

- Accot, J. and Zhai, S. (1999). Performance evaluation of input devices in trajectory-based tasks: an application of the steering law. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 466–472.
- Brey, P. and Søraker, J. H. (2009). Philosophy of computing and information technology. In *Philosophy of technology and engineering sciences*, pages 1341–1407. Elsevier.
- Chang, K.-H. (2016). *e-Design: computer-aided engineering design*. Academic Press.
- Corporation, H. (2018). *VIVE Pro HMD User guide*.
- Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. (1992). The cave: audio visual experience automatic virtual environment. *Communications of the ACM*, 35(6):64–73.
- Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381.
- Haas, J. K. (2014). A history of the unity game engine.
- Harris, D., Wilson, M., and Vine, S. (2020). Development and validation of a simulation workload measure: the simulation task load index (sim-tlx). *Virtual Reality*, 24(4):557–566.
- Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.
- Hedegaard, M., Støttrup, N., Sørensen, F. F., Langer, T. H., and Samani, A. (2019). Evaluation of five steering input devices in terms of muscle activity, upper body kinematics and steering performance during heavy machine simulator driving. *International Journal of Industrial Ergonomics*, 72:137–145.
- Hyperkin (2019). *S Wheel Wireless Racing Controller User Manual*.
- Kountouriotis, G. K., Wilkie, R. M., Gardner, P. H., and Merat, N. (2015). Looking and thinking when driving: The impact of gaze and cognitive load on steering. *Transportation research part F: traffic psychology and behaviour*, 34:108–121.
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmer test package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.
- Logitech (2015). *Logitech G920 Driving Force Racing Wheel User Guide*.
- Luxton, D. D., June, J. D., Sano, A., and Bickmore, T. (2016). Intelligent mobile, wearable, and ambient technologies for behavioral health care. In *Artificial intelligence in behavioral and mental health care*, pages 137–162. Elsevier.
- Manufacture, N. Windridge city Unity Asset.
- McMahan, R. P., Bowman, D. A., Zielinski, D. J., and Brady, R. B. (2012). Evaluating display fidelity and interaction fidelity in a virtual reality game. *IEEE transactions on visualization and computer graphics*, 18(4):626–633.
- Microsoft. *Xbox One Wireless Controller*.
- Sportillo, D., Paljic, A., Boukhris, M., Fuchs, P., Ojeda, L., and Roussarie, V. (2017). An immersive virtual reality system for semi-autonomous driving simulation: a comparison between realistic and 6-dof controller-based interaction. In *Proceedings of the 9th International Conference on Computer and Automation Engineering*, pages 6–10.
- van Leeuwen, P. M., Gómez i Subils, C., Ramon Jimenez, A., Happee, R., and de Winter, J. C. (2015a). Effects of visual fidelity on curve negotiation, gaze behaviour and simulator discomfort. *Ergonomics*, 58(8):1347–1364.
- van Leeuwen, P. M., Happee, R., and de Winter, J. C. (2015b). Changes of driving performance and gaze behavior of novice drivers during a 30-min simulator-based training. *Procedia Manufacturing*, 3:3325–3332.
- Vinnikov, M. and Allison, R. S. (2014). Gaze-contingent depth of field in realistic scenes: The user experience. In *Proceedings of the symposium on eye tracking research and applications*, pages 119–126.

Towards Evaluating e-Commerce Sites Purchase Intention using Affective Computing, a Preliminary Study

Hayri Yigit¹, Javier De Andrés²^a, Martin González-Rodríguez³^b, Daniel Fernández-Lanvin³^c,
Kathia Marçal de Oliveira⁴ and Emmanuelle Grislin-Le Strugeon^{4,5}

¹Department of Computer Science, Mehmet Akif Ersoy University, Antalya Burdur Yolu, 15030 Yakaköy/Burdu, Turkey

²Department of Accounting, University of Oviedo, c/San Francisco 3, 33001, Oviedo, Spain

³Department of Computer Science, University of Oviedo, c/San Francisco 3, 33001, Oviedo, Spain

⁴Univ. Polytechnique Hauts-de-France, LAMIH, CNRS, UMR 8201, F-59313 Valenciennes, France

⁵INSA Hauts-de-France, F-59313 Valenciennes, France

hayriyigit@protonmail.com, {martin, jdandres, dflanvin}@uniovi.es, {kathia.oliveira, emmanuelle.grislin}@uphf.fr

Keywords: Usability, Affective Computing, e-Commerce.

Abstract: The evaluation of an e-commerce website usability and effectiveness is traditionally measured through surveys. These tools provide subjective information about the user perception of the website, and their use is expensive and involve some drawbacks, since it requires the user to collaborate in the study and to control the sample. In this work, we explore the relation between what the users perceives and shares when they are asked about their experience, and the emotion they show while using the system under study. We use Affective Computing algorithms to analyse user expressions, and we compare the results with the information provided by means of a TAM based survey.

1 INTRODUCTION

The success of any e-commerce website is strongly determined by the usability of its user interface, and thus, by the satisfaction of the user that access the site (Sahi, 2015). The most popular and traditional way to measure the satisfaction of the user is applying Technology Acceptance Models (TAMs) (Marangunić & Granić, 2015), that is, using surveys that the subject must fulfil after using the product. Although this method is well tested and accepted, it has many drawbacks that make it difficult and expensive to use, especially in frequently updated web sites, since it requires the subject collaboration and a controlled sample. Moreover, the results of the evaluation are based on the subjective perception from the user, who can be influenced by many factors (familiarity, learning effect, sympathy for the website, reluctance to changes, etc.) that could bias the results, and favour a website over others with better design and usability.

In order to remove the subjective component, there are other strategies like, for example, the use of eye-tracking technologies, which measure the performance of the user with objective parameters (Cooke, 2004; Lasa et al., 2015; Rakoczi & Pohl, 2012). However, they usually require onsite evaluation and technological infrastructure.

On the other hand, one of the emerging and most active lines of research in Human Computer Interaction is the one related to the perception of user emotions through artificial vision algorithms. This is part of what we know as Affective Computing (AC) (Ma & Yarosh, 2021). AC algorithms can process the picture of a subject and classify their emotions analysing their facial expression. The relation between emotions, usability and purchase intentions has been already demonstrated by previous research works (Eroglu et al., 2001).

In this paper, we try to explore whether the user emotion captured by AC algorithms while using an e-commerce site is consistent with the user subjective perception of usability according to responses to a

^a <https://orcid.org/0000-0001-6887-4087>

^b <https://orcid.org/0000-0002-9695-3919>

^c <https://orcid.org/0000-0002-5666-9809>

TAM based survey. We designed a prototype that simulates the purchase process in an e-commerce website, gathering pictures of the subject and evaluating them with three different AC algorithms, and at the same time, we asked users to fulfil the survey at the end of the process. The results we obtained, despite the reduced size of the sample, point out that there is a significant correlation between the output of the AC algorithms and some questions of the survey.

2 RELATED WORK

Results from earlier studies demonstrate a strong and consistent association between user emotions and purchase behaviour. The most important factor affecting user emotion on an e-commerce website is undoubtedly the user interface and user experience (Eroglu et al., 2001). These authors developed an exemplary e-commerce site in their study and conducted a survey with 328 people who tried the site. They concluded that the environment of the online shopping site will affect the emotion and attitude of the user, which in turn will affect the shopping results (Eroglu et al., 2001). Makkonen et al. emphasized that, in cases of successful sales, users are generally in a positive mood. Moreover, they found that positive emotions are more effective than negative ones in situations such as satisfaction, repurchase and recommendation (Makkonen et al., 2020). In addition, Kemppainen et al. found that the user interface is a critical factor that can cause negative emotions (Kemppainen et al., 2020).

Numerous studies have been conducted to evaluate interface performance based on user feelings. Although the selected methods are different from each other, the main idea is to detect the emotions of the user with the help of sensors while trying a product, and then to direct a survey to the users and analyse the data obtained to measure the performance of the user interface.

2.1 Emotion Detection

Several different dimensions can be used to represent emotions. For example, the *Pleasure, Arousal and Dominance* (PAD) emotion model developed by Mehrabian et al. uses 3 dimensions to detect all emotions (Mehrabian & Russell, 1974). In contrast, Fontaine and Plutchick examined emotions using more than two dimensions (Fontaine et al., 2007; Plutchick, 1980). Bhandari et al. used arousal and valence dimensions to analyse user emotions in their

study examining the effects of design factors (Bhandari et al., 2017). Johanssen et al., on the other hand, developed a framework to define usage issues in mobile applications based on user sentiment, and tended to divide sentiment into 7 different dimensions (Johanssen et al., 2019). In their study, Eroglu et al. experienced that the PAD scheme could not capture most of the effects of the site atmosphere and suggested the use of a more inclusive scheme in nature that includes reactions such as interest, anger, surprise and disgust (Eroglu et al., 2003). In our study, similar to such suggestion, we examined emotions in 7 different dimensions.

Furthermore, many methods such as Galvanic Skin Response (GSR), Electrodermal Activity (EDA), Electroencephalogram (EEG), Electromyography (EMG) and Facial Expression Recognition (FER) have been used to detect emotions in research. Gasah et al. used an EEG device to detect emotions in their study, which was aimed to determine whether a user interface would create positive emotions in students (Gasah et al., 2019). Bhandari et al. used more than one method: Facial-EMG for valence and EDA for arousal in their study (Bhandari et al., 2017). Lasa et al. developed a UX evaluation tool using FER and an eye tracker (Lasa et al., 2015). Psychophysiological sensors (GSR/EDA, EEG, EMG etc.) can give the user's emotional state in real time, but there are some disadvantages associated with the use of such sensors. Firstly, measurements made with them can be expensive. Moreover, the use of these sensors is so intrusive that can disrupt human nature by distracting the user and hence change the results of the experiment (Maia & Furtado, 2016). In our study, we used the facial expression detection method to detect emotions. No physical contact was made with the participants, the user carried out the experiment in his own personal environment, and only the user's own computer and camera were used to collect data.

2.2 Measuring Satisfaction with Surveys

The Technology Acceptance Model (TAM) pretends to understand how users accept and use a technology (Marangunić & Granić, 2015). TAM has been used in many studies covering the subject of Human Computer Interaction (Tang & Hsiao, 2016). It was developed by Fred Davis in his study, "What causes people to accept or reject information technology?" (Davis, 1989). In response to the question, he proposed the following bivariate approach.

Perceived Usefulness: The degree to which a user believes that a new technology will improve business performance.

Perceived Ease of Use: The degree to which users believe that they will not have difficulty using a new technology.

We used a questionnaire survey method based on the Technology Acceptance Model in our work. Questionnaires are useful tools for collecting subjective data containing users' feelings and thoughts, and can be used to support the results of the experiment. To the best of our knowledge, there is not any *de facto* standard survey model used in studies that perform facial expression analysis for emotion detection. For example, in the research driven by McDuff et al., two questionnaires were administered, one of which was conducted prior to the experiment (McDuff et al., 2015). The prior experiment questions were mostly about whether the user liked the ad or bought the product featured in the ad. The questionnaire used by Staiano et al. consisted of 3 different sections: (i) UX evaluation; (ii) user's media-player usage information and (iii) demographic information. In the user experience evaluation section, they examined 6 different dimensions of web site design (Staiano et al., 2012). Faisal et al. developed a 26-question questionnaire, in which they assessed web design features (typography, colour, content quality, interactivity, and navigation) to determine trust, satisfaction, and loyalty (Faisal et al., 2017). In this study we used a questionnaire based on this last paper. The final set of items, which were measured using a 7-point Likert scale, was the following:

1. It is easy to read the text on this website with the used font type and size.
2. The font color is appealing on this website.
3. The text alignment and spacing on this website make the text easy to read.
4. The color scheme of this website is appealing.
5. The use of color or graphics enhances navigation.
6. The information content helps in buying decisions by comparing the information about products or services.
7. The information content provided by this website meets my needs.

8. Contents and information support for reading and learning about buying process.

9. This website provides adequate feedback to assess my progression when I perform a task.

10. This website offers customization.

11. This website offers versatility of ordering process.

12. This website provides content tailored to the individual.

13. In this website everything is consistent.

14. Navigation aids serve as a logical road map for buying.

15. Obviousness of buying button and links in this website.

16. It is easy to personalize or to narrow buying process.

17. It is easy to learn to use the website.

18. This website supports reversibility of action.

19. Overall I am satisfied with the interface of this website.

20. My current experience with this website is satisfactory.

21. Overall. I am satisfied with the amount of time it took to complete the tasks for buying products.

22. Overall. I am satisfied with accuracy for this website related to the buying process.

23. I trust the information presented on this website.

24. This website is credible for me.

25. I would visit this website again.

26. I would recommend this website to my friend.

3 OUR APPROACH

Our purpose was to check whether it is possible to measure the satisfaction of the user with a website through the analysis of the expressions of their face, processed by means of affective computing algorithms. We developed a shopping site simulation to collect data. To know their subjective perception of the website, participants were asked to complete two tasks, browsing the website and completing the questionnaire about their experience using the website. Figure 1 shows the method pipeline.

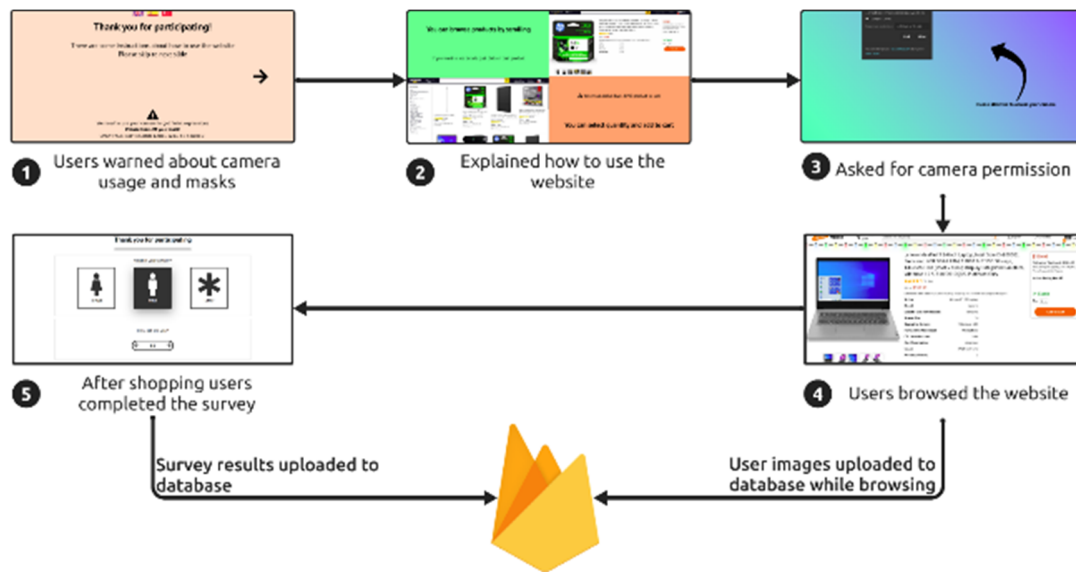


Figure 1: Method pipeline.

3.1 Data Collection

To develop the website user interface, we used the popular JavaScript library ReactJS⁴. We adapted the design from a popular shopping website and created a different look and feel by changing some features in the design, such as the logo, colours and layout, so as not to bias participants. Before starting the simulation, we included a start page that explains how the participant must use the website, asks them to remove their masks (when this is relevant) while using the website, and for the necessary permissions from the participant so that we can use the camera and take snapshots.

After obtaining the necessary permissions, the participant is redirected to the web site and the purchase simulation starts. To distinguish the participants, an artificial unique id is assigned to each of them. The participant is asked to do some simple tasks: browse the products and add at least 1 product to the shopping cart to complete the simulation. A photo of the subject is taken every 2 seconds while they are browsing the website.

Additionally, photos are taken when the subject adds a product to the shopping cart and finish the shopping. The captured photos are saved in the database in Base64 format in real-time. Moreover, the information of the page viewed by the user at the time the photo is taken, and the time information are also recorded in the database. Due to possible errors, simulation sessions are limited to 10 minutes in order

to avoid taking too many photos of the participants and therefore to fill the database unnecessarily.

At the end of the simulation, the questionnaire is shown to the subjects. The questionnaire is presented in 3 different languages, English, Spanish and Turkish. This questionnaire is designed based on the one used by Faisal et al. in a similar study (Faisal et al., 2017), which was designed based on well tested previous relevant TAM studies.

In addition to these, Google Analytics web analytics service, was used to analyse website traffic and audience. The unique ids we assigned to the participants enabled us to track each subject separately.

The Google Cloud Firestore service is used to store collected data. The website is hosted on Vercel, a free platform for hosting static websites.

We kept the website open for 3 months. At the end of the experiment, we wrote a simple JavaScript script to retrieve data from the database. Survey data was stored in CSV format. Subjects' photos were converted from Base64 format to PNG image format using a Python script and stored temporarily for use in the analysis.

3.2 Participants

At the end of the data gathering process, a total of 119 participants visited the site. Although we offer the survey in 3 different languages (English, Spanish and Turkish), we made the site accessible to all countries. Most of the participants were from Pakistan and

⁴ <https://es.reactjs.org/>

Spain. We did not try to balance factors such as the gender of the participant or the platform they used so that the shopping process would proceed naturally. Figure 2 shows the distribution of participants by gender, age, and the platform they use.

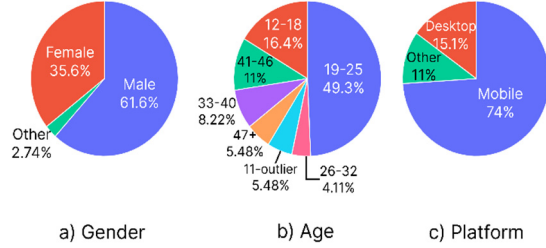


Figure 2: Distribution of participants by gender, age, and the platform they use.

Not all respondents who visited the website completed the survey or were willing to allow their photos and responses to be recorded. In these cases, participants' data were not used in the analysis. 61% of the participants who visited the web site allowed their photos to be saved and completed the survey. In addition, not all the participants had a properly working webcam or they did not read the warnings carefully and take off their masks before starting the simulation. After examining the collected photographs and cleaning the data, 49.3% of the participants who completed the survey (30.25% of all participants) provided valid data to be used in the analysis.

In summary, data from 36 participants, with an average of 57 photos, were used for the analysis.

3.3 Detecting Facial Expressions

To predict facial expressions, we first detected the faces of the participants from the captured photos and then saved these facial photos in the subjects' folders. For the sentiment prediction, we used 3 pre-trained deep learning models that were previously trained on the facial expression dataset. We mentioned this process in the Emotion Detection section. Figure 3 shows the facial expression recognition pipeline.

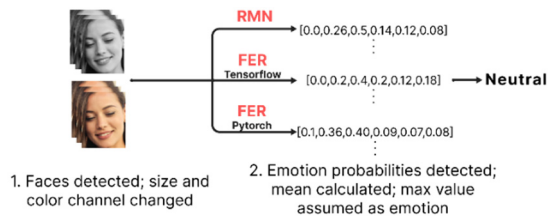


Figure 3: The facial expression recognition pipeline.

3.3.1 Face Detection

MediaPipe is a framework that provides cross-platform machine learning solutions. We used the MediaPipe framework's face detection tool to detect faces. We set the MODEL_SELECTION parameter to 0.0 value, which works better at finding faces up to 2 meters, and parameter 1, better at finding faces up to 5 meters. For the MIN_DETECTION_CONFIDENCE parameter we set 0.5, which means that face detection is considered successful if the value is greater than 0.5.

In some photos, more than one face was detected, in which case we saved the detected faces in a separate folder and then manually filtered the faces that did not belong to the participant. We manually reviewed the obtained data again and removed any invisible, blurred or duplicate faces.

3.3.2 Facial Expression Detection

We used 3 pre-trained models, trained with ready-made datasets, to detect facial expressions.

RMN: It was developed using Pytorch, trained using the FER2013 and VEMO datasets, and achieved 76.82% and 65.94% accuracy on these datasets, respectively (Pham et al., 2020).

FER-Tensorflow: It was developed using TensorFlow, trained using the FER2013 dataset and achieved 65% accuracy (*GitHub - Xionghc/Facial-Expression-Recognition: Facial-Expression-Recognition in TensorFlow. Detecting Faces in Video and Recognize the Expression(Emotion).*, n.d.).

FER-Pytorch: It was developed using Pytorch, trained using the FER2013 and CK+ datasets and achieved 73.12% and 94.64% accuracy on these datasets, respectively (*GitHub - WuJie1010/Facial-Expression-Recognition.Pytorch: A CNN Based Pytorch Implementation on Facial Expression Recognition (FER2013 and CK+), Achieving 73.112% (State-of-the-Art) in FER2013 and 94.64% in CK+ Dataset.*, n.d.).

Each model predicts 7 emotions, namely happy, sad, neutral, angry, fearful, disgusted, surprised. To get the final result from 3 different predictions, we averaged each emotion and assumed the emotion with the highest probability to be the overall emotion of the participants while shopping. We combined the results with the survey results and saved them in CSV format.

3.4 Data Analysis

Once the data had been gathered, we assessed whether certain facial expressions were statistically associated with certain responses to the questions in the questionnaire. To do this we discarded all the facial expressions except neutral and sad, as they did not have enough observations to conduct statistical tests. For the case of neutral there are 16 cases in the sample and for the sad expression 15.

As we have only two valid expressions and the total sample is of small size, we used the non-parametric Mann-Whitney test (Wilcoxon rank sum test) instead of the usual t-tests to assess whether the responses of the participants with sad expression were significantly different from those of individuals with neutral expression. Also, as the prior theoretical developments do not suggest a direction in the relationship, we used two-tailed tests.

4 RESULTS

Table 1 shows the results of the Mann-Whitney tests. The results of each test (one for each question) are displayed in a row in the table. We include the mean rank for neutral and sad expressions, and the p-value associated to the z statistic associated to the Mann-Whitney's U in the test. Tests evidencing significant differences ($p < 5\%$) are displayed in bold typeface while tests suggesting slight differences ($5\% < p < 10\%$) are displayed in italics.

Table 1: Results of the Mann-Whitney tests.

Question	Mean rank for neutral (n=16)	Mean rank for sad (n=15)	p-value
1	10.44	21.93	<0.001
2	<i>13.00</i>	<i>19.20</i>	<i>0.054</i>
3	12.19	20.07	0.014
4	13.56	18.60	0.115
5	11.06	21.27	0.002
6	13.56	18.60	0.117
7	14.34	17.77	0.284
8	14.50	17.60	0.337
9	17.22	14.70	0.438
10	17.47	14.43	0.348
11	15.56	16.47	0.779
12	16.69	15.27	0.660
13	16.28	15.70	0.856
14	14.50	17.60	0.330
<i>15</i>	<i>13.25</i>	<i>18.93</i>	<i>0.078</i>
16	16.94	15.00	0.549
17	14.66	17.43	0.386

18	15.97	16.03	0.984
19	16.31	15.67	0.842
20	14.69	17.40	0.400
21	14.47	17.63	0.324
22	14.53	17.57	0.347
23	15.34	16.70	0.673
24	14.88	17.20	0.467
25	13.75	18.40	0.148
26	15.91	16.10	0.952

It is evidenced that for questions 1, 3 and 5 the mean rank for the individuals with sad expression are significantly higher than the rank for those with neutral expression. This means that the values obtained in the questionnaire were significantly higher for individuals with a sad expression. Then, this kind of users finds easier to read the text with a) the used font type and size and b) the text alignment and spacing. Users with sad expression have also a stronger conviction that colour / graphics in the website enhance navigation. For questions 2 and 15 the test evidences that differences, if not so strong, are still present. So, the font colour is slightly more appealing for sad users and they find a bit more obvious the buying button and links in the website. For the rest of the questions, tests do not show significant differences.

5 CONCLUSION

These results of the study are initially confusing. The three questions that evidence a strong relation with subjects showing a sad expression are emphasizing positive aspects of the design (regarding font, text layout and readability and colours selection). That is, results suggest that users who appreciate these aspects of the design as positive are sad. After analysing the results, we reviewed the pictures classified as sad by the algorithms, and what we found out is that most of them seem to be concentrated instead of sad. Thus, it seems more reasonable to think that users that found out the application more useable where more focused on the task and so, were classified as sad.

Obviously, this is just a conjecture, and should be checked with a new experiment, so we cannot assume that as a valid result of the experiment. Nevertheless, results suggest that there is a relation between the subjective perception of some of the aspects of the design and the information we gather using AC algorithms. This encourages us to repeat the experiment with a bigger and controlled sample of users.

ACKNOWLEDGEMENTS

This work was funded by the Scientific and University Department – SCAC of the Embassy of France in Spain (French Institute of Spain REF: Grant 02/2021) and the Department of Science, Innovation and Universities (Spain) under the National Program for Research, Development and Innovation (Project RTI2018-099235-B-I00).

REFERENCES

- Bhandari, U., Neben, T., Chang, K., & Chua, W. Y. (2017). Effects of interface design factors on affective responses and quality evaluations in mobile applications. *Computers in Human Behavior*, 72, 525–534. <https://doi.org/10.1016/j.chb.2017.02.044>
- Cooke, L. (2004). Improving usability through eye tracking research. *IEEE International Professional Communication Conference*, 195–198. <https://doi.org/10.1109/IPCC.2004.1375297>
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*, 13(3), 319–339. <https://doi.org/10.2307/249008>
- Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2001). Atmospheric qualities of online retailing: A conceptual model and implications. *Journal of Business Research*, 54(2), 177–184. [https://doi.org/10.1016/S0148-2963\(99\)00087-9](https://doi.org/10.1016/S0148-2963(99)00087-9)
- Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2003). Empirical Testing of a Model of Online Store Atmospherics and Shopper Responses. *Psychology and Marketing*, 20(2), 139–150. <https://doi.org/10.1002/mar.10064>
- Faisal, C. M. N., Gonzalez-Rodriguez, M., Fernandez-Lanvin, D., & Andres-Suarez, J. De. (2017). Web Design Attributes in Building User Trust, Satisfaction, and Loyalty for a High Uncertainty Avoidance Culture. *IEEE Transactions on Human-Machine Systems*, 47(6), 847–859. <https://doi.org/10.1109/THMS.2016.2620901>
- Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12), 1050–1057. <https://doi.org/10.1111/j.1467-9280.2007.02024.x>
- Gasah, M., Hidayah, N., Zain, M., & Baharum, A. (2019). An approach in creating positive emotion for children's e-learning based on user interface design. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(3), 1267–1273. <https://doi.org/10.11591/ijeecs.v13.i3.pp1267-1273>
- GitHub - WuJie1010/Facial-Expression-Recognition.Pytorch: A CNN based pytorch implementation on facial expression recognition (FER2013 and CK+), achieving 73.112% (state-of-the-art) in FER2013 and 94.64% in CK+ dataset. (n.d.). Retrieved May 17, 2022, from <https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch>
- GitHub - xionghc/Facial-Expression-Recognition: Facial-Expression-Recognition in TensorFlow. Detecting faces in video and recognize the expression(emotion). (n.d.). Retrieved May 17, 2022, from <https://github.com/xionghc/Facial-Expression-Recognition>
- Johanssen, J. O., Bernius, J. P., & Bruegge, B. (2019). Toward usability problem identification based on user emotions derived from facial expressions. *Proceedings - 2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering, SEmotion 2019*, 1–7. <https://doi.org/10.1109/SEmotion.2019.00008>
- Kemppainen, T., Makkonen, M., & Frank, L. (2020). Exploring online customer experience formation: How do customers explain negative emotions during online shopping encounters? *32nd Bled EConference Humanizing Technology for a Sustainable Society, BLED 2019 - Conference Proceedings, June*, 655–675. <https://doi.org/10.18690/978-961-286-280-0.35>
- Lasa, G., Justel, D., & Retegi, A. (2015). Eyeface: A new multimethod tool to evaluate the perception of conceptual user experiences. *Computers in Human Behavior*, 52, 359–363. <https://doi.org/10.1016/j.chb.2015.06.015>
- Ma, H., & Yarosh, S. (2021). A Review of Affective Computing Research Based on Function-Component-Representation Framework. *IEEE Transactions on Affective Computing*, 1–1. <https://doi.org/10.1109/TAFFC.2021.3104512>
- Maia, C. L. B., & Furtado, E. S. (2016). A study about psychophysiological measures in user experience monitoring and evaluation. *ACM International Conference Proceeding Series, Part F1280*. <https://doi.org/10.1145/3033701.3033708>
- Makkonen, M., Riekkinen, J., Frank, L., & Jussila, J. (2020). The effects of positive and negative emotions during online shopping episodes on consumer satisfaction, repurchase intention, and recommendation intention. *32nd Bled EConference Humanizing Technology for a Sustainable Society, BLED 2019 - Conference Proceedings, June*, 931–953. <https://doi.org/10.18690/978-961-286-280-0.49>
- Marangunic, N., & Granic, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Universal Access in the Information Society*, 14(1), 81–95. <https://doi.org/10.1007/s10209-014-0348-1>
- McDuff, D., Kaliouby, R. El, Cohn, J. F., & Picard, R. W. (2015). Predicting Ad Liking and Purchase Intent: Large-Scale Analysis of Facial Responses to Ads. *IEEE Transactions on Affective Computing*, 6(3), 223–235. <https://doi.org/10.1109/TAFFC.2014.2384198>
- Mehrabian, Albert., & Russell, J. A. (James A. (1974). *An approach to environmental psychology*. 266.
- Pham, L., Vu, T. H., & Tran, T. A. (2020). Facial expression recognition using residual masking network. *Proceedings - International Conference on Pattern*

- Recognition*, 4513–4519. <https://doi.org/10.1109/ICPR48806.2021.9411919>
- Plutchick, R. (1980). A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion* (pp. 3–33). Academic Press. <https://doi.org/10.1016/b978-0-12-558701-3.50007-7>
- Rakoczi, G., & Pohl, M. (2012). Visualisation and analysis of multiuser gaze data: Eye tracking usability studies in the special context of e-learning. *Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012*, 738–739. <https://doi.org/10.1109/ICALT.2012.15>
- Sahi, G. (2015). User satisfaction and website usability: Exploring the linkages in B2C E-commerce context. *2015 5th International Conference on IT Convergence and Security, ICITCS 2015 - Proceedings*. <https://doi.org/10.1109/ICITCS.2015.7293034>
- Staiano, J., Menéndez, M., Battocchi, A., De Angeli, A., & Sebe, N. (2012). *UX_Mate: From Facial Expressions to UX Evaluation*. 741. <https://doi.org/10.1145/2317956.2318068>
- Tang, K., & Hsiao, C.-H. (2016). The Literature Development of Technology Acceptance Model. *International Journal of Conceptions on Management and Social Sciences*, 4(1), 1–4.

Design Requirements for the Definition of Haptic Messages for Automated Driving Functionalities

Joseba Sarabia^{1,2}^a, Sergio Diaz¹^b, Asier Zubizarreta²^c and Joshué Perez¹^d

¹Basque Research and Technology Alliance, Tecnalia, Arteaga street, 25, 48160 Derio, Bizkaia, Spain

²University of the Basque Country UPV/EHU, Faculty of Engineering in Bilbao, Spain
jsarabia003@ehu.ikasle.eus, joseba.sarabia@tecnalia.com

Keywords: Haptic and Tangible Devices, Automated Driving, Design and Evaluation.

Abstract: As the number of advanced driving assistance systems grows, there is an increasing number of interactions between the driver and the automated system of the vehicle. Requests to transfer control from the automated system to the driver, new information sources increasing driver workload, and safety-critical situations imply new challenges in the communication between the driver and the automated system. In this context, haptic feedback for steering-wheel control has proven to be a valuable strategy. This work aims to propose a novel description of the requirements needed for a driver-in-the-loop system capable of ensuring safety while providing haptic feedback to the driver. Furthermore, a set of haptic patterns for the steering wheel are proposed based on the described requirements, to be evaluated in future studies. As future steps of this study, a continuation of this study will be published focused on human centered factors of the driver.

1 INTRODUCTION

In recent years, a rising number of Advanced Driving Assistance Systems (ADAS) are being implemented in modern vehicles (Ziebinski et al., 2017). Current ADAS take operational functionalities from the driver such as longitudinal (Adaptive Cruise-Control or Emergency Collision Avoidance) and lateral control (Lane Keeping).

However, with the increase in the level of automation of vehicles as defined by SAE J3016, vehicle control systems take charge over more and more functionalities. Many works have suggested that reducing human intervention in automated vehicles will increase road safety (Lv et al., 2018), (Li et al., 2019) and (Morales-Alvarez et al., 2020). Hence, an important effort has been conducted by the automotive industry to develop functionalities with higher levels of automation.

In the transition to fully automated vehicles, the driver still has a fundamental role. When considering Automated Driving Levels (ADL), only ADL 4 and 5, conditional and high automation, completely

remove the driver from the driving task. Functionalities with a medium or conditional level of automation (ADL 2 and 3) require the driver to supervise or partially control the vehicle (Marcano et al., 2020a). In these functionalities, the Takeover Request (TOR) plays a key role in the transition between the automated mode (controlled by the vehicle) and the manual mode (controlled by the driver) (Sasangohar and Cummings, 2010) (Mulder et al., 2008).

In this context, new challenges arise to ensure safe cooperation and interaction between the driver and the vehicle control systems, so that ADL changes such as TOR are properly executed by all agents. The driver-automated vehicle interface currently relies on approaches based on auditory or visual channels, such as Heads Up Displays (HUD), infotainment displays, and auditory alarms (Politis et al., 2015). However, these approaches can overstimulate the driver (Mehler et al., 2021), reducing their effectiveness for time-critical tasks (Strayer et al.,), such as TOR (Zhang et al., 2022).

As an alternative, haptic messages have been proposed, which use the actuators of the vehicle as feedback to the driver. Haptic messages can provide swift control transfer between the driver and the automated vehicle in cases such as TOR (Morales-Alvarez et al., 2020). Moreover, it has been demonstrated that haptic guidance considerably reduces the cognitive work-

^a <https://orcid.org/0000-0002-1759-4365>

^b <https://orcid.org/0000-0002-2617-2121>

^c <https://orcid.org/0000-0001-6049-2308>

^d <https://orcid.org/0000-0001-8328-9978>

load of the driver (Wang et al., 2018) (Shakeri et al., 2016), and that it improves handover transitions especially when applied together with visual feedback (Vito et al., 2020). A good illustration of this is the study conducted by the NHTSA (National Highway Traffic Safety Administration) where they compared different types of signals to alert drivers of a TOR. There, it was shown that the haptic feedback together with visual alerts generated reaction times with an average of 1.3 seconds, while only visual alerts took 4.8 seconds (Marinik et al., 2014). Regarding auditory cues, (Harrison et al., 2010) shows that they have a considerably short response time, especially when used together with visual feedback.

Looking for short response time channels, several authors have proposed works about situational awareness focused on retrieving the driver from the distracted to attentive in the fastest possible way. Auditory and haptic channels show the best time results (Ploch et al.,). However, while there is an abundance of auditory cues, the haptic channel is rarely overloaded. This facilitates the haptic signals to be more rapidly perceived (Strayer et al., 2003). Furthermore, haptic perception appears to be resilient to high cognitive load conditions (Törnros and Bolling, 2005). However, this is still an open research area, as strategies to define the haptic messages, and the definition of these in terms of evaluation of their acceptability by users in real and simulated scenarios is still missing. In particular, note that when considering a medium level of automation, the driver is generally required to have at least one hand on the steering wheel. This additionally opens a gap for ADAS to be complemented with haptic feedback through the steering wheel.

About the type of equipment used for the haptic feedback, the current literature is mainly focused on small actuators installed on the steering wheel. However, there are fewer studies where the main steering actuator of the EPS (Electric Power Steering) is used. This is generally used to provide guidance (Beruscha et al., 2011), or used with low frequency profiles (Talamonti et al., 2017).

In the context of human machine interaction, the HADRIAN European Project has been proposed (had,). This project, funded by the EU Horizon 2020 program, aims at creating a novel interaction approach between the driver, the vehicle, and the road infrastructure. This interaction pretends to be adaptive towards the state of the driver and to the operational design domain, conveying an idea of fluid interaction with different HMIs (display, alarms, HUD, haptic feedback, etc.). In this sense, haptic devices are proposed as a tool to communicate with the driver regarding drive mode transitions, take-over requests, or mes-

sage delivery. Hence, in this framework, two main contributions are proposed in this work: 1) a driving simulation framework, which ensures human-in-the-loop testing for haptic messages, as a tool to evaluate future works in this area; and 2) a set of design specifications for haptic messages using the steering wheel.

The rest of the work is structured as follows. Section II describes the equipment used to study the haptic feedback. Section III points out the requirements for safe use of the haptic feedback in steering wheels. Section IV proposes a methodology to design a set of haptic icons that fulfils the previous requirements. Section V shows the resultant haptic icons. Section VI offers a concluding overview of the obtained results and points to future works where user-centered studies validate the proposed icons.

2 DRIVING SIMULATOR FOR HAPTIC HMI TESTING

As stated in the introduction, there is a need to further test and evaluate approaches that consider the driver and automated vehicle interaction. In this work, a framework based on a driving simulator with a haptic steering wheel has been proposed, developed in the context of the HADRIAN project.

The proposed setup is detailed in Fig. 1, which is based on an off-the-shelf racing simulator, in which an Augury H kit haptic steering wheel has been integrated.

As seen in the Figure, the simulation setup includes three 32 inches LDC screens, a 120W stereo sound system, racing simulator pedals with mechanical damping, a fully equipped PC that runs the simulated vehicle and environment, a touch display to emulate secondary activities by the driver and a racing seat.

The simulation environment is based on Dynacar (Marcano et al., 2020b), developed by Tecnalia, which allows realistic vehicle dynamics modeling in real-time, including force feedback for the driver. Dynacar simulation framework includes a map editor and allows 3D representation of the vehicle in the selected scenario, so that it can be easily used as a driving simulation. A Mercedes Class E vehicle has been selected as the study case implemented in the setup (Marcano et al., 2020a). The inertia and damping of the steering system have been simulated based on the previously described case, then the actuator provides a more accurate response.

As previously detailed, the simulation setup has been complemented with the introduction of a haptic steering wheel, with the aim of testing and evaluating



Figure 1: Simulator setup.

vehicle-driver shared control approaches, and interaction strategies. The haptic wheel setup is composed of three elements: a wheel, a motor controller, and a motor.

The servomotor is a brushless DC motor in which a current sensor and an incremental encoder have been attached. The actuator is a MiGE model 130ST, which can exert a maximum torque of 18 Nm and provides 1.5 KW.

The racing steering wheel is a standard metal-core and leather-cover steering wheel used in common driving simulators. Thanks to its metallic core and its tight fixation to the actuator, all vibrations are directly transmitted to the hands of the driver. The steering wheel includes a small Bluetooth 3.0 controller fixed at the right spoke, allowing it to be used with the right thumb. Its purpose is to change the driving mode from Automated to Manual modes (and vice versa).

Finally, the motor controller is composed of a Simucube motherboard, an IONI PROHC driver control unit with a maximum intensity of up to 25A, a safety emergency button and all the necessary interfaces to connect it to the simulator computer.

3 DESIGN REQUIREMENTS FOR THE DEVELOPMENT OF HAPTIC MESSAGES

The setup will be considered as the basis for developing and evaluating haptic messages for the driver. However, some requirements must be considered when defining haptic feedback signals for steering wheels.

The steering wheel is the lateral controller of the vehicle. Modifying its dynamics through vibrations can lead to potential control loss or deviations from

the vehicle trajectory. Steering wheels can be coupled or uncoupled (Marcano et al., 2020a). If it is uncoupled (or drive-by-wire), there is no such constraint, as vibrations do not necessarily need to be transmitted to the movement of the wheels. If the steering is coupled, however, safety issues must be considered, among others.

From one side, haptic feedback should be felt in the hands of the driver but must not affect the trajectory of the vehicle. However, this goes against the nature of the working principle of the electric-driven steering wheel, as its movement is generated through its shaft. Hence, haptic signals must be centered around zero, so that the net movement produced by haptic vibrations is counterbalanced. This way, no lateral deviation from the trajectory will be caused by the haptic pattern.

From the other, haptic patterns in the steering wheel must be distinguishable from vibrations in the vehicle caused by road roughness. At the end of the day, the steering wheel is by itself an interface to feel the road. If haptic patterns may induce the driver to think that there was a bump in the road, or to believe that something happened to the tires, this could have a degrading effect on the perception of the road and, consequently, on vehicle control.

Regarding the functionality of the steering wheel, haptic patterns should not block or reduce the driver's control over it. When a haptic signal is activated by the system, the driver must be able to keep controlling the trajectory of the vehicle as if there was no such signal. Thus, the haptic pattern must not only be balanced around zero considering the entire pattern but also in small intervals of time, so that there is no torque applied in the steering wheel long enough to change its trajectory or to be excessive to be overridden by the driver.

In addition, regarding the durability of the system, the haptic patterns must not damage the mechanical components of the steering column. Vibrations are a widely studied field of mechanics and can lead to serious damage if they are not previously considered (Rouillard, 2014). When tuning the frequencies of the haptic patterns, natural frequencies of the system must be avoided. In this work, this is not the main focus, as a driving simulation environment is used. Future works require a study case when the proposed haptic patterns are brought to real vehicles.

4 DESIGN METHODOLOGY FOR HAPTIC MESSAGES

Based on the requirements analyzed in the previous section, in this one, the three main design strategies followed to design the set of haptic messages proposed in this work are defined.

4.1 Definition of Zero-centered Patterns

Regarding the zero-centered haptic patterns, in this work, two types of patterns are proposed. On the one side, oscillatory motion patterns. That is, haptic patterns composed of sine-based vibrations. And on the other side, short pulses that are felt like little strokes on the steering. The latter also have the same positive and negative component. This way, the total applied torque of the actuator will produce a null amount of work in both cases, generating no total displacement.

To achieve the desired haptic patterns, the actuator must be capable of reproducing the desired pattern properly. In addition, internal frictions and nonlinearities of the system may cause lateral displacements. To evaluate these effects, the following experiments have been performed.

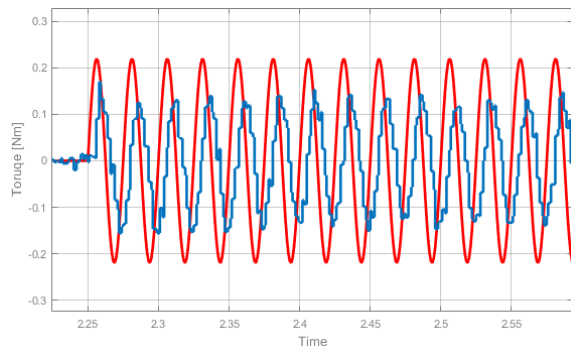


Figure 2: Torque response of the 1.5 s vibration (red-reference /blue-response).

First, the response obtained when sending a torque sine input in the steering wheel has been analyzed. Fig. 2 shows the torque reference input in red and the measured torque output in blue. Even if the measured signal is slightly discretized, and there is a phase shift, it can be observed that the response of the actuator is capable of following the set input in terms of frequency. Note that the main focus in this section is the haptic feeling provided by the system. In addition, this figure shows the worst-case scenario, this is, the response of the motor to the highest frequency at the reference, which sets the maximum frequency of the system (considering the simulated inertia and damping are part of the measured net torque) at 40

Hz. Regarding the amplitude, a higher value does not provide a difference in the output, as the torque output of the actuator has a gradient saturation, meaning that for such frequency, this is the highest value it will reach.

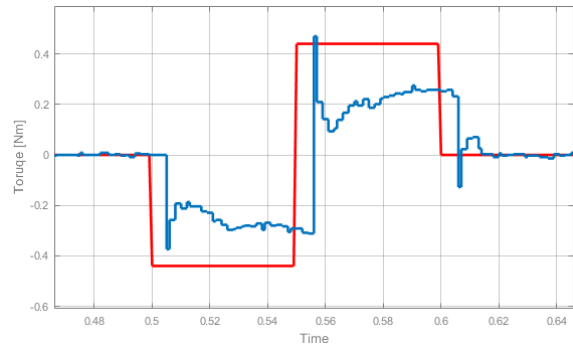


Figure 3: Torque response of the pulse signal.

Second, pulse responses have been tested, as shown in Figure 3. Similar to the previous case, the haptic feeling of the reference can be executed by the motor, as transitions (peak values), and then a constant torque area, can be defined. Note that apart from the reference torque sent, there are also the components of the simulated damping and the inertia, which are responsible for the transitory effects.

Hence, the aforementioned patterns (sine and pulse) are appropriate to develop haptic messages in the proposed setup.

4.2 Definition of Easily Distinguishable Patterns

If vehicle vibrations and haptic patterns are considered, as previously stated, those two sources of information must be distinguishable. Unfortunately, road vibrations may not have a regular pattern, as sporadic bumps or road roughness changes can cause unexpected vibrations in the vehicle. However, most common vibrations in vehicles are caused by the interaction between the roughness of the road and the surface of the tires.

Road roughness may be described as the elevation profile obtained along the wheel tracks over which vehicles pass. Such profiles fit the general category of broad-band random signals. When passed by a vehicle, its velocity determines the frequency at which such profile will be collided, generating vibrations on the wheels of the vehicle. Some studies (Rouillard, 2014) suggest that for a velocity range between 20 and 100 km/h, the predominant road vibration is between 1 and 10 Hz. But this is the vibration generated in the floor of the vehicle, after being filtered by the

damping of the vehicle. However, as (Agostinacchio et al., 2013) shows, the vibrations generated by the road roughness at driving speeds between 25 to 100 km/h, go up to 100 Hz and higher. And those are the ones that are felt in the steering wheel in coupled steering columns, as the system is in direct contact with the vehicle tires. Indeed, this is easily demonstrable just by taking a sound system and reproducing signals increasingly between 30 and 200 Hz. The obtained sound is equivalent to that heard while driving a car.

In (Gillespie and Sayers, 1981), a study was performed evaluating frequencies between 0 and 200 Hz. There, it is shown that a frequency variation to be haptically perceivable requires intervals of at least 10 Hz. Thus, a couple of conclusions can be obtained from here: first, having a 40 Hz upper limit established by the actuator, there are few frequencies left to design the haptic feedback patterns. And second, the haptic feedback will be in the range of the typical vibrations found in the vehicle steering.

For that reason, the approach to obtain some distinguishable haptic patterns will not be focused on specific frequency ranges. Instead, 3 other variables will need to be prioritized: first, the regularity of the signal; second, the duration; and third, its amplitude. As road vibrations have chaotic components, regular structured signals may provide a more trustworthy feeling of their presence. Regarding the duration and the amplitude, a balance between what is perceivable and what is uncomfortable will need to be obtained.

4.3 Definition of Patterns to Reduce the Effect on the Steering Wheel

As previously detailed, the applied haptic patterns must not affect the drive. However, it could be considered that applying specific torques in the steering column could derive in deviations of the vehicle trajectory and possible control lost. To check whether the proposed haptic patterns could really affect the safety of the drive, the following tests have been conducted.

First, a sequence of equal sine waves (0.4 Nm of amplitude, 40 Hz of frequency and 1.5 s of duration) has been sent as input to the system repeatedly during 300 s. The aim was to evaluate the angular displacement generated on the steering wheel. Results are shown in Fig 4.

Each of the repeated patterns zoomed in Figure 4 are equal to the ones presented in Figure 2. The aim of Figure 4 is to show the total deviation of the steering wheel after applying the haptic icon a repeated number of times (133 times in this case). After the test, the total deviation of the steering wheel shown in the

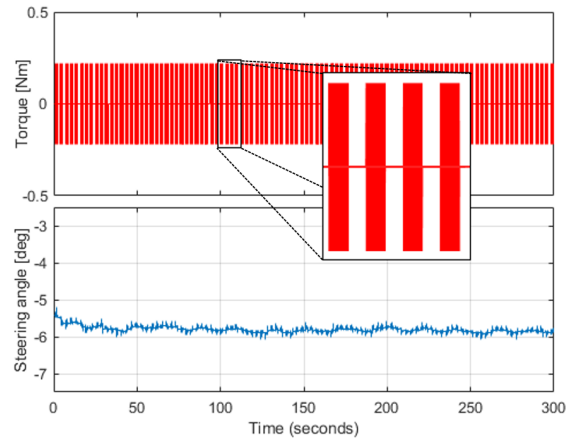


Figure 4: Sequence of sine patterns and angular deviation of the steering angle.

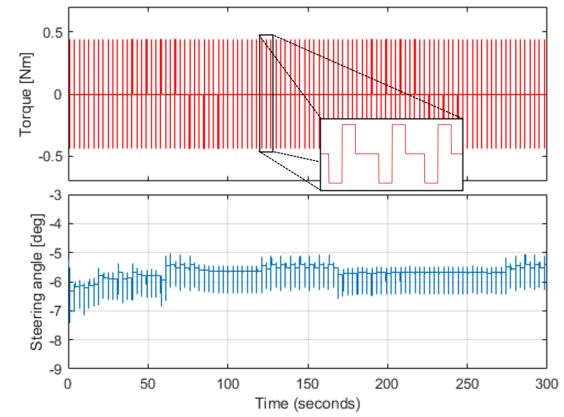


Figure 5: Sequence of pulse patterns and the angular deviation of the steering wheel.

second graph of the layout is less than 1°.

This same test has also been conducted for the other haptic pattern proposed, for the pulse (Figure 5). Results show, with the same amount of repetitions, an overall deviation of up to 2°. However, the instantaneous deviation of the steering angle is much higher for the pulse than for the sine-based vibration. This is because the amplitude achieved in the pulse is higher than in the first pattern, and the pulses are longer in time. To obtain a feeling about how potentially dangerous these deviations could be, a theoretical and a practical approach is proposed.

From the theoretical perspective, the effect of the steering wheel on vehicle orientation can be calculated by a model, such as the bicycle model formula. This is a stationary approximation, without considering the transitory damping effects of the tires that could happen during the haptic pattern duration. The equation goes as follows,

$$\beta = \arctg(l_r \tan(\delta)/L) \quad (1)$$

where β is the slip angle, l_r the distance between the center of gravity of the vehicle to the rear axle, L the distance between the rear and front axles, and δ the front wheels angle. As δ is proportional to the steering wheel angle, it can be directly related the slip angle with the steering wheel angle. The typical ratio between δ and the steering wheel angle tends to be around 15:1, meaning that a turn of 15° in the steering wheel implies 1° in the tires.

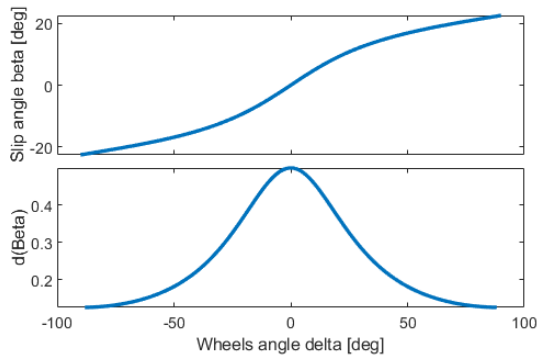


Figure 6: Graphic representation of the bicycle model formula and its derivative.

Figure 6 shows the graphic representation of the bicycle model. The maximum variation of the slip angle is at the maximum point of its derivative, that is, when δ is closer to zero. Thus, computing the calculation of the obtained 2° deviation for the pulse pattern, a slip angle of 0.0667° is obtained. This implies that, for instance, driving at a speed of 90 km/h (25m/s), an instantaneous deviation caused by the haptic pulse would deviate the trajectory of the vehicle 0.03m after one second. This is considering no damping of the system nor the driver hands.

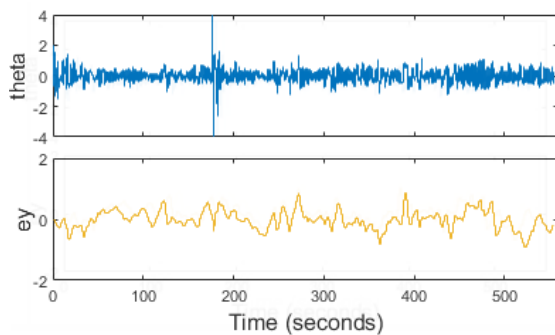


Figure 7: Steering wheel angle and lateral error recorded during a 560 s drive in a straight line in the simulator.

From an experimental perspective, a 10-minute ride has been tested in the simulator on a straight-line highway to measure the periodic fluctuation of a driver. As it can be seen in Figure 7, except for an outlier in second 190, the rest of the drive is within

a range between -2 and 2 degrees. In the lower sub-plot, the lateral deviation (in meters) is shown for the lane center. Considering the mean frequency of the driver making corrections, it is calculated that every 0.895 s a lateral correction cycle is performed. In other words, the lateral oscillation of the driver is approximately around 1 Hz.

This implies that the maximum lateral deviation caused by the haptic pattern would be inside the range of a deviation of a generic ride. Thus, as long as the driver has the hands on the steering wheel, the effect on the drive is negligible.

Based on the aforementioned design strategies, the following set of haptic messages have been proposed for three main approaches: message notification, take-over requests (TOR) and hand-over transitions (Figure 8).

Message notification patterns are designed as short, and mainly composed of pulses. They are intended to provide quick, non-intrusive and non-critical messages to the driver. That is, they are supposed not to be urgent, but to catch the attention of the driver easily.

TOR patterns are mainly composed of long vibrations, whose amplitude increases in steps as the urgency to take over the control increases. They are all 40 Hz in frequency and mainly last about 5 seconds. They are supposed to increase the level of urgency during the time, providing a countdown feeling.

Finally, the hand-over transition patterns are designed in two different ways. The first four approaches are designed looking for a continuous increase or decrease of the amplitude, while the resting four have a continuous change in frequency. This duality has been proposed to prove that amplitude-based transitions are more reliable than frequency-based transitions, as frequency jumps are less noticeable. Their design aim is that the driver receives a transition feeling, as if something was switching on and off.

This haptic set is defined so that the actuator system proposed in the driving simulation environment can reproduce the required vibrations to the driver. However, note although the haptic steering wheel can follow the defined haptic messages, user-centered studies will be required to evaluate the acceptance and effectiveness of the messages.

5 CONCLUSIONS AND FUTURE WORKS

Haptic feedback on the steering wheel will be highly beneficial in future ADAS, as there is a strong need

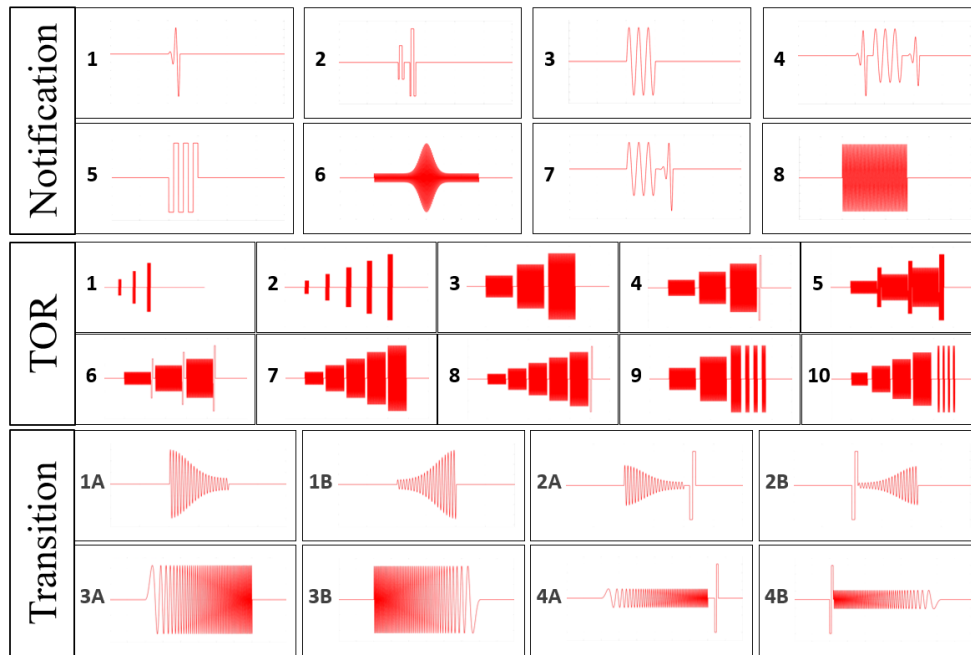


Figure 8: 1 to 8 icons: Message notification icons. 1 to 10 icons: Take-Over Request icons. 1A to 4B icons, hand-over transition icons.

for safe critical interfaces that ensure rapid intervention from the driver. Besides, haptic feedback will not imply a safety issue regarding vehicle control, even in the edgiest situations. Indeed, it will increase safety by improving the interaction between the driver and the vehicle and reducing driver workload. The main contribution of this work is a new HMI concept, based on a set of haptic messages for ADS, together with the defined requirements for the implementation of safe haptic messages.

The simulator shows how an off-the-shelf racing steering wheel can work as a haptic feedback interface, as it fulfills all the previously defined requirements. Some interactions with the haptic messages show their potential for ADS specific cases. Nevertheless, user-centered tests will be needed to evaluate and improve the proposed set of haptic patterns.

ACKNOWLEDGMENTS

This work is supported by the EU Commission HADRIAN project and the Government of the Basque Country by means of AUTOEV@L project (KK-2021/00123). HADRIAN has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875597. This document reflects only the author's view, the Innovation and Networks Executive Agency

(INEA) is not responsible for any use that may be made of the information it contains.

REFERENCES

- The hadrian project. <https://hadrianproject.eu/>. Accessed: 2022-02-01.
- Agostinacchio, M., Ciampa, D., and Olita, S. (2013). The vibrations induced by surface irregularities in road pavements-a Matlab® approach.
- Beruscha, F., Augsburg, K., and Manstetten, D. (2011). (9) (PDF) Haptic warning signals at the steering wheel: A literature survey regarding lane departure warning systems.
- Gillespie, T. D. and Sayers, M. (1981). Transportation Research Record 836 15 Role of Road Roughness in Vehicle Ride.
- Harrison, N. R., Wuerger, S. M., and Meyer, G. F. (2010). Reaction time facilitation for horizontally moving auditory-visual stimuli. *Journal of Vision*, 10(14):16–16.
- Li, Y., Lv, C., and Xue, J. (2019). A novel predictive haptic control interface for automation-to-human takeover of automated vehicles. *IEEE Intelligent Vehicles Symposium, Proceedings*, 2019-June:994–999.
- Ly, C., Wang, H., Cao, D., Zhao, Y., Sullman, M., Auger, D. J., Brighton, J., Matthias, R., Skrypchuk, L., and Mouzakitis, A. (2018). A Novel Control Framework of Haptic Take-Over System for Automated Vehicles. *IEEE Intelligent Vehicles Symposium, Proceedings*, pages 1596–1601.

- Marcano, M., Diaz, S., Perez, J., and Irigoyen, E. (2020a). A Review of Shared Control for Automated Vehicles: Theory and Applications. *IEEE Transactions on Human-Machine Systems*, 50(6):475–491.
- Marcano, M., Diaz, S., Perez, J., and Irigoyen, E. (2020b). A Review of Shared Control for Automated Vehicles: Theory and Applications. *IEEE Transactions on Human-Machine Systems*, 50(6):475–491.
- Marinik, A., Bishop, R., Fitchett, V., and NHTSA (2014). Human factors evaluation of level 2 and level 3 automated driving concepts: Concepts of operation. Technical report, NHTSA, Washington DC.
- Mehler, B., Brookhuis, K., Meteier, Q., Khaled, A. O., Capallera, M., Ruffieux, S., Angelini, L., Khaled, O. A., Mugellini, E., Widmer, M., and Sonderegger, A. (2021). Article 596038 (2021) classification of drivers' workload using physiological signals in conditional automation. *Frontiers in Psychology — www.frontiersin.org*, 12:596038.
- Morales-Alvarez, W., Sipele, O., Léberon, R., Tadjine, H. H., and Olaverri-Monreal, C. (2020). Automated driving: A literature review of the take over request in conditional automation.
- Mulder, M., Abbink, D. A., and Boer, E. R. (2008). The effect of haptic guidance on curve negotiation behavior of young, experienced drivers. In *2008 IEEE International Conference on Systems, Man and Cybernetics*, pages 804–809.
- Ploch, C. J., Bae, J. H., Ploch, C. C., Ju, W., and Cutkosky, M. R. Comparing Haptic and Audio Navigation Cues on the Road for Distracted Drivers with a Skin Stretch Steering Wheel.
- Politis, I., Brewster, S., and Pollick, F. (2015). Language-based multimodal displays for the handover of control in autonomous cars. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '15, page 3–10, New York, NY, USA. Association for Computing Machinery.
- Rouillard, V. (2014). Quantifying the Non-stationarity of Vehicle Vibrations with the Run Test. *Packaging Technology and Science*, 27(3):203–219.
- Sasangohar, F. and Cummings, M. (2010). Human-system interface complexity and opacity.
- Shakeri, G., Ng, A., Williamson, J. H., and Brewster, S. A. (2016). Evaluation of haptic patterns on a steering wheel. *AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, *Proceedings*, pages 129–136.
- Strayer, D. L., Cooper, J. M., Goethe, R. M., Mccarty, M. M., Getty, D. J., and Biondi, F. Assessing the visual and cognitive demands of in-vehicle information systems.
- Strayer, D. L., Drews, F. A., and Johnston, W. A. (2003). Cell Phone-Induced Failures of Visual Attention During Simulated Driving. *Journal of Experimental Psychology: Applied*, 9(1):23–32.
- Talamonti, W., Tijerina, L., Blommer, M., Swaminathan, R., Curry, R., and Ellis, R. D. (2017). Mirage events & driver haptic steering alerts in a motion-base driving simulator: A method for selecting an optimal HMI. *Applied Ergonomics*, 65:90–104.
- Törnros, J. E. and Bolling, A. K. (2005). Mobile phone use - Effects of handheld and handsfree phones on driving performance. *Accident Analysis and Prevention*, 37(5):902–909.
- Vito, P. D. C. S., Brown, E., Brewster, S., Pollick, F., Thompson, S., Skrypchuk, L., and Mouzakitis, A. (2020). Haptic Feedback for the Transfer of Control in Autonomous Vehicles. *Adjunct Proceedings - 12th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications*, *AutomotiveUI 2020*, pages 34–37.
- Wang, Z., Zheng, R., Kaizuka, T., and Nakano, K. (2018). Influence of haptic guidance on driving behaviour under degraded visual feedback conditions. *IET Intelligent Transport Systems*, 12(6):454–462.
- Zhang, H. ., Zhang, Y. ., Xiao, Y. ., Wu, C., Xie, K., Ma, X., Zhang, H., Zhang, Y., Xiao, Y., and Wu, C. (2022). Analyzing the Influencing Factors and Workload Variation of Takeover Behavior in Semi-Autonomous Vehicles. *International Journal of Environmental Research and Public Health* 2022, Vol. 19, Page 1834, 19(3):1834.
- Ziebinski, A., Cupek, R., Grzechca, D., and Chruszczyk, L. (2017). Review of advanced driver assistance systems (ADAS). *AIP Conference Proceedings*, 1906(1):120002.

RetoñosApp: Work in Progress on a Platform to Support the Teaching of Programming in CS through the Automation and Customization of Learning Processes Guided by Artificial Intelligence

David Mauricio Valoyes-Porras^a, Juan Sebastián Rodríguez-Obregón^b,
David Steven Salamanca-Sánchez^c and Miguel Alfonso Feijóo-García^d

Program of Systems Engineering, Universidad El Bosque, Bogotá, Colombia
{dvaloyes, jsrodriguez, dsalamancas, mfeijoog}@unbosque.edu.co

Keywords: User-centered Interaction, User Experience, Artificial Intelligence, Recommendation System, Computer Science Education.

Abstract: Learning difficulties in Computer Science (CS) are a multicausal problem that promotes student dropout in CS undergraduate programs. This results from the students' psychological, emotional, and motivational implications, affecting their academic performance. We present RetoñosApp, as a web-based and user-centered platform supported by Artificial Intelligence (AI) that assists the teaching and learning processes for CS. It fosters the students' autonomous learning, and provides accompaniment and feedback to students during their academic term, and CS instructors on their students' learning processes. This web-based tool uses a Conversational Bot as an autonomous and synchronous virtual tutor, and a Content-based Recommendation System to generate customized reports with "educational routes" to students and instructors, based on their needs. We evaluated this web-based tool, and reported the findings and results, considering its efficiency and effectiveness, based on the participants' interaction. This, in order to answer how the platform supported and complemented the teaching and learning processes of programming in CS, evaluating its potential to be part of the educational methodology of further CS undergraduate courses. Our findings from the pilot study suggest that RetoñosApp effectively provides a friendly user-centered asynchronous assistance and enhancement to learning processes, and frequent feedback on teaching processes.

1 INTRODUCTION

Computer Science (CS) is the discipline that gathers the fundamentals, methods, and theory to support the development of new informatic solutions (Sommerville and Torres, 2011). Consequently, CS education constitutes a considerably high degree of complexity. Programming skills, as basic and fundamental knowledge of CS, present one of the main challenges that professionals in this field often face. These difficulties result from the requirement of teaching-learning processes to be incremental and evolving in CS, based on curricular approaches that gather previous skills and knowledge. This has motivated the CS Education community to evaluate teaching and learning strategies often to face this educational challenge.

There are numerous technologies and research with scaffolds on the following topics: (1) interactive platforms as a complement to the educational processes that are synchronous-based, (2) conversational bots as virtual tutors, and (3) recommendation systems for the abstraction of shortcomings and difficulties in educational processes.

In the first place, interactive e-learning tools for synchronous assistance of academic activities encourage the students' participation in constructing their knowledge (Mutiaiwani et al., 2014). On the other hand, integrating conversational bots as virtual tutors promotes student learning in educational environments mediated by information and communication technologies (ICTs) (Wellnhammer et al., 2020). In addition, implementing recommendation systems in education (Cano and Alarcón, 2021) allows identifying students' learning difficulties to support and provide feedback on educational processes based on their needs (Harris and Kumar, 2018).

^a  <https://orcid.org/0000-0003-2893-0261>

^b  <https://orcid.org/0000-0001-7039-244X>

^c  <https://orcid.org/0000-0001-9355-3955>

^d  <https://orcid.org/0000-0001-5648-9966>

We introduce a web-based tool called *RetoñosApp* as a friendly and user-centered platform supported by Artificial Intelligence (AI) that assists the teaching and learning processes for CS. It fosters the students' autonomous learning, and provides frequent accompaniment and feedback to students during their academic term, and CS instructors on the overall learning progress of their students. Like LEGO (McNamara et al., 1999), this approach pretends the students to guide their customized “learning route”, following significant learning. The Significant Learning Model (Baque-Reyes and Portilla-Faican, 2021) constitutes a constructivist teaching model, where the student is not only a receiver, but the protagonist of their learning process. The constructivist theory (Saldarriaga-Zambrano et al., 2016) asserts that teaching-learning processes should be based on the construction of knowledge from enriching experiences (i.e., individual needs), beyond the simple transmission of concepts or skills.

This paper presents *RetoñosApp* and the results of our first pilot study using the web-based tool. We piloted the study with first-year undergraduate students (i.e., CS1 and CS2) at Universidad El Bosque, Colombia. We present our findings and results on students' perceptions based on their interaction with the tool, and the impact of our approach on their learning processes. We also describe its limitations, addressing the pros and cons of this preliminary pilot approach.

With this study, we intended to validate the following hypothesis: “Using a teaching-learning platform based on Artificial Intelligence (AI) satisfactorily benefits, supports, and complement the educational processes of programming, providing customized feedback on particular topics to the students, and the overall groups' progress to the instructors”. Thus, we ask the following questions: How do we promote a clear customized learning roadmap in CS? How do we make this roadmap to be learner-centered?

Our work contributes to CS Education and Computer-Human Interaction literature, evaluating a web-based approach to support teaching and learning processes in CS in a customizable way. This approach uses the benefits of autonomous, synchronous, and customized educational processes addressing CS concepts and skills.

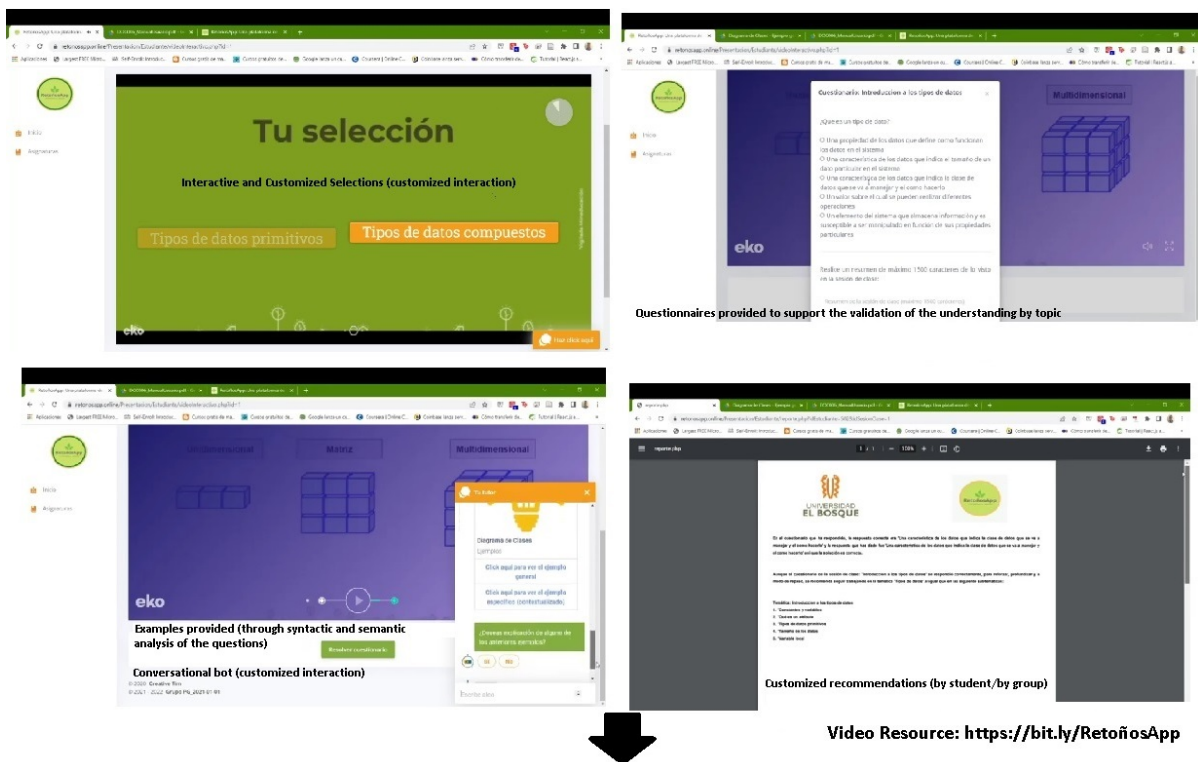
2 *RetoñosApp*: DESCRIPTION

We introduce a web-based platform that supports and complements the teaching and learning processes in

CS programming courses. It provides a friendly and interactive GUI that allows the user to guide customized “educational routes” through asynchronous assistance and frequent feedback. *RetoñosApp* features a conversational bot that fulfills the task of an autonomous and synchronous virtual tutor— we implied SAP Conversational AI (Adamopoulou and Moussiades, 2020), which involves Natural Language Processing (Chowdhary, 2020) (i.e., NLP, involving syntactic and semantic processing). Moreover, it features a content-based recommendation system (Pazzani and Billsus, 2007) fed by the information collected by the platform (i.e., user entries) and the conversational bot (e.g., topic-centered questions, personalized user interaction). *RetoñosApp* provides a customized “educational route” to students, or the overall progress of a group of students of a particular course to the instructors. The customized report is based on the students' particular needs, doubts, and difficulties (see Figure 1).

This approach used techniques and concepts of CS, Data Analytics, and Intelligent Systems. *RetoñosApp* is structured by three main modules: (1) Administrator, (2) Teaching, and (3) Learning. The first module (i.e., Administrator) presents the management and audit features of the platform (e.g., parameter configuration, user management, incident management, traceability verification, display or exportation of reports). In addition, it features a friendly and user-centered GUI that implies that administrators (i.e., users) do not necessarily have to understand technology to interact with these features. On the other hand, the second module (i.e., Teaching) provides features to instructors (e.g., creation of subjects, creation of topics, creation of class sessions, analysis/exportation of customized “educational routes” of a group of students— strictly related to the recommendation system). Finally, the third module (i.e., Learning) presents to the students the asynchronous work proposed as a support and complement to the teaching process in their programming courses in CS. This third module guides and presents monitoring to the students' customized educational progress— an individual educational roadmap. It also features the interaction with a conversational bot that works as a virtual tutor and a “synchronous” and “autonomous” guide that intends to accompany the student during their learning process on the platform.

Architecturally, *RetoñosApp* presents a simultaneous, coordinated, and complimentary use of a (1) Clean Architecture (based on a hexagonal architecture) (Martin et al., 2018), (2) a layered architecture— consisting of a combination of traditional layered architecture (Richards, 2015) and a



EDUCATIONAL ROADMAP COSTUMIZATION - RECOMMENDATIONS (CUSTOMIZED INDIVIDUAL/GRUPAL REPORTS) SUPPORTED BY AI

Figure 1: Features/Functionalities of RetoñosApp (preliminary customization features).

Model–View–Controller (MVC) architectural style (Deacon, 2009), and (3) an attribute-based architecture (Klein et al., 1999)— representing the McCall’s quality model (Moreno et al., 2010).

The three architectural styles or approaches selected constitute three different perspectives of the application architecture design. Each of the styles used allows presenting and emphasizing a particular aspect of the general architecture of the application, dividing it into: (1) Technological infrastructure and deployment of the artifact, (2) Architectural design based on the internal and specific components of the application, and (3) Determination of the quality model selected as a guide for the quality assurance of the project.

This web-based approach proposes three main differentiator features: (1) Interactivity as an element that gamifies the learning experience, promoting motivation, retention of information, ease in understanding new topics and concepts, and individualizing learning forms. (2) Using conversational bots to perform virtual tutoring tasks in real-time. (3) Using of recommendation systems to provide feedback on the learning processes, consequently allowing adapting or reconciling the teaching processes based on in-

dividual and group needs of students (See Figure 2 and Figure 3).

3 DATA ACQUISITION

The participants who gathered the experimentation voluntarily were Junior Undergraduates (i.e., first-year CS students). We expected to have 188 participants in the experimentation. However, we had a total of 136 students ($N=136$), representing 72.34%: 71.3% ($N=97$) for CS1, and 28.7% ($N=39$) for CS2. The participants had an average age between 18 and 19 years old. We considered this a significant population sample to support the claims presented in this experience report.

We sought to answer how the platform supported and complemented the teaching and learning processes of programming in CS, evaluating its potential to be part of the educational methodology of further CS undergraduate courses. Hence, we evaluated the tool’s effectiveness, considering the participants’ perceptions of their experience using RetoñosApp. We designed a questionnaire that allowed the participants to evaluate their user experience (Hassenzahl

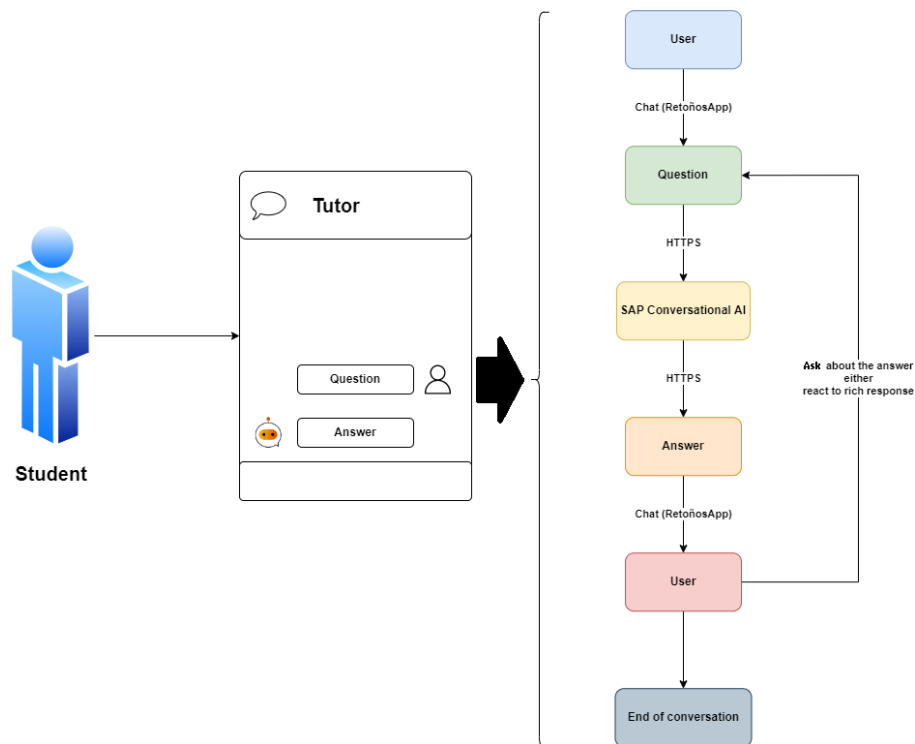


Figure 2: Integration of the Conversational Bot to RetoñosApp.

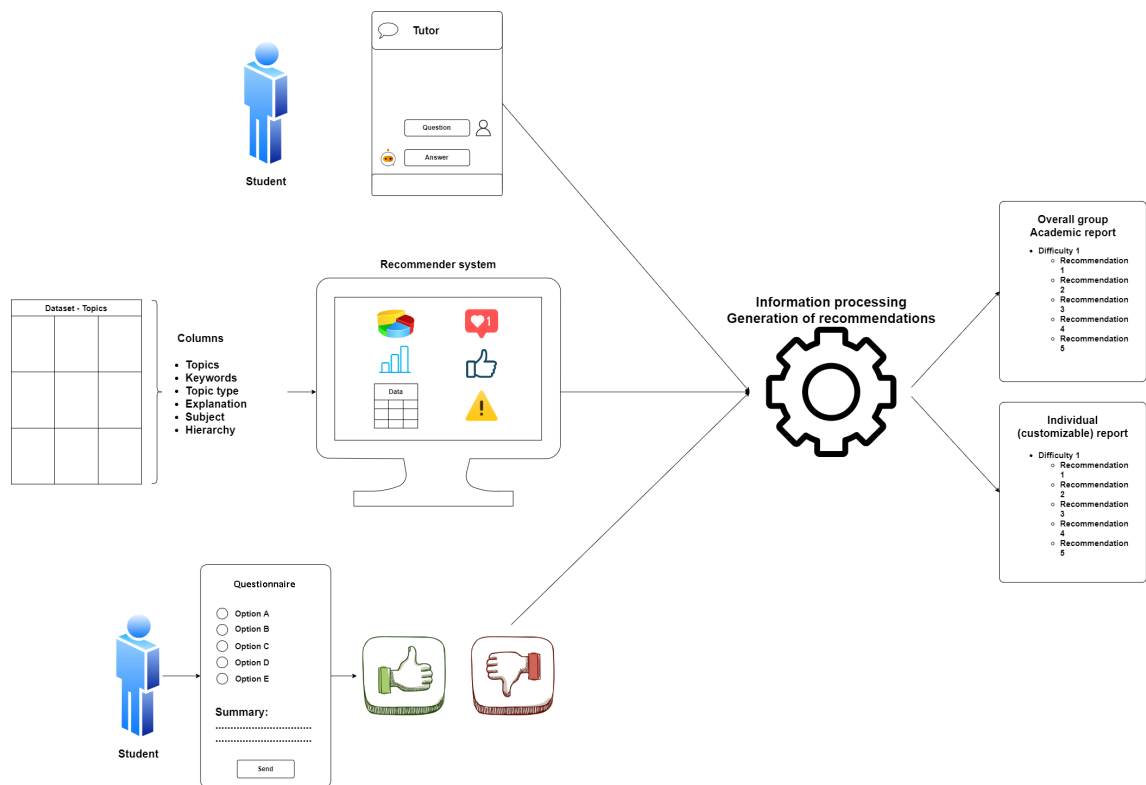


Figure 3: Integration of the Content-based Recommendation System to RetoñosApp.

and Tractinsky, 2006) by interacting with the web-based application. The questionnaire had an initial question of acceptance, guaranteeing the voluntary and uncoerced participation of the students in this pilot study. In addition, it had 5 demographic questions and 22 additional questions focused on measuring the user experience of the target population against the web-based platform (see Table 1).

Perceptions of usability and satisfaction in this study were calculated by applying the following metrics: (1) SUS (i.e., 1-5 System Usability Scales) (Grier et al., 2013), (2) NPS (i.e., 1-10 Net Promoter Scales) (Mandal, 2014), and (3) 1-10 Rating-Scales (Harpe, 2015). For this pilot study, we defined User-Experience as the relation (i.e., synergy) between the participants' perceptions of satisfaction and usability of RetoñosApp (see Table 2).

4 EXPERIMENTAL APPROACH

Each participant followed the experimentation process: (1) Introduction of the experimentation activity based on RetoñosApp. (2) Unsupervised testing phase of the developed application, providing a user manual as a guide for use. This phase lasted between 20 to 30 minutes, depending on the size of each group of CS1 and CS2. (3) Application of a questionnaire to evaluate their user experience on the web-based platform (see Table 1), once they interacted with it.

The user test was uncontrolled (i.e., unsupervised). Apart from the introduction around the experimentation at the beginning of each session, we provided each participant with the system guide of RetoñosApp (i.e., user manual) to test out each of the functionalities freely—they interacted with the platform in the way each participant preferred. Once they interacted with the web-based tool, the participants were given a questionnaire to evaluate their perception of user experience on RetoñosApp through questions regarding Usability (i.e., Functionality, Accessibility, Visual Design, Error Rate, Ease of Learning, Efficiency of Use), and Satisfaction. We asked in this questionnaire (1-5) System Usability Scales (SUS), (1-10) Rating Scales, and (1-5) Net Promoter Scales (NPS) (see Table 1).

5 FINDINGS AND RESULTS

We present our findings and results based on the experience reported by the participants who gathered the experimentation and evaluation of RetoñosApp—the

methodological process and instruments used, were described in sections 3 and 4.

To analyze the User Experience (UX) of RetoñosApp, we contemplated some calculations (see Table 2), based on the average results by the participants of the questions posed in the questionnaire designed for the experimentation phase (see Table 1).

We sought to reach or exceed a threshold for User Experience of 65% acceptance by the participants of this pilot study. This indicates that end-users (i.e., target population) perceive an average user experience of 6.5 out of 10, which we could interpret as a perception of an acceptable User Experience.

We present on Table 3 the results obtained on the perception's evaluation, based on the questionnaire presented on section 3 (see Table 1). This analysis helped us identify how effectively the participants perceived RetoñosApp based on their experience through the experiment. From these results (see Table 3), the participants perceived a User-Experience of 8.16 out of 10, considered outstanding, 1.65 points above the expected result (i.e., 6.51 was expected). The participants evaluated the system usability (i.e., by calculating the SUS) with a score of 8.15, which is 2.16 points higher than expected (i.e., 5.99 was expected). Also, 81.81% of the participants would probably recommend RetoñosApp (i.e., by calculating the NPS), 31.81% above what was expected.

Moreover, from the comments and perceptions of the participants, the customized reports (i.e., educational roadmap) that RetoñosApp provides, encourage academic continuity and promote a deepened participation of students in their education (Baque-Reyes and Portilla-Faican, 2021). Also, we found that RetoñosApp promotes new learning and teaching strategies through the on-time customized “learning routes” provided to students and instructors—strategies aligned with defined curricular and methodological structures.

6 DISCUSSION

From the preliminary results and comments received by the participants of this pilot study, RetoñosApp effectively provides a friendly and user-centered asynchronous assistance and enhancement to learning processes, and feedback on teaching processes. Moreover, we also found that the preliminary participants' comments about their web-based platform experience were positive and constructive, guiding us to improve the tool on their comments for future iterations.

Unlike traditional algorithms, artificial intelligence (AI) does not seek to be 100% assertive regard-

Table 1: Questionnaire for the Perception's evaluation.

Question	Variable	Scale: SUS ⁽¹⁾ , NPS ⁽²⁾ , RS ⁽³⁾
Q1: "I think I would like to use RetoñosApp more often in the course."	Usability	(1-5) SUS
Q2: "I found RetoñosApp to be unnecessarily complex to be used without prior instruction."	Usability	(1-5) SUS
Q3: "I thought RetoñosApp was easier to use than it really was."	Usability	(1-5) SUS
Q4: "I think I would need someone's technical support to be able to use RetoñosApp."	Usability	(1-5) SUS
Q5: "I found the features of RetoñosApp to be well integrated/developed according to what I would expect them to do."	Usability	(1-5) SUS
Q6: "I thought that there were too many inconsistencies, and contradictions in RetoñosApp in terms of the functionalities that I would expect the system to have in relation to its purpose."	Usability	(1-5) SUS
Q7: "I imagine that most people would learn to use RetoñosApp very quickly."	Usability	(1-5) SUS
Q8: "I found RetoñosApp very complicated to use."	Usability	(1-5) SUS
Q9: "I felt very confident or comfortable when using RetoñosApp."	Usability	(1-5) SUS
Q10: "I needed to learn a lot of things before I could use the RetoñosApp."	Usability	(1-5) SUS
Q11: "How precise, coherent and pertinent do you consider the functionalities provided by the application to be compared to the functionalities you expected?"	Usability	(1-10) RS
Q12: "How consistently do you think the graphic style (colors, shapes, images, and elements of the graphical user interface) of the application is concerning the institutional image of Universidad El Bosque?"	Usability	(1-10) RS
Q13: "How nicely do you think the application is in terms of its graphical user interface?"	Usability	(1-10) RS
Q14: "How complicated do you think it was for you to use the application?"	Usability	(1-10) RS
Q15: "How friendly do you think the application can be for any user?"	Usability	(1-10) RS
Q16: "How efficient do you think the application was regarding the number of errors it presented?"	Usability / Error rate	(1-10) RS
Q17: "How easy to learn to use do you think the application is?"	Usability	(1-10) RS
Q18: "How efficient do you think the application was during the ended user test?"	Usability	(1-10) RS
Q19: "How satisfied do you feel after using the application?"	Satisfac- tion	(1-10) RS
Q20: "How satisfied do you think you would be in the long term if the application were implemented as one more resource within the teaching and learning exercises of the subject?"	Satisfac- tion	(1-10) RS
Q21: "How much would you recommend your colleagues or friends use the application if there were modules for their respective interests, careers, subjects, or topics?"		(1-5) NPS
Q22: "How much would you recommend your instructor use the application as a complementary tool for the subject's methodology, fostering autonomous or independent study?"		(1-5) NPS

Notes: ⁽¹⁾SUS: System Usability Scale, ⁽²⁾NPS: Net Promoter Score, ⁽³⁾RS: Rating Scale

Table 2: Calculations of the contemplated metrics.

Metric	Calculation used
Usability (<i>U</i>)	$U = Avg(Avg.Rating + Avg.SUS + Avg.NPS) \quad (1)$ Avg.Rating corresponds to the average of questions 11-18, Avg.SUS corresponds to the average of questions 1-10, and Avg.NPS corresponds to the average of questions 21 and 22 (see Table 1).
User Satisfaction (<i>S</i>)	$S = Avg(Q19 + Q20) \quad (2)$ Q19 and Q20 correspond to questions 19 and 20, respectively (see Table 1).
User Experience (<i>UX</i>)	$UX = Avg(U + S) \quad (3)$ U and S correspond to the result of the metrics Usability and User Satisfaction, respectively. This calculation for User Experience (UX) is a preliminary approximation based on U and S calculated information, for this paper. Nevertheless, we are aware that UX contemplates usability (e.g., efficiency, perspicuity, dependability) and experience (e.g., originality, stimulation, interactivity) (Knijnenburg et al., 2012).

Table 3: Results of the perception's evaluation.

Metric	Expected Value	Score	Difference of the score vs. the minimum expected value
Usability [SUS]	8,15	2.14	5.99
Satisfaction [Rating Scales]	8,16	2.16	6.00
Recommendation [NPS]	81,81%	50%	31.81%
User-Experience (Usability & Satisfaction)	6.51	8.16	1.65

ing the results of its processes (Chen et al., 2020). Probabilistic factors determine the accuracy of AI models in CS education, since it learns from the context, environment, and users— all these elements are significantly malleable and volatile. Also, there is always a degree of uncertainty that affects the results when addressing teaching and learning strategies involving AI.

We evaluated the AI models part of RetoñosApp (i.e., Conversational Bot (NLP model) and Content-Based Recommendation System). As a result, the conversational bot had an accuracy of 82%, and the recommendation model had an accuracy of 83%. For these evaluation processes we used a preliminary N=40 training set. In the first place, both accuracies obtained by each model correspond accordingly as the content-based recommender model is naturally fed by the information collected by the platform (e.g., questionnaires and selections), and the conversational

bot (e.g., questions, enriched answers, interaction). Moreover, we consider both integrated models' accuracy reliable as it conforms to the valid range of accuracies between 63.1% and 89.3% as reported in the Teaching Academic Survival Skill (TASS) community (Díaz-Galiano et al., 2019).

We find that a web-based tool, such as RetoñosApp, is a considerable way for users (i.e., CS students or instructors) who require accessibility, analysis, complement, and support to ease their teaching and learning processes in a customizable way. This is similar to how the LEGO (McNamara et al., 1999) model works, based on the participants' particular needs, doubts, and difficulties. Furthermore, these results also guided us to reflect on the difficulties perceived (e.g., how to improve the GUI in each module to get the most of the web-based tool), helping us identify elements to enhance for future iterations of RetoñosApp. In further research, a deepened

analysis of the metrics that were calculated (see Table 2), will be addressed to complement the preliminary claims.

Therefore, we can preliminarily claim that RetoñosApp positively fosters the educational (i.e., teaching and learning) processes. The platform benefits both the teaching and learning processes of programming in CS. Moreover, unlike other available educational tools, RetoñosApp has the advantage that it can nurture the teaching and learning processes since it incorporates a conversational bot (i.e., virtual tutor) and a recommendation system (i.e., providing a customizable “educational roadmap” and frequent feedback). Further research on the tool will be conducted to support and complement all claims in this paper.

ACKNOWLEDGMENTS

The authors thank all the participants who voluntarily and actively collaborated in evaluating RetoñosApp. Also, the authors thank the Undergraduate Program of Systems Engineering instructors at Universidad El Bosque, Colombia, who willingly permitted this preliminary experimentation in their introductory programming courses in the early academic semesters.

REFERENCES

- Adamopoulou, E. and Moussiades, L. (2020). An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer.
- Baque-Reyes, G. R. and Portilla-Faican, G. I. (2021). El aprendizaje significativo como estrategia didáctica para la enseñanza-aprendizaje.
- Cano, P. A. O. and Alarcón, E. C. P. (2021). Recommendation systems in education: A review of recommendation mechanisms in e-learning environments. *Revista Ingenierías Universidad de Medellín*, 20(38):147–158.
- Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278.
- Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.
- Deacon, J. (2009). Model-view-controller (mvc) architecture. *Online*[Citado em: 10 de março de 2006.] <http://www.jdl.co.uk/briefings/MVC.pdf>, 28.
- Díaz-Galiano, M. C., García-Cumbreras, M. Á., García-Vega, M., Gutiérrez, Y., Cámara, E. M., Piad-Morffis, A., and Villena-Román, J. (2019). Tass 2018: The strength of deep learning in language understanding tasks. *Procesamiento del Lenguaje Natural*, 62:77–84.
- Grier, R. A., Bangor, A., Kortum, P., and Peres, S. C. (2013). The system usability scale: Beyond standard usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 187–191. SAGE Publications Sage CA: Los Angeles, CA.
- Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.
- Harris, S. C. and Kumar, V. (2018). Identifying student difficulty in a digital learning environment. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 199–201. IEEE.
- Hassenzahl, M. and Tractinsky, N. (2006). User experience—a research agenda. *Behaviour & information technology*, 25(2):91–97.
- Klein, M. H., Kazman, R., Bass, L., Carriere, J., Barbacci, M., and Lipson, H. (1999). Attribute-based architecture styles. In *Working Conference on Software Architecture*, pages 225–243. Springer.
- Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. (2012). Explaining the user experience of recommender systems. *User modeling and user-adapted interaction*, 22(4):441–504.
- Mandal, P. C. (2014). Net promoter score: a conceptual analysis. *International Journal of Management Concepts and Philosophy*, 8(4):209–219.
- Martin, R. C., Grenning, J., Brown, S., Henney, K., and Gorman, J. (2018). *Clean architecture: a craftsman’s guide to software structure and design*. Number s 31. Prentice Hall.
- McNamara, S., Cyr, M., Rogers, C., and Bratzel, B. (1999). Lego brick sculptures and robotics in education. In *1999 Annual Conference*, pages 4–369.
- Moreno, J. J., Bolaños, L. P., and Navia, M. A. (2010). Exploración de modelos y estándares de calidad para el producto software. *Revista UIS Ingenierías*, 9(1):39–53.
- Mutiawani, V. et al. (2014). Developing e-learning application specifically designed for learning introductory programming. In *2014 International Conference on Information Technology Systems and Innovation (IC-ITSI)*, pages 126–129. IEEE.
- Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer.
- Richards, M. (2015). *Software architecture patterns*, volume 4. O’Reilly Media, Incorporated 1005 Gravenstein Highway North, Sebastopol, CA . . .
- Saldarriaga-Zambrano, P. J., Bravo-Cedeño, G. d. R., and Lóor-Rivadeneira, M. R. (2016). La teoría constructivista de jean piaget y su significación para la pedagogía contemporánea. *Dominio de las Ciencias*, 2(3 Especial):127–137.
- Sommerville, I. and Torres, J. A. D. (2011). *Ingeniería de software*. Addison-Wesley Pearson Educación, México, 9 edition.
- Wellnhammer, N., Dolata, M., Steigler, S., and Schwabe, G. (2020). Studying with the help of digital tutors: design aspects of conversational agents that influence the learning process.

Popular Music Production Trend Analysis and Prediction Research

Mingyue Zheng* and Jiarui Jin*
Keystone Academy in Beijing, Beijing, China
dlencyzheng@163.com

Keywords: Pattern Recognition, Machine Learning, GAN, Popular Music Production, Trend Analysis and Prediction.

Abstract: Music is the most popular art form performed and listened to by billions of people every day. There are many kinds of music, such as pop music, classical music, jazz music, folk music, etc., among which pop music is the most widely listened type of music. When pop music creators are willing to improve their own music production level, they usually get inspiration by listening to and learning a lot of other people's songs. This is a time-consuming job and requires the person to have strong judgment and prediction ability, which requires years of experience accumulation. Using big data and artificial intelligence technology to analyse and speculate on the trend of popular music in a certain area over a period of time can help creators quickly find valuable data, such as the most popular song melody, changes in the corresponding artist's style. Music samples are long and complex and are composed of multi-channel mixed signals of different vocals and musical instruments. Therefore, the popularity of music is a difficult pattern recognition problem, and it has great research and application value. This project will create a GAN-based machine learning model for finding the multi-dimensional quantification of the most popular pop music features in different regions in the current time period from multiple music samples, which can be used for pop music creators to discover popular trends and improve their own music level and popularity.

1 INTRODUCTION

Adolescents in urban and rural areas around the world, whether they are in school or not, spend a lot of time listening to music, and more time listening to music than they do playing with friends outside of school. This phenomenon is easy to verify, observe teenagers wearing headphones in the subway or walking, and it is obvious that music is very important to teenagers, music changes their emotions, dominates their conversations and determines them. most of the ideas. Digital music and online streaming have become very popular with young people in recent years, with around 40 million songs in digital libraries and more than 100,000 songs released annually on digital music services.

Adolescents enjoy music for different motivations and reasons, ranging from a desire to obtain information about the adult world, or to foster friendships in a group of friends, or to avoid socializing through music. For example, rock music is everywhere, especially the online platforms that provide rock music continue to grow and develop.

Nearly 30% of 16-19 year olds in the United States consume rock music, and the rock music genre is very popular among teens around the world. At first, young people embraced rock music because it could represent a rebellion against social norms, because the society at that time emphasized conformity, and rock music seemed to give teenagers another way of life. In fact, rock music has always been created for the same purpose, it provides teens with the opportunity to express themselves to society. Rock music is very dynamic and provides the fun that young people want to experience.

Music is so powerful that teens have become a key segment of the target market. The face of music is constantly changing, and music producers must stay on top of the latest trends. Teens are an important part of music consumers, so it is important to understand their tastes and preferences. Not just for teens, but for adults, the fusion of YouTube, iTunes, and Grey's Anatomy has led to a new approach to music marketing: Advertisers purposefully speculate on songs and predict their popularity. In other words, brands want their ad to be your first exposure to a

* Both authors contributed equally to this work

specific song that a user ends up downloading. This is a big departure from past advertising strategies.

But for young music creators, how to correctly grasp the trend of music is a very difficult task. For young talented writers, it is very possible to devote all their efforts and energy to creating music that they are intoxicated with, but due to the inability to grasp the trend and direction of music development, they will not be able to get due for a long period of time. attention and appreciation, long-term in the low ebb stage of the career. The research goal of this paper is to use machine learning to find and quantify the style trends of these music data by analysing the music pieces that are popular with listeners on various charts at the current or a period of time, and use adversarial neural networks to generate music pieces that are similar to popular trends for the reference and rapid improvement of young music start-ups, while discovering popular trends and combining with their own musical talents to create musical works that are more popular and loved by a wide range of audiences.

2 RELATED WORKS

2.1 Introduction to Generative Adversarial Networks

GAN (Generative Adversarial Networks) is an unsupervised deep learning model whose kernel philosophy is taken from game theory. The most special property of the GAN model is that its training process simultaneously trains two models: a generative model and a discriminative model. In the process of playing against each other, the effects of the two models can promote and improve each other. GAN consists of two models, called Generative Model and Discriminative Model. Fig.1 shows the structure diagram of the GAN model.

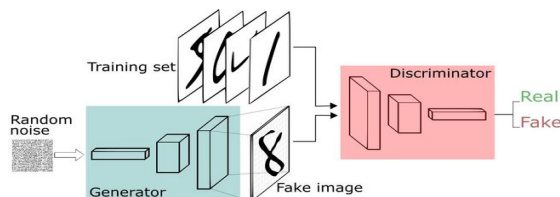


Figure 1: The Structure diagram of the GAN model.

A generative model attempts to produce a "fake" data that fits a certain probability distribution. For this model, it takes an arbitrary noise data and generates a "fake data as close as possible to the real data" based on the noise. However, it cannot directly know

whether its data is "close to the real", but needs a discriminative model to determine the generation effect of the generative network. A discriminative model is used to accept a set of data as input, and it will output whether this set of data is real data. In other words, the job of the discriminative model is to determine whether the input it accepts is from the dummy data of the generative model or the real data from the dataset.

The training process of the GAN model is a process in which the generative model and the discriminative model are mutually trained and confronted. The specific process is as follows:

- A generative model takes a noisy data and produces a set of generative images.
- Train a discriminative model using generated and real images so that it can decide whether an image is a generated image or a real image.
- Feedback the conclusions of the discriminative model to the generative model to guide it to generate more realistic images.
- Repeat the above process

In the actual training process, the above second step (that is, the process of training the discriminant model) will be repeated k times before entering the third step, where the k value is a hyperparameter. With the progress of training, the discrimination ability of the discriminant model will become stronger and stronger, which will guide the generated pictures to be more realistic, and then the more realistic generated pictures will test the discrimination ability of the discriminant model with higher requirements. Through this mode, both the generative model and the discriminative model will continuously optimize each other in this adversarial process. The final ideal situation is that the samples generated by the generative model are the same as the real samples, the discriminant model can no longer distinguish whether the samples are real or not, and the judgment result of the authenticity of the picture is always 50%.

When the GAN model was proposed in 2014, there was no requirement for the specific forms of these two models. Any algorithm that can complete sample generation and sample discrimination can be applied to GAN as a generative model or a discrimination model. However, with the continuous improvement of the structure and algorithm of GAN, the continuous clarification of the application scope, and the continuous exploration of application prospects, in the current practical operation, the generative model and the discrimination model are usually deep learning models. The DCGAN (deep

convolutional generative adversarial networks) model integrates CNN and the original GAN model, which improves the generation effect of GAN, and the discriminative ability of the discriminative model itself is also better than other unsupervised methods.

2.2 Music-based Generative Adversarial Networks

Artificial intelligence is being used in art and creativity to generate music, images and text. In recent years, computer-generated music has gradually become the focus of attention. "Hello World" is SKYGGE's first album to involve artificial intelligence technology in the creative process. By introducing AI technology into music production, it creates unusual harmony and melodies, opening a new era of pop music and AI. The author believes that the future trend of AI in music production and other related industries is very clear, and it may become mainstream not only in music production, but also in social media platforms. SKYGGE's "Hello World" is the product of these new forms and an important turning point in the understanding of human-machine collaboration.

As more and more direct applications of deep learning are discovered and developed, music generation is becoming one of the areas of research interest, and many large tech companies have recently begun to investigate related research areas. Google's Magenta project has been released to help programmers and artist's experiment. Extracting useful information from tunes analysed in the past, the created model learned how to make key changes and melody loops. Somewhat like the well-known AlphaGo, AlphaGo's approach, called reinforcement learning, allows animals to pick up a new skill. Researchers from Google's AI research group believe that this reinforcement learning can be used for more complex affairs and works of art, and the resulting creativity can be used in many areas of life. Researchers believe that in the future, artificial intelligence can generate more complex artworks than today.

With the development of deep learning music generation, new challenges are starting to emerge. There are some problems with deep learning-generated music, such as control over tonal consistency and rhythm, musical structure, interactivity, and creativity, which means that the generated music tends to imitate training samples rather than creating new musical pieces. The method is through sampling adjustment, audio style transfer, and reinforcement learning. Some researchers have

carried out research on the effect of music generation on several deep learning network architectures: Generative Adversarial Network (GAN), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The conclusion is that GAN outperforms all other network architectures because it uses RNN and CNN effectively, and of course the generated music is still incomparable to real music.

Another problem is that the originality of music is closely related to the economics of copyright. And creative adversarial neural networks based on deep learning usually do not have such a dilemma. Many studies have used generative adversarial networks (GANs) to generate symbolic music in the form of piano-rolls. But the generated music is often weak in temporal feature extraction, and the music sounds unnatural. DMB-GAN is a new structure for music generation GANs, which makes three improvements to existing GAN-based music generators: First, DMB-GAN implements a self-attention mechanism to extract space from the given data and temporal characteristics; secondly, it has a multi-track architecture that allows for multi-instrumental music and enhances harmony between different tracks. Third, by implementing switchable normalization, GMB-GAN supports more stable music generation. Practice has proved that GMB-GAN can potentially handle multi-track music generation with higher stability and natural coherence.

GANs have been shown to be an effective method for training deep generative neural networks, among which the continuous recurrent neural network C-RNN-GAN model trained by generative adversarial networks uses GANs to train recurrent neural networks on continuous sequence data. C-RNN-GAN uses specially defined generators and discriminators, aiming to reach a Nash equilibrium in a zero-sum game of training, where the discriminator cannot distinguish generated data from real data. The model also employs feature matching to encourage greater variance in the generator and reduce overfitting. By comparing the C-RNN-GAN model with the baseline model, it is found that the music generated by C-RNN-GAN is increasingly complex and achieves higher scores on pitch span, intensity span, and number of unique tones. On the other hand, the baseline model shows a relatively low pitch span, indicating that it uses less pitch variation. Furthermore, the music generated by the C-RNN-GAN model has the same intensity span as real music.

3 PROPOSED MUSIC SYNTHESIS BASED ON GAN NETWORK

At present, GAN-based generative networks have made considerable progress in the fields of pictures, videos, and speech, but the work in these areas is different from the music generation we study. The closest form to music synthesis is probably speech synthesis (TTS). The difficulty of music synthesis is: firstly, music is based on time series; secondly, notes are under the control of chords, arpeggios, melody, polyphony, etc.; at the same time, a song is composed of multiple instruments and multiple tracks.

Music has its own time-based architecture, as shown in Figure 2. Like many audio signals, the various parts of music are highly time-dependent. But unlike ordinary audio, which only has a direct correspondence with time, concerts involve paragraphs, phrases, measures, and beats. Not only does the time develop from front to back, but there are also cyclic changes between phrases, and between paragraphs. echo relationship, etc.

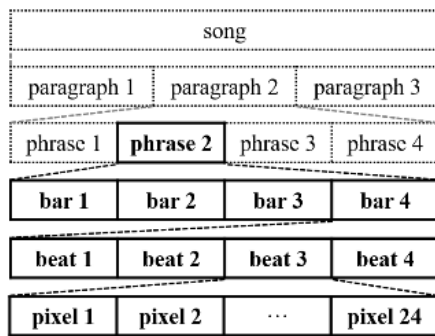


Figure 2: Time frame of music.

Music is multi-instrument and multi-track. Modern orchestras often have 4 parts: brass, strings, woodwinds, and percussion. Rock bands often use bass, drums, guitar, and possibly vocals. Music theory requires that these elements unfold in harmony and in counterpoint over time.

Control of notes in chords, arpeggios, melody, polyphony and other rules. Therefore, if you want to emulate the method of natural language processing (NLP), it is not advisable to generate a certain note for each beat position. Because this can only generate monophonic music, not polyphonic music. There are multiple notes at the same time, chords are stipulated by artificial rules, and the melody has no clear semantics, which has become a breakthrough in machine learning. These are all problems that are

difficult to solve by modifying the NLP method for music synthesis.

This paper refers to the principle and structure of the MuseGan model, and considers building our generative network from the difficulties mentioned above as a starting point:

(1) Build correlations between tracks. Mainly tried the following composition methods:

Jamming Model: Each track has its own set of generators and discriminators, and an independent latent space variable Z .

Composer Model: A global set of generators and discriminators that share Z to generate all tracks.

Neither approach is quite as expected. We expect the whole composition to be harmonious, so we need the discriminator to look at the effect of the generated music uniformly for all tracks, which the Jamming Model does not do. Each track has its own characteristics, and the shared latent space variable Z cannot achieve such an effect, so the Composer Model also has shortcomings, so the Hybrid model is a better choice.

Hybrid Model: Mixing the above two modes, each track has an independent generator, accepting the input vector formed by the independent latent space variable Z_i and the global latent space variable Z , and sharing a discriminator to generate the entire music. Compared with the Composer Model, the blending mode is more flexible and can use different parameters (such as the number of layers, convolution kernel size, etc.) in the G model to combine the independent generation of audio tracks and global harmony.

(2) Build timing dependencies

The purpose of the structure mentioned above is how to generate a single syllable in different audio tracks, and there is still some cyclic or progressive relationship between syllables and syllables, which is processed by using a hidden variable to expand. By dividing the generator into two sub-networks G_{temp} and G_{bar} , z is first mapped into a sequence of latent space vectors $G_{temp}(z)$, which can carry some timing information (the relationship of each subsection in time), and then $G_{temp}(z)$ is sent to the generator G_{bar} to serially generate piano-roll.

(3) Improve the time frame

Finally, we combine the previous designs to form a multi-track music generator that can build a music structure. The input Z of the model consists of 4 parts:

The top row is the global time correlation vector between tracks. Global is relative to a single track, which refers to the fusion of all tracks, and considers

the relationship between different bars, that is, time correlation. This vector mainly adjusts the relationship between each instrument and each track, and is closely related to chords.

The second row is a global time-independent vector between orbits. This vector sets the tone for the style of the entire composition considering all the tracks, but not the composition's relationship to time. The third row is a single time-independent vector within the orbit. This vector considers factors that represent time independent of a single track, such as tempo.

The last row is the correlation vector of individual events within the orbit. This vector considers factors that represent the time-dependent nature of a single track, such as the melody of an individual instrument.

4 EXPERIMENTS AND RESULT ANALYSIS

4.1 Dataset

We use the Lakh Pianoroll dataset (LPD) for experiments. The Lakh Pianoroll dataset is a collection of 174,154 multi-track Piano-rolls derived from the Lakh MIDI dataset (LMD), a commonly used electronic music dataset, as shown in the Fig.3 for the music contained in the Lakh MIDI dataset style.

Comparison of different subsets

subset	size	matched to MSD	one entry per song
full	174,154	X	X
matched	115,160	O	X
cleansed	21,425	O	O

Comparison of different versions

version	number of tracks	tracks
LPD	-	-
LPD-5	5	Drums, Piano, Guitar, Bass, Strings
LPD-17	17	Drums, Piano, Chromatic Percussion, Organ, Guitar, Bass, Strings, Ensemble, Brass, Reed, Pipe, Synth Lead, Synth Pad, Synth Effects, Ethnic, Percussive, Sound Effects

Figure 3: Lakh MIDI dataset.

The data representation uses the multiple-track piano-roll representation. piano-roll is a score-table matrix of binary values representing notes at different time steps. Multiple-track piano-roll is defined as a set of different tracks. piano-rolls. One has M tracks, and each track has R time nodes. The matrix size of each X is fixed, which is good for CNN training features.

4.2 Experimental Results and Analysis

The experiment randomly selects 10 music data in the dataset. Using the GAN structure based on the

mixture model mentioned above, the code execution steps for generating new music data are shown in Fig.4.

```
[aurora-c5 music]$ ls
input run.py src
[aurora-c5 music]$ ls input
midi mp3 npz
[aurora-c5 music]$ ls input/mp3
10.mp3 1.mp3 2.mp3 3.mp3 4.mp3 5.mp3 6.mp3 7.mp3 8.mp3 9.mp3
[aurora-c5 music]$ python run.py input/npz
success!
you can change the midi file into MP3 manually
[aurora-c5 music]$ ls
input output.mid results run.py src
[aurora-c5 music]$ lets listen to the output file
```

Figure 4: Steps to perform to generate new music.

Fig.5(a) and Fig.5(b) shows the waveform comparison between the original music data and the generated music data. The generated music is subjectively listened to by the human ear and has unnatural traces that are difficult to distinguish. By using tools for comparison, the results are shown in Fig.6(a) and Fig.6(b). There are certain differences between the output generated music data and the original music data.

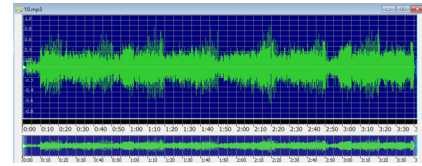


Figure 5(a): Music Input Sample Waveform.

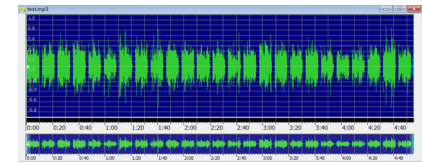


Figure 5(b): Music Output Sample Waveform.

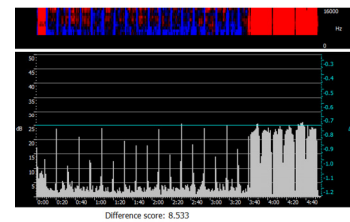


Figure 6(a): Comparison of input and output.

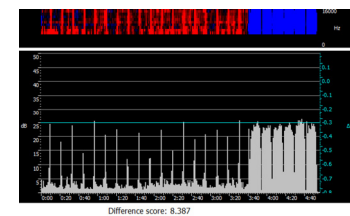


Figure 6(b): Comparison of another pair of input and output.

5 CONCLUSIONS

Music production is a skill that is very dependent on talent and ability. Traditionally, the level of music production is generally directly related to the music creator's music foundation and innovation ability. In order to further enhance the ability of new music creators to grasp the trend of popular music and gain the audience's continuous attention, this paper builds a music generation model based on adversarial neural network, which is characterized by learning and simulating multiple popular music pieces, extract the current popular music features, generate basic music data with popular style, help pop music creators to discover the trend of popular music as soon as possible, and further improve the popularity of their own music.

REFERENCES

- Parker, E. C. (2020). Adolescents on music. *Oxford University Press*.
- Campbell, P. S., Connell, C., & Beegle, A. (2007). Adolescents' expressed meanings of music in and out of school. *Journal of Research in Music Education*, 55(3), 220-236.
- Freed, E. N. (1990). An exploratory study of the meaning of music to adolescents. *Massachusetts School of Professional Psychology*.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1), 53-65.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Liu, M. Y., & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in neural information processing systems*, 29.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. *In Proceedings of the IEEE international conference on computer vision* (pp. 2794-2802).
- Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.
- Guan, F., Yu, C., & Yang, S. (2019, July). A gan model with self-attention mechanism to generate multi-instruments symbolic music. *In 2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.
- Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*.
- Kulkarni, R., Gaikwad, R., Sugandhi, R., Kulkarni, P., & Kone, S. (2019). Survey on deep learning in music using GAN. *International Journal of Engineering Research & Technology*, 8(9), 646-648.
- Avdeeff, M. (2019, October). Artificial intelligence & popular music: SKYGGE, flow machines, and the audio uncanny valley. *In Arts* (Vol. 8, No. 4, p. 130). MDPI.
- Qin, T., Liu, Y., Wang, C., Yang, T., & He, C. (2017, July). From online to offline: Charactering user's online music listening behavior for efficacious offline radio program arrangement. *In 2017 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1219-1224). IEEE.
- Wang, Z., Ye, C., & Wang, W. (2019, June). Music Trend Forecast Based on LSTM. *In 2019 4th International Conference on Computational Intelligence and Applications (ICCIA)* (pp. 30-35). IEEE.
- Dong, H. W., Hsiao, W. Y., Yang, L. C., & Yang, Y. H. (2018, April). Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- <http://dy.163.com/v2/article/detail/D896DT3L0511AQHO.html>

Fynex: Work in Progress on a Web-based Approach That Implements a Hybrid Recommendation System for Preventing and Treating Diseases based on Eating Disorders

Bríter Andrés González-Daza^a and Miguel Alfonso Feijóo-García^b

*Program of Systems Engineering, Universidad El Bosque, Bogotá, Colombia
{bgonzalezd, mfeijoog}@unbosque.edu.co*

Keywords: Usability, Healthcare Application, Recommendation System, Machine Learning.

Abstract: Diseases based on eating disorders have been considered a health issue worldwide. It is worrying due to the increased complexity of treatments for diseases such as Diabetes, Hypertension, Obesity, and Anorexia. We present Fynex as a web-based expert system that implements a hybrid recommendation system that supports Healthcare Professionals and Patients with recommendations on nutritional plans and physical activity. It provides functionalities to have detailed follow-up and control regarding the evolution of a particular treatment regarding these diseases. This solution poses a decision-making system for Healthcare Professionals and Patients, fostering the medical processes (i.e., treatments, progress, and decisions) in an integrated and systemic way. We evaluated this tool following a pilot study based on pre-post experimentation, and reported the findings and results considering participants' performance and perceptions, basing our analysis on (0-5) Likert scales, open-ended and YN responses, regarding their experience interacting with Fynex— analyzing the users' perceptions on satisfaction and usability on the web-based application. Our preliminary findings suggest that Fynex is effectively a friendly user-centered approach that successfully increases medical processes, healthcare control, and treatment satisfaction.

1 INTRODUCTION

The worldwide population (i.e., the youth population) usually does not have correct eating habits, which increases the chances of being diagnosed with a disease based on eating disorders (Candela, 2016). These diseases based on eating disorders such as Diabetes, Hypertension, Obesity, and Anorexia are considered a global problem, as this could result in possible cardiac complications, glucose intolerance, and insulin resistance, among others (Thomas et al., 2019). Moreover, medical controls are not generally customized, monitoring of these diseases is limited, and the lack of communication with healthcare professionals (HcP) could become critical beyond the South American context.

There are numerous solutions with scaffolds for health applications in this context. For example, My-Diabetes.diet (García et al., 2001) or DialBetics (Waki et al., 2014), have a similar approach to treating this type of diseases. Furthermore, the research work “In-

ternet of Things based on electronic and mobile health systems for blood glucose continuous monitoring and management” (Barata et al., 2019), applies Internet of Things (IoT) on glucose monitoring to be measured in real-time to provide patients control over their signs. However, these solutions restrict the monitoring of the treatment through single applications, focusing either on the patient or HcP.

Hybrid Recommender Systems combine different techniques (i.e., recommendation systems) to produce outputs to complete or complement their best features and make better conjoint recommendations (Seth and Sharaff, 2022). Due to the hybrid technique's usefulness, many recommendation systems have technically integrated it (e.g., implementation of solutions in e-commerce). However, unfortunately, there is a lack of hybrid recommendation systems focused on medicine, with only 5% adopting this approach (Danilova and Ponomarev, 2017). Recommendation systems focused on medicine use different methods, such as K-Means and association rules. For example, supervised models define which drugs have helped certain patients with specific symptoms— the supervised model is one of the most common methods

^a  <https://orcid.org/0000-0002-9587-205X>

^b  <https://orcid.org/0000-0001-5648-9966>

applied to recommendation systems. However, they usually point to multiple diseases without delving into them (Stark et al., 2019).

We present Fynex as a friendly and user-centered web-based application, that supports HcP and patients with treatments regarding diseases based on eating disorders. It has a hybrid recommendation system, a messaging chat, and a visual system to visualize the evolution of a patient, among others. Furthermore, it integrates Artificial Intelligence (AI), pretending to support medical on-time decision-making processes. Also, it focuses on customized treatment follow-up, which both HcP and patients could use.

This healthcare-based solution seeks to increase the satisfaction of both HcP and patients regarding medical treatments and decision-making processes. Hence, this paper presents Fynex and the result of our first pilot study using the web-based tool. We evaluated the participants' perception of satisfaction before and after using the application through questionnaires to analyze the behavior and impact of the solution preliminary—we followed a pre-post experimentation approach (Miller et al., 2020). We also describe the pros and cons of this pilot. We report the preliminary findings and results of this pilot study, pretending to answer and analyze how the satisfaction of Patients and HcP is impacted regarding the treatment of diseases based on eating disorders, when supported by an intelligent system based on a hybrid recommendation system.

Our work contributes to Healthcare Applications and User-Centered Interaction literature. We evaluate Fynex as an information and communication technology (ICT) for healthcare and medicine, customizing HcP and Patient's experience to support decision-making processes (i.e., home care monitoring, or diagnostic support) regarding treatments, and on-time medical control. This approach uses the benefits of autonomous, e-health-based, and customized medical processes addressing diseases based on eating disorders.

2 FYNEX: DESCRIPTION

We introduce a web-based application called Fynex that provides a friendly, user-centered, and interactive Graphical User Interface (GUI) to support HcP and patients in medical processes regarding eating disorders. Fynex is presented to end-users who are particularly involved in the treatment of one of the following preliminary diseases: (1) Diabetes, (2) Hypertension, (3) Obesity, and (4) Anorexia Nervosa. Therefore, the platform provides features and functionalities to im-

prove the satisfaction perceived by these actors (i.e., HcP and patients), by providing a visually attractive GUI that eases the user's interaction on the application, getting the most out of it. It implements a hybrid recommender system (Çano and Morisio, 2017) to present possible nutrition and exercise plans for the patient to follow and apply them in their life (See Figure 1). Besides, the web-based tool presents a monitoring system to keep track of the patient's evolution throughout time. This resulting information is helpful to the patient as to its HcP. Shneiderman's mantra (Munzner, 2014) presents how a tool is intended to manage the information visually through: (1) overview first, (2) zoom and filter, and (3) details on-demand, as required by a user (i.e., HcP and patients). Hence, Fynex provides a dashboard with health-based variables related to the patient's treatment, a system to upload and download test results, a messaging system between the HcP and its patients, and a visual representation of the similarities between the patients, among others (See Figure 2).

Fynex is a monolithic application using Django as its web framework, integrating a Model-View-Template architecture (MVT) (Rull et al., 2009), and Angular JS (Green and Seshadri, 2013) as a front-end framework to improve the dynamism and quality of the web application. Additionally, the application integrates third-party tools such as Scikit-learn (Kramer, 2016), Redis (Carlson, 2013), and IBM Cloud's Object Storage service (Sampé et al., 2018), to support the preprocessing of the information and the development of the web-based tool's management, as of its hybrid recommender system (Çano and Morisio, 2017). The data processing for the information system (i.e., Fynex) supports the recommendation models' implementation to enhance its results and reinforce further decision-making processes of a user. Fynex implements an ETL (Vassiliadis and Simitsis, 2009) to process the information from the Fynex database and be able to use the already trained Machine Learning (ML) models. Once the data is transformed, and loaded, we used a hybridization method to obtain a final recommendation, which we present to the user through the application—memory-based recommendation model, content-based recommendation model, and the ML model-based recommendation model (see Figure 3).

3 DATA ACQUISITION

The participants who gathered the experimentation voluntarily had roles of HcP and Patients. In this pilot study, we had a total of 14 participants (N=14)—50%

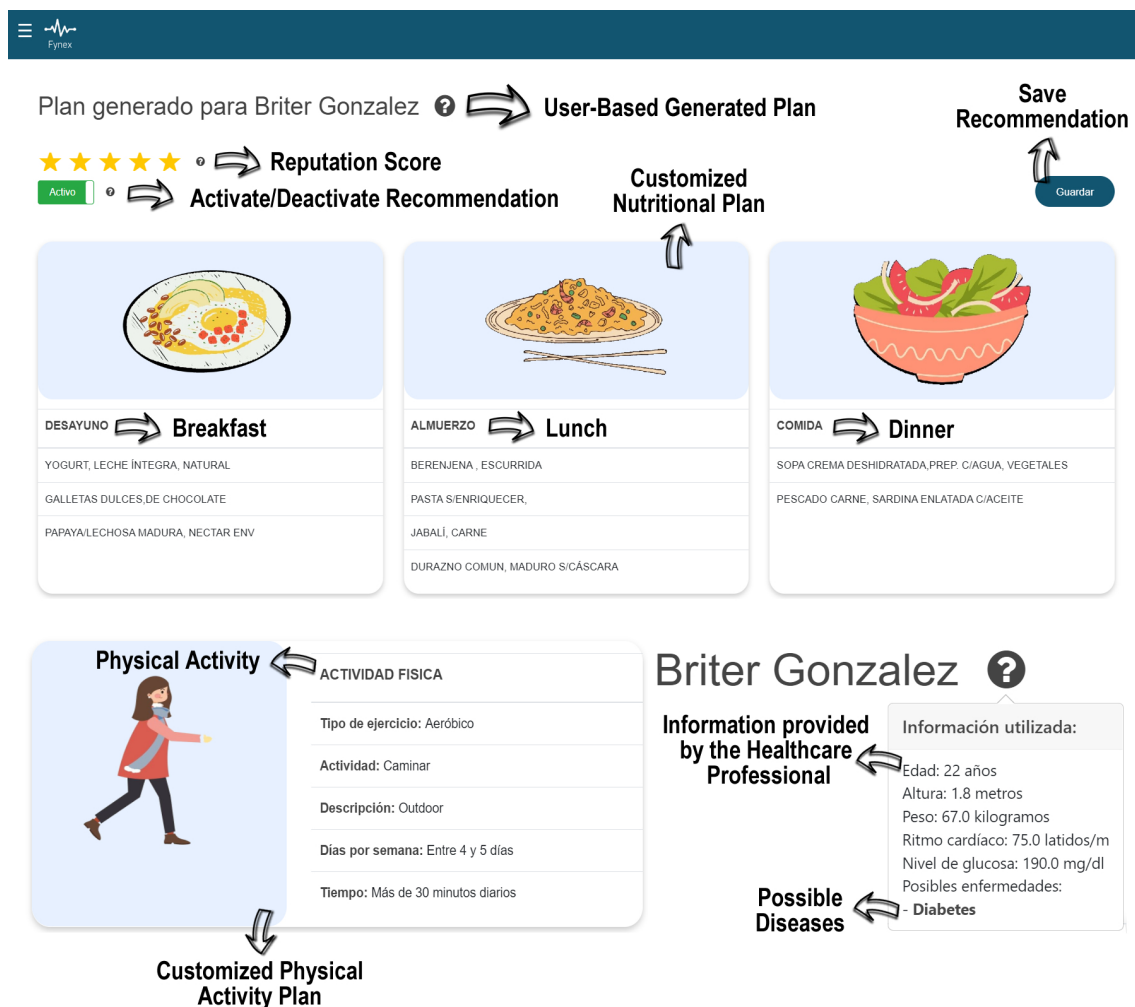


Figure 1: Features/Functionalities of Fynex - Part 1.

(n=7) HcP, and 50% (n=7) patients. We had 64.29% (n=9) male participants (i.e., 85.7% (n=6) HcP, and 42.9% (n=3) Patients), and 35.1% (n=5) female participants (i.e., 14.3% (n=1) HcP, and 57.1% (n=4) Patients).

Each participant performed specific tasks in the application. At the end of the tasks, the participants in the experimentation phase had to fill out a questionnaire that contained questions regarding the application and how much they recommended Fynex.

We evaluated the tool's effectiveness, usability, and acceptance, considering the participant's perceptions when using Fynex. We followed pre-post experimentation (Miller et al., 2020) based on two questionnaires [quantitative— 0-5 Likert scales (Joshi et al., 2015), and qualitative— Y/N and Open-Ended Questions], which were responded before and after interacting with the web-based tool (See Table 1). To the evaluation of usability, we applied an ad-

ditional questionnaire based on SUS— System Usability Scale [quantitative— 1-5 scales] (Grier et al., 2013). Moreover, we asked the participants to answer the NPS— Net Promoter Score [quantitative— Percentage from 0-100%] (Mandal, 2014) (See Table 2).

4 EXPERIMENTAL APPROACH

We sought to understand the behavior change of the satisfaction perceived by HcP and Patients, before and after the implementation of Fynex. To understand the context involved, we conducted our research on eating disorders and the diseases they might cause (i.e., Diabetes, Hypertension, Obesity, and Anorexia).

We considered in our approach the Biopsychosocial and Cultural Model (BPSCM) (Cruz and Buitrago, 2017) to understand the context based on



Figure 2: Features/Functionalities of Fynex - Part 2.

the identification of (1) artifacts, (2) means, (3) habits, and (4) beliefs involved, from each participant's perspective (i.e., HcP and Patients), helping in the understanding and enhancement of complex needs— context based on medical processes. Then, we defined how these elements might change if the actors used Fynex, so we can verify if those changes contribute to fulfilling the project's purpose.

Once the hybrid recommendation system of Fynex was developed, we defined three phases for testing and evaluating the model. The first phase was focused on evaluating the recommendation model, using metrics like the ROC curve (de Ullibarri Galparsoro and Fernández, 1998), F1-Score (Lipton et al., 2014), and results of simulations, among others. In the second phase, each participant (i.e., HcP and Patients) was asked to answer how the experience and

satisfaction changed (e.g., treatment, support) when using the tool. We calculated the percentage of the increment or decrement of perceived satisfaction when interacting with the tool based on the ACSI (Fornell et al., 1996) (i.e., American Customer Satisfaction Index), as well as the Probit and Logit (Chen and Tsurumi, 2010)— regression model based on independent variables such as age and time with treatment. Moreover, in the third phase, we used Loop11 (Bustamante, 2010) to establish the tasks to perform and to obtain certain metrics (i.e., percentage of completed tasks, usability calculated by the SUS (Grier et al., 2013)— System Usability Scale, and the acceptance calculated by the NPS (Mandal, 2014)— Net Promoter Score).

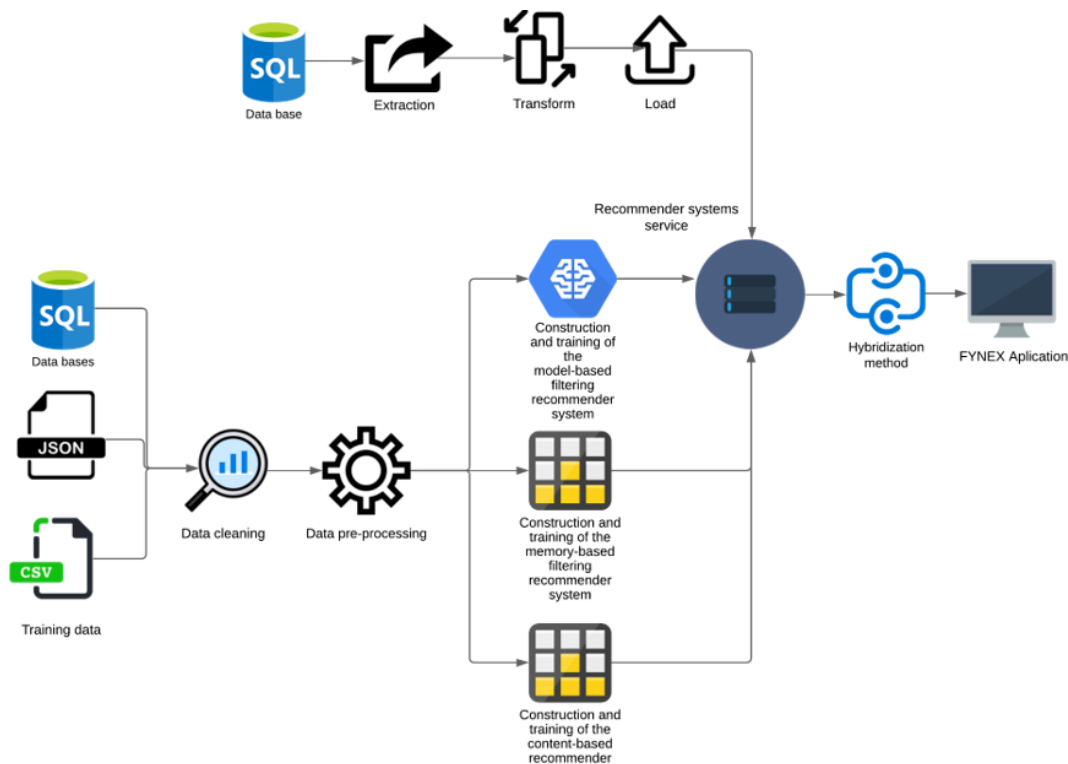


Figure 3: Concept design of Fynex.

5 FINDINGS AND RESULTS

We present our findings and results based on the participants' experience on Fynex reported, following the experimental methodology—the methodological process and instruments used, were described in sections 3 and 4.

In the first place, we evaluated the predictive AI models of Fynex, calculating their precision, accuracy, sensitivity, and the F1-Score indicators. In addition, we complemented each model's confusion matrix with the calculation of the ROC curve, which allowed knowing the AUC (i.e., Area Under the Curve) (Fawcett, 2006) to know the probability of a model detecting the disease in a patient in the context—diseases regarding eating disorders.

The diabetes predictive model offered an 85.7% probability of detecting the disease in patients. The predictive model of Arterial Hypertension offered a probability of 83.2% of detecting the disease in patients.

Furthermore, over 25 different recommendations provided by the hybrid recommender system of Fynex for this pilot study, we confronted the results against the ideal nutritional plans for each disease. We considered ideal nutritional plans considering the litera-

ture and the particular context—diseases regarding eating disorders. Patients treating Diabetes typically consume 15% to 20% protein, 45% to 50% carbohydrate, and 20% to 35% fat (Gray and Threlkeld, 2015). To treat Arterial Hypertension, there is a standard diet known as DASH (Dietary Approaches to Stop Hypertension), based on consuming between 15% and 20% protein, between 52% and 58% carbohydrates, and between 25% and 30% fat (Campbell, 2017). On the other hand, the diet to treat Obesity is based on low-calorie, low-fat, and low-carbohydrate (Fock and Khoo, 2013). Finally, people treating Anorexia Nervosa usually consume between 15% and 20% protein, between 60% and 65% carbohydrates, and between 20% and 25% fat (Baskaran et al., 2017).

Obtaining an F1-Score between 70% and 85% (i.e., we obtained 77%), we can claim that there is no case of Overfitting or Underfitting, as there is no overtraining of the data when having an evaluation far from perfect. However, there is a sufficiently successful evaluation to deduce that there is also no poor data training (Ying, 2019). In addition, from the analysis of the ROC curves, the probability of correctly detecting Diabetes and Hypertension in patients was 85.7% and 83.2%, respectively, which lets us infer it is a high

Table 1: Questionnaire regarding HcP's and Patient's pre/post-experience using Fynex.

Question	Type of Question: MC ⁽¹⁾ , OE ⁽²⁾ , LS ⁽³⁾ or YN ⁽⁴⁾	Role: HcP ⁽⁵⁾ or P ⁽⁶⁾	Moment of Application: B ⁽⁷⁾ or A ⁽⁸⁾
Q1: Provide your gender.	MC	HcP & P	B & A
Q2: Provide your age.	OE	HcP & P	B & A
Q3: For how many years have you been a HcP?	OE	HcP	B & A
Q4: How secure do you feel about the medical recommendations you give to your patients?	LS	HcP	B & A
Q5: Justify your provided answer in Q4.	LS	HcP	B & A
Q6: How easy do you consider is it to monitor your patient's health condition?	LS	HcP	B & A
Q7: How easy is it to communicate with your patient/HcP?	LS	HcP & P	B & A
Q8: How useful are the nutritional and exercise plans recommended by Fynex, for the treatment of diseases based on eating disorders?	LS	HcP	A
Q9: Is the website comfortable and attractive, allowing easy use of the application?	YN	HcP	A
Q10: Is the information offered by the application clear and sufficient?	YN	HcP	A
Q11: The nutrition and exercise plans recommendations are generated fast?	YN	HcP	A
Q12: For how many years have you been a medical treatment?	OE	P	B & A
Q13: Do you think you eat well?	YN	P	B & A
Q14: Are you in treatment with a nutrition professional?	YN	P	B & A
Q15: How satisfied are you with your treatment?	LS	P	B & A
Q16: How much do you trust your HcP's recommendations?	LS	P	B & A
Q17: Justify the provided answer in Q17.	OE	P	B & A
Q18: How easy is it to monitor your own health?	LS	P	B & A
Q19: Is Fynex comfortable and attractive, allowing easy use of the application?	YN	P	A
Q20: Is the information provided by Fynex clear and sufficient?	YN	P	A
Q21: How easy is it to use Fynex?	LS	HcP & P	A
Q22: Would you recommend Fynex to others?	YN	HcP & P	A
Q23: Provide additional comments	OE	HcP & P	B & A

Notes: ⁽¹⁾MC: Multiple Choice, ⁽²⁾OE: Open-Ended Question, ⁽³⁾LS: [0-5] Likert Scale, ⁽⁴⁾YN: Y/N Questions, ⁽⁵⁾HcP: Healthcare Professional, ⁽⁶⁾P: Patient ⁽⁷⁾B: Before using Fynex, ⁽⁸⁾A: After using Fynex

probability of preventing and treating these diseases.

We calculated the usability and acceptance tests through the SUS (i.e., System Usability Scale) and the NPS (i.e., Net Promoter Score). As a result, HcP obtained an average SUS of 76.69, and the patients

obtained a SUS of 90.36. Additionally, the NPS obtained by HcP was 14.29%. Even if the resulting value was not that significant, we considered it a positive possibly recommendation (Mendieta Giron et al., 2017). Likewise, patients resulted with an NPS of

Table 2: Questionnaire to calculate System Usability Scale (SUS) and Net Promoter Score (NPS) of Fynex.

Question	Scale Option
I think that I would like to use this system frequently.	[1-5] SUS
I found the system unnecessarily complex.	[1-5] SUS
I thought the system was easy to use.	[1-5] SUS
I think that I would need the support of a technical person to be able to use this system.	[1-5] SUS
I found the features and functionalities in this system were well integrated.	[1-5] SUS
I thought there was too much inconsistency in this system.	[1-5] SUS
I would imagine that most people would learn to use this system very quickly.	[1-5] SUS
I found the system very cumbersome to use.	[1-5] SUS
I felt very confident using the system.	[1-5] SUS
I needed to learn a lot of things before I could get going with this system.	[1-5] SUS
How likely are you to recommend Fynex?	[1-10] NPS

85.71%, which is a significant value due to the relationship between promoters and detractors.

6 DISCUSSION

From the preliminary results and comments received by the participants of this pilot study, Fynex effectively provides friendly and user-centered support to medical processes and healthcare control, regarding diseases based on eating disorders. It integrates AI models and content-based, ML model-based, and memory-based recommendation systems supporting nutrition and physical activity (i.e., medical treatment control or process). Regardless of the results' volatility, satisfaction increased in both roles (i.e., HcP and Patients). From the preliminary results, both feel more satisfied with the treatment process when using Fynex as a support tool. Although both roles consider that the application does offer support, it was more significant in patients than HcP. Hence, we could claim that patients are more open to new technologies and trying different solutions to address their problems more directly. Moreover, the Harris Poll, in association with Stanford University, analyzes these circumstances, particularly HcP satisfaction with health-focused technologies (Wuerdeman et al., 2005). They claim that 55% of HcP are willing to experiment and use new technologies only if other HcP have used them before and have recommended their use.

The results (i.e., before and after interacting with Fynex) show that the HcP's satisfaction increased by 32.95% in the experiment. However, when applying the Probit and Logit technique, the satisfaction increased by 57.6%. On the other hand, the patients' satisfaction increased by 56.96% in the experiment. Also, the increase obtained when comparing the predictions of the patients' models used in the Probit and Logit was 55.47%.

We find that a web-based tool such as Fynex supports and complements medical processes (i.e., healthcare control, diagnoses, and treatments) for HcP and Patients. Based on the participants' comments regarding Fynex, in addition to the different similar tools that exist, the participants found satisfactory the possibility that both HcP and patients could interact in the medical treatment processes synchronously on a single platform. Fynex allows both end-users (i.e., HcP and patients) to have on-time information regarding the available treatment process independently. HcP can monitor their patients and make decisions to support medical treatments. Likewise, patients can access and monitor their health status (i.e., healthcare control), hand in hand with the HcP medical control and decisions. In further research, a deepened analysis will be addressed to complement the preliminary claims presented in this paper.

ACKNOWLEDGEMENTS

The authors thank Dayan Alejandra Hincapié-Cortés for her dedication, performance, and dedication to the successful development of Fynex, which was vital to achieving the results and findings obtained and reported in this pilot study. Also, the authors thank all participants who voluntarily and actively contributed to the preliminary experimentation of Fynex.

REFERENCES

- Barata, J. J. R., Munoz, R., Silva, R. D. D. C., Rodrigues, J. J., and De Albuquerque, V. H. C. (2019). Internet of things based on electronic and mobile health systems for blood glucose continuous monitoring and management. *IEEE Access*, 7:175116–175125.

- Baskaran, C., Carson, T. L., Campoverde Reyes, K. J., Becker, K. R., Slattery, M. J., Tulsiani, S., Eddy, K. T., Anderson, E. J., Hubbard, J. L., Misra, M., et al. (2017). Macronutrient intake associated with weight gain in adolescent girls with anorexia nervosa. *International Journal of Eating Disorders*, 50(9):1050–1057.
- Bustamante, J. (2010). La herramienta de tests de usabilidad a distancia loop11. *Profesional de la Informacion*, 19(4):425–430.
- Campbell, A. P. (2017). Dash eating plan: an eating pattern for diabetes management. *Diabetes Spectrum*, 30(2):76–81.
- Candela, J. M. (2016). ¿cuáles son los factores de riesgo para desarrollar diabetes mellitus tipo 2? *GUÍA DE*.
- Çano, E. and Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6):1487–1524.
- Carlson, J. (2013). *Redis in action*. Simon and Schuster.
- Chen, G. and Tsurumi, H. (2010). Probit and logit model selection. *Communications in Statistics—Theory and Methods*, 40(1):159–175.
- Cruz, O. L. and Buitrago, C. H. O. (2017). Formación de ingenieros informáticos como gestores de innovación: Transformadores de vidas, constructores de futuros posibles. *ANFEI Digital*, (7).
- Danilova, V. and Ponomarev, A. (2017). Hybrid recommender systems: The review of state-of-the-art research and applications. In *Proceedings of the 20th Conference of FRUCT Association*.
- de Ullibarri Galparsoro, L. and Fernández, P. (1998). Curvas roc. *Atención Primaria en la Red*, 5(4):229–35.
- Fawcett, T. (2006). Introduction to receiver operator curves. *Pattern Recognit. Lett.*, 27:861–874.
- Fock, K. M. and Khoo, J. (2013). Diet and exercise in management of obesity and overweight. *Journal of gastroenterology and hepatology*, 28:59–63.
- Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., and Bryant, B. E. (1996). The american customer satisfaction index: nature, purpose, and findings. *Journal of marketing*, 60(4):7–18.
- Garcia, M. A., Gandhi, A. J., Singh, T., Duarte, L., Shen, R., Dantu, M., Ponder, S., and Ramirez, H. (2001). Esdiabetes (an expert system in diabetes). In *Proceedings of the twelfth annual CCSC South Central conference on The journal of computing in small colleges*, pages 166–175.
- Gray, A. and Threlkeld, R. J. (2015). Nutritional recommendations for individuals with diabetes.
- Green, B. and Seshadri, S. (2013). *AngularJS*. ” O’Reilly Media, Inc.”.
- Grier, R. A., Bangor, A., Kortum, P., and Peres, S. C. (2013). The system usability scale: Beyond standard usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 187–191. SAGE Publications Sage CA: Los Angeles, CA.
- Joshi, A., Kale, S., Chandel, S., and Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.
- Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer.
- Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*.
- Mandal, P. C. (2014). Net promoter score: a conceptual analysis. *International Journal of Management Concepts and Philosophy*, 8(4):209–219.
- Mendieta Giron, C. F. et al. (2017). Ventajas del net promoter score para la atención al cliente en las empresas de telecomunicaciones.
- Miller, C. J., Smith, S. N., and Pugatch, M. (2020). Experimental and quasi-experimental designs in implementation research. *Psychiatry Research*, 283:112452.
- Munzner, T. (2014). *Visualization analysis and design*. CRC press.
- Rull, G., Farré, C., Teniente, E., and Urpí, T. (2009). Mvt: a schema mapping validation tool. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 1120–1123.
- Sampé, J., Vernik, G., Sánchez-Artigas, M., and García-López, P. (2018). Serverless data analytics in the ibm cloud. In *Proceedings of the 19th International Middleware Conference Industry*, pages 1–8.
- Seth, R. and Sharaff, A. (2022). A comparative overview of hybrid recommender systems: Review, challenges, and prospects. *Data Mining and Machine Learning Applications*, pages 57–98.
- Stark, B., Knahl, C., Aydin, M., and Elish, K. (2019). A literature review on medicine recommender systems. *International journal of advanced computer science and applications*, 10(8).
- Thomas, R., Halim, S., Gurudas, S., Sivaprasad, S., and Owens, D. (2019). Idf diabetes atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018. *Diabetes research and clinical practice*, 157:107840.
- Vassiliadis, P. and Simitsis, A. (2009). Extraction, transformation, and loading. *Encyclopedia of Database Systems*, 10.
- Waki, K., Fujita, H., Uchimura, Y., Omae, K., Aramaki, E., Kato, S., Lee, H., Kobayashi, H., Kadowaki, T., and Ohe, K. (2014). Dialbetics: a novel smartphone-based self-management support system for type 2 diabetes patients. *Journal of diabetes science and technology*, 8(2):209–215.
- Wuerdeman, L., Volk, L., Pizziferri, L., Tsurikova, R., Harris, C., Feygin, R., Epstein, M., Meyers, K., Wald, J. S., Lansky, D., et al. (2005). How accurate is information that patients contribute to their electronic health record? In *AMIA annual symposium proceedings*, volume 2005, page 834. American Medical Informatics Association.
- Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

User Empowering Design: Expanding the Users' Hierarchy of Needs

David Gallula¹, Hadar Ronen², Ido Shichel³ and Adi Katz⁴

¹Department of Information Systems, The Yezreel Valley Academic College, Israel

²Department of Education and Society, Ono Academic College, Israel

³Department of Learning Disabilities, University of Haifa, Israel

⁴Usability Center, Industrial Engineering and Management Department, SCE, Ashdod, Israel
davidga@yvc.ac.il, hadar.ro@ono.ac.il, shichel003@gmail.com, adis@sce.ac.il

Keywords: User Empowerment, User Empowering Design, User-Centred Design, Usability, User Experience, Human Computer Interaction.

Abstract: We offer a comprehensive framework for HCI designers to empower users. Our purpose is to lay the foundations for a complete design method: User Empowering Design (UED). Empowerment is not a new human need, and the ability of technologies to empower users has already been discussed and deemed important by many experts. UED concentrates on the human need to fulfil the potential for growth, increase strengths, and reduce weaknesses. It does not replace user-centred design (UCD) but expands it to a higher-level need. Thus, we suggest that UED is the third revolution in developing interactive technology. Just as user experience (UX) has expanded usability, UED expands the user experience. The UX design method focuses on the user experiences while interacting with a digital product (e.g., emotions, feelings, aesthetics, hedonic needs). The UED approach takes a step further, as it focuses on the impact that goes beyond the scope of the digital product with an emphasis on aspects that are in the user's main life path. Within this framework, we offer a methodology that includes eight distinct models, each referring to a variety of technologies that have an impact on personal, community and organisational empowerment. We explain the current need for a design that is intended to empower users, present the different models by describing their mechanisms, and offer practical examples to implement the empowering design.

1 INTRODUCTION

Technological advancement in recent decades has led to fundamental changes in the way users think about and use their digital products. Today, users are no longer settling for a merely functional product, or for having a pleasant experience when using the product. It seems that as technology progresses, users also expect more from their digital products. In this paper, we argue that product designers should address an additional need, which is at a higher level than user experience. Therefore, we call for **User Empowering Design** (hereafter UED). In his seminal paper, Shneiderman (1990) declared that users need to be empowered by technology and stated that they should “be able to apply their knowledge and experience to make judgments that lead to improved job performance and greater personal satisfaction.” The

term “Empowerment” denotes reinforcing and enabling people so that they realize their abilities, increase their strengths, have a high level of belief in themselves, and perceive themselves as more self-efficient (Brendtro et al., 2002). Our research follows the idea that there should be design methods that intentionally achieve user empowerment (Gallula and Frank, 2014) and aims to encourage innovative ideas and tools for empowering users with technologies.

In recent years, we have witnessed a rise of the concept of empowerment in various fields, such as employment and learning. In the field of employment, one of the most prominent factors that impact employees’ engagement, their commitment to the organisation, and their level of performance at work, is employee empowerment (Osborne and Hammoud, 2017; Hanaysha, 2016). In the past, work was regarded as a source of livelihood and financial security. Later generations sought interest and self-

^a <https://orcid.org/0000-0002-7065-2819>

^b <https://orcid.org/0000-0002-8143-709X>

satisfaction in the workplace (including a good atmosphere, and fascinating tasks). Nowadays, people are looking for a workplace that allows them to develop abilities and skills, grow and feel empowered. In a similar vein, in the field of learning, the rhetoric of empowerment has been linked to formal education (Selwyn, 2011). This is evident in the literature related to learning technologies and pedagogic approaches, where the Internet is considered an appropriate space for autonomous learning and where students can exercise their power and experience with a high degree of freedom and control (see, e.g., Prinsloo and Slade, 2017; Bennett and Folley, 2019).

In this paper, we will address the manner in which digital products can provide empowerment to their users. We begin by describing the user empowerment framework, followed by a discussion of the applied implications of the model, and conclude with recommendations for future research. Before we broaden the scope of user empowerment and discuss how it can be applied in practice in designing digital products, we must first understand the building blocks on which user empowerment is laid, i.e., the current designers' focus on user's interaction with digital products and determine whether existing trends fulfil the aim of user empowerment.

1.1 Time to Address a Higher Order of User Needs

According to the User-Centred Design (hereafter UCD) approach, the purpose of a system is to serve its users, and the needs of the users should lead the design of the system's interface (Norman, 1986). Currently, the UCD approach concentrates on two main design methods that meet the users' needs: Usability and User Experience (UX). Usability emphasizes the effectiveness, efficiency, and satisfaction of execution (Nielsen, 1993) and UX emphasizes the feelings, sensations, and experiences of a user (Norman, 2004).

The UED framework is inspired by the ERG (existence, relatedness, growth) theory (Alderfer, 1969). This refers to an order of needs, in which certain needs serve as building blocks for higher needs. If we relate to Alderfer's theory, existence is the basic material required for living and is the cornerstone of the hierarchy pyramid; relatedness emphasizes the importance of social interactions and positive emotions, and growth describes the urge to continue personal development and therefore is at the top of the pyramid. We propose an extension to the UCD model by adding a third and higher level to the

users' needs hierarchy. We claim that it is time to consciously add user empowerment when designing digital products.

1.1.1 Usability: The First Need

Usability addresses the users' need to perform well in tasks and successfully complete the tasks conducted via the interfaces of digital systems (Nielsen, 1993). It refers to the extent to which a system, product, or service "can be used by specified users to achieve specific goals with effectiveness, efficiency (in minimum time, with minimum errors), and satisfaction in a specified context of use" (Bevan et al., 2015; ISO FDIS 9241-210). Nielsen referred to usability as a method for improving ease of use during the design process. The usability emphasizes the user's need to achieve pragmatic goals such as successful and efficient performance in tasks supported by the computerized system, with a focus on cognitive and task-oriented aspects, emphasizing the instrumental value of technological products (Hassenzahl and Tractinsky, 2006).

1.1.2 User Experience: The Second Need

A User Experience (UX) goes beyond meeting the instrumental (pragmatic) user's needs for usefulness, efficiency, and satisfaction. It strives to optimize the user's holistic experience, which includes the need to achieve hedonistic goals (Hassenzahl and Tractinsky, 2006). Among these goals, users require aesthetic pleasure, fun, being surprised, and entertainment (Gaver and Martin, 2000). The holistic UX includes emotions, beliefs, preferences, psychological responses, and physical behaviours (Bevan, 2009).

1.1.3 Empowerment: The Third Need

Shneiderman (1990) declared that users need to be empowered by technology. Kim and Gupta (2014) demonstrated the critical role of user empowerment in regulating IS infusion in organizations. Schneider et al. (2018) claim that empowering people through technology is of increasing concern in the HCI community and reviewed 54 CHI full papers using the terms empower and empowerment. Inspired by these previous works, we too concentrate on the human need to be empowered.

1.1.4 User Empowering Design (UED)

User empowering design (UED) goes a long way beyond the concept of UX. While UX strives to increase feelings and emotions of pleasure and

enjoyment, these remain within the limited context of interacting with the product or system. In contrast, UED focuses on the effects of the interaction beyond its occurrence, not in limited terms of reflection on the user experience, but in a way that benefits aspects of the user's life that are outside and beyond the product's use experience. UED is a powerful idea, because it is based on the understanding that technological product designers can positively impact the users' well-being. UED strives to strengthen the users and minimize their weaknesses, which are reflected in the central path of their life. Therefore, it significantly expands the ramifications that a product, system, or service has for its user by fulfilling the user's need to be empowered in their personal life.

We define UED as a method built from a collection of design rules and principles aiming to design HCI systems that meet the need for user empowerment. These systems will have positive effects on the perceptions or abilities that are in the main path of the user's life, in a way that minimizes the user's weaknesses and/or emphasizes the user's strengths. While it is possible that the existing models for designing user-technology interaction might provide user empowerment unintentionally as a most welcome side effect, there is more to aspire to. We argue that the existing situation in HCI should not be left intact, and the growing need for empowerment must be addressed; The fact that there is currently no orderly design method that consciously provides an answer for this need motivated us to explore and offer a preliminary basis for such a method. As the previous two methods have evolved over decades of research and practice, we do not claim to offer in this paper a complete and encompassing method, but only raise awareness of this need and suggest implementation options.

Just as the need for good UX has not replaced the need for usability, so the need for empowerment will not replace the two previous needs. An empowering product should also be usable and provide a good UX. Different user need should not come at the expense of each other. However, different products may emphasize a particular need, according to their purpose and the context of their use.

1.1.5 Why Now?

To this date, there is no orderly method designed to empower users. We believe that the conditions for empowering users have matured, and this is the time for a qualitative leap towards a design that is consciously dedicated to user empowerment. This argument is based on four components:

First, it is the *Zeitgeist* of the current generation who grew up with technology, which leads its users to take functionality for granted, cease to be satisfied with positive experiences, and expect technology to improve their lives. As in the other areas of life we mentioned earlier - work and education - expectations from technology are undergoing a revolution. Usability and UX are already considered foundational and the need for a transition to the next level has arisen.

The second factor is technological maturity. Nowadays, the hybridity and friction between humans and technology (e.g., mobile and wearable devices, augmented reality, virtual reality, IoT) create a symbiosis between users and the technological products that surround them. Advanced digital developments such as Metaverse will increase this friction. New digital products will continue to adjust and adapt themselves to the user (e.g., by using AI and machine learning). While these technological developments integrate and keep the users more and more inside the digital matrix, the concept of user empowerment emphasizes positive changes in the users' well-being outside the virtual space.

The third factor is based on the fact that, at present, steadiness, in terms of not regenerating products, is detrimental to digital companies, and therefore there is a constant demand for continuous innovation. Big tech companies that have reached maturity and technological extraction are seeking their next breakthrough. Having a competitive innovation edge is a key aspect of survival and growth. This makes 'innovate or perish' the new motto of the global knowledge economy era (Shakina and Barajas, 2020). Thus, one way to overcome technological barriers and continue to innovate is by addressing the human high-quality need to be empowered.

Lastly, Netflix' "Social Dilemma" documentary, which came out in 2020, raised public awareness of the negative sides of social networks, including their ability to act as an "addictive drug" (see also Eyal, 2014 on the addictive patterns underlying technologies). We use the term "Junk Tech" (Bally and Desmaison, 2021) to describe technology as tasty but not healthy for us users. That is, we consume digital products in a way that is equivalent to eating unhealthy fast food. Technological products capture our attention and without being aware, we fall into the rabbit hole. This, for example, causes us to stay awake late at night, and thus affects us negatively the next day. Usability, user experience, and empowerment can be analogous to food consumption influences. Just as consumers have realized that they

can and should demand food that is not only fast and tasty, but also healthy, so will users of digital products undergo a transformation and stop settling for just the experience. They will demand products that will also improve their well-being, lead them to personal growth and help them fulfil the important goals in their lives. In other words, people will look for products that empower them to manage their well-being and health (Blandford, 2019; Seshadri et al., 2019). Illustrative examples are apps for body self-acceptance to elevate the self-esteem of young people (Paay et al., 2018), for raising the sense of self-efficacy among job seekers such as "review me" and "Interview4" (Dillahunt and Hasiao, 2020), and for dealing with social anxiety, such as the app "CogniShift" (Elefsen and Khaliq, 2021).

2 HOW TO IMPLEMENT UED

We propose to expand the UCD theory and add practical tools that enable user empowerment.

To develop a digital product, one must start from the premise of the steps within the UCD. Within them, we will focus on the first two processes of the UCD - research and concept, as a thorough examination of the design and test stages are beyond the scope of the current paper. When developing a new digital product or updating an existing one, the designers of the product must keep in mind that the product can improve the user's strengths, and we encourage them to think of ways to help the users reach their potential in the central path of their life.

2.1 Research

We consider the common tool of personas as an effective UCD method (Cooper, 1999; Jansen et al., 2022; Goodwin, 2011). In UED, the persona's description (user card) includes additional layers that describe the importance of the product for the user in terms of the technological product's potential impact on the user's life beyond the course of interaction with the product itself.

The UED approach introduces additional important aspects to the process of building personas, including the user's personal potential mapping tool. Using this tool, we ask about the life of the user and define their "main path," i.e. their important goals and ambitions, strengths and weaknesses in the context of that main path. Applying our mapping tool, for each persona we add several Likert-type scales, each representing an attribute that is relevant to the main path. The attributes are important characteristics or

abilities that the persona wants to have or fulfil. On each attribute scale, we mark the persona's current state and their potential future state. In addition, we describe the factors that currently inhibit or prevent the persona from realizing their potential.

2.2 Concept

We suggest that the product designer employ a systematic usage of our methodology. After the designer defines the trait/ability that the user wishes to emphasize and identifies the factor that inhibits the user's potential to advance that trait (the "preventive factor") – the methodology systematically leads the designers through certain paths to the most suitable design solutions.

The methodology we offer includes eight distinct models, demonstrated in Table 1. Each model appears in a separate column and represents a different "preventive factor." We suggest a suitable design solution for each of these factors, which may empower the user beyond the course of interaction with the technological product. We divided the models into two groups: four models see technology as the solution for dealing with the preventing factor, and the other four models identify the source of the preventing factor in the technology itself.

Due to space constraints, we will demonstrate only two models out of the eight described in Table 1, the Walls and the Ripples. For each, we suggest design features specifically intended to empower users using a popular instant messaging application, WhatsApp.

2.2.1 The Wall Model: Intergenerational Dictionary

Many adults experience technology as an obstacle in their lives. Age-related barriers are derived from normal changes that a person undergoes in cognition and physical conditions, along with affective barriers (age-related perceptions and attitudes towards technology). In addition, technologies are usually not designed to be age-friendly. The combination of these factors leads to intergenerational gaps and negative consequences, including digital exclusion that may, in turn, lead to social exclusion and deterioration of the sense of belonging (Katz and Sophia, 2021). The acquisition of online literacy for digital communication creates a linguistic gap that distinguishes generation Y from the baby boomer generation (Subramaniam and Razak, 2014). The use of nonstandard language, emojis and new forms of slang in social networks may lead to misunderstandings and communication barriers.

Table 1: The eight model that guides designers to empower users.

		Technology offers the solution				Technology causes the problem			
		mental aspects		lack of abilities		unintentional		intentional	
		motivation	barriers	transient	permanent	exclusion	side effects	to addict	to harm
		Engines	Breaks	Trainers	Bridges	Walls	Ripples	Hooks	Bad Wolf
Explanation		Technology deals with lack of motivation that prevents the user from persevering and succeeding	Technology deals with fears and changes thoughts that hinder the user from realising his or her potential	Technology offers proper guidance and coaching, so users will be able overcome transient difficulties and inabilities	Technology can help users overcome a permanent or inherent disability	Technology was not designed to fit the user's needs, and therefore becomes an obstacle, excluding users.	While technology solves a problem, it may create others, by causing unexpected by-products such as loss or degrade of capabilities	Technology was designed to addict the user or negatively affects the user	Technology was designed to allow for evil or is exploited by users in immoral ways
		Create a habit by encouraging existing motivations and by sparking hidden motivations	Encourage small successes and dissolve fears in a safe environment	Build a training process that will encourage the user to succeed without the product	Build a tool that will provide a permanent solution	Consciously include the user in the design process.	Refine and weaken the by-product outcomes without impeding the main product	Neutralize an addiction or create an alternative positive conduct	Block the possibility to harm and protect the users

One of the weaknesses of the WhatsApp application is faulty communication caused by the user being unfamiliar with unique slang and terms that are typical to the medium of instant messaging correspondence. Through the lens of the Walls Model, Katz and Sophia (2021) identified this problem and suggested a solution by designing “WhatsThat,” which provides a synchronous automatic translation for dialogue participants.

WhatsThat offers a solution that reduces the gap between participants and minimizes possible negative consequences for the well-being of all family members, especially older adults. Figure 1 shows a screen that demonstrates message exchanges between Grandpa David and his grandson, Tamir. The figure presents Grandpa David’s view. Grandpa David clicked on an embedded link for the word “sup” that the system generated automatically and received contextual knowledge (an explanation) retrieved from the system’s dictionary (Figure 1).

The design contributes to fulfilling the user’s communication potential with loved family members, especially grandchildren, by building common ground, increasing mutual understanding and overcoming potential miscommunications. This design focuses on empowering older adults’ self-efficacy, a meaningful goal that is on the main path of their lives.

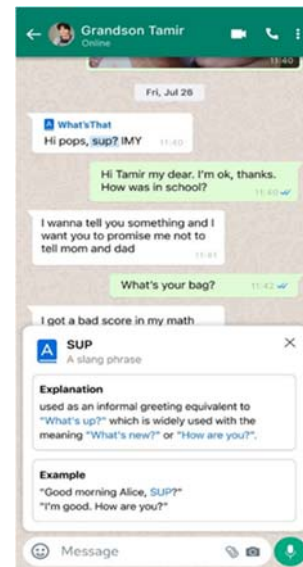


Figure 1: A synchronous automatic “translation” for dialogue participants.

2.2.2 The Ripples Model: Displaying the “Important” along with the “Recent”

Nowadays, people make extensive use of instant messaging applications throughout almost every waking moment. They constantly check, read, and

send messages in the context of various relationships and activities: family, friends, work, leisure, colleagues and more.

The Ripples Model portrays situations in which technology that was invented to solve a certain problem inadvertently creates other problems. For example, nowadays people hardly remember telephone numbers due to the high availability of digital personal contact apps. Another example is navigation apps, which help people efficiently move from one geographic point to another, freeing them from the need to remember the directions required to reach their destination. However, this technological assistance may come at the cost of the loss of intrinsic navigational skills and lack of attention to landmarks situated in a route. Thus, researchers list many negative side effects and ramifications of technological designs, including loss of navigational skills, memory loss, lack of attention, inability to form interpersonal relationships, loneliness and more (Amichai-Hamburger and Etgar, 2018; Turkel, 2017; Goodman, 2021).

For the sake of continuity, we will once again use WhatsApp to demonstrate another ripple effect. In order to meet user needs, WhatsApp designers surely applied the rules of usability and UX. To fulfil users' needs for efficiently detecting and reading messages, chats with the most recent incoming messages are presented at the top of the screen. This design strategy also reinforces the emotional UX level, since people become more engaged when seeing a new message. This emotional response is related to the widespread *Fear Of Missing Out* (FOMO), which triggers the need to be constantly updated. Presenting the most recent chats at the top supports the immediate need to know what is happening here and now at any given time.

Although the designers' good intentions to provide a user-friendly app that meets the users' need for usability and a good experience, WhatsApp's current design may impede the need of many users to preserve and promote certain relations that are important in their lives. As in various social media platforms, many users feel that their time and attention have been hijacked by the abundant social groups to which they belong and that intensive involvement in social media negatively affects their well-being (Sagioglou & Greitemeyer, 2014; Vorderer et al., 2017; Appel et al., 2020). A mundane message recently received in the kindergarten parents' group finds its way to the top of the chat list. An inflated volume of chats that find their way to the top of the screen may push important contacts (such as family members and dear friends) down the chat history and thereby lead to the "out of sight – out of

mind" phenomenon. This may blur the distinction between recent and important, leading users to focus on recent chats while forgetting about the important contacts that are out of sight. WhatsApp developers identified this problem and attempted to offer various solutions, such as muting groups and allowing users to pin up to three chats at the top of the chat list. However, enabling users to pin more than three chats is not possible. In addition, WhatsApp's current design does not create a continuous awareness of the people that are important to the user. The user's ability to see their important contacts at any moment is crucial for the user's sense of control (Ng, 2004).

We propose a solution that does not require the user to give up the convenience of having recent messages appear at the top of the screen. Rather, we focus on adding a permanent horizontal bar of important contacts to the top of the screen (see Figure 2). We enable and encourage users to select and add their important contacts to the bar that we label "Important to me". This microcopy will appeal to the user's emotions and act as a subtle reminder of who is truly important to them. The ordinary vertical order of chats preserves the recent, while our new horizontal bar simultaneously pins the important to the top. Together, the vertical order of chats and the horizontal bar that we offer allow the user to find the balance between what happened recently (changes dynamically) and what is important (relatively constant), thus restoring their sense of control over the important things in life. This design will contribute to the user's main life path outside the product - maintaining and strengthening the relationships with the people who are truly important to them.

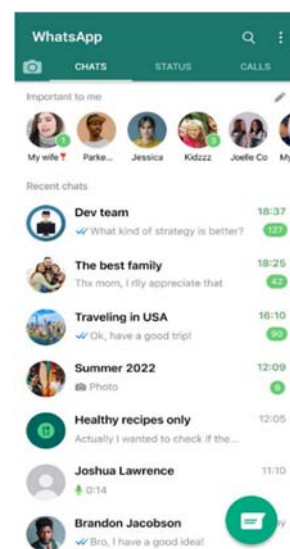


Figure 2: Fixed bar of important contacts in the upper part of the screen.

3 CONCLUSIONS

In this paper, we lay the foundations for a design method – UED, motivated by the fact that there is currently no orderly design method that consciously provides an answer for empowering users. We described the user empowerment framework and explained how it expands UCD to a higher-level user need. We also described practical solutions that implement the empowering design. Our framework offers a methodology that identifies eight models, each representing a variety of situations that may have an impact on personal, community and organisational empowerment.

Empowerment is certainly not a new concept. The innovation we offer is UED, designing digital products intentionally to empower users. UED is a powerful idea, as it emphasises the understanding that HCI designers can positively affect the users' well-being and aspects that are in the central path of the users' lives (behaviours, worldviews, feelings of self-worth, self-efficacy and even perceptions of the meaning of life). While most designers focus on providing influences that occur during the interaction (usability and emotions at the time of interacting), UED is meant to achieve positive effects way beyond the interaction. We wish not only to highlight this ability of designers to enhance the impact of their designs, but also call on designers to take responsibility for the characteristics of the product they produce and design products that will better the lives of their users.

There are already some very interesting and inspiring projects by Maes in MIT Media Lab that show how new body-integrated technologies may assist people by strengthening cognitive skills and “soft” skills such as attention, learning, decision-making, creativity, communication, and emotion regulation. Maes' works with her students and colleagues (e.g., Rosello et al., 2016; Leong et al., 2021) demonstrate how technologies can assist users in overcoming their limitations and realizing their goals. Some applications include intelligence augmentation, memory augmentation, motor augmentation, augmented decision-making and more. These works adopt a design approach to create and study new wearable and immersive systems that can help people reach a life with well-being, and while doing so, they are mindful of ethical and social issues.

In future works, we will further expand the description of the design stages (research, concept, design and test) under the UED lens. For example, we will elaborate on practical instructions for building a persona card that focuses on the user's weaknesses

and strengths. We will also offer design principles for each of the eight models and include a solution plan for various problematic situations. In addition, we will expand the discussion beyond personal empowerment, to social and organizational empowerment. To these ends, we will refine the measures of user empowerment and test the usefulness of UED by means of empirical experiments.

REFERENCES

- Alderfer, C. P. (1969). An empirical test of a new theory of human needs. *Organizational behavior and human performance*, 4(2), 142-175.
- Amichai-Hamburger, Y., & Etgar, S. (2018). Internet and well-being. In *The Social Psychology of Living Well* (pp. 298-318). Routledge.
- Appel, M., Marker, C., & Gnambs, T. (2020). Are social media ruining our lives? A review of meta-analytic evidence. *Review of General Psychology*, 24(1), 60-74.
- Bally, J., Desmaison, X. (2021). *Junk tech - how Silicon Valley won the marketing war*. Hermann
- Bennett, L., Folley, S. (2019). Four design principles for learner dashboards that support student agency and empowerment. *Journal of Applied Research in Higher Education*.
- Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods? In *Proceedings of the Workshop UXEM 2009, Uppsala, Sweden*
- Bevan, N., Carter, J., Harker, S. (2015, August). ISO 9241-11 revised: What have we learnt about usability since 1998? In *International conference on human-computer interaction* (pp. 143-151). Springer, Cham.
- Blandford, A. (2019). HCI for health and wellbeing: Challenges and opportunities. *International Journal of Human-Computer Studies*, 131, 41–51
- Brendtro, L., Brokenleg, M. (2002). *Reclaiming youth at risk: Our hope for the future*. Solution Tree Press.
- Cooper, A. (1999). *The inmates are running the asylum*. Indianapolis, IA: SAMS. Macmillan.
- Dillahun, T. R., & Hsiao, J. C. (2020) Positive Feedback and Self-Reflection: Features to Support Self-efficacy among Underrepresented Job Seekers. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (588, 13 pages)
- Elefsen, K., Khaliq, A. (2021). CogniShift: Smartphone Application for Public Speaking Anxiety. In *Proceedings of Aalborg University (AAU). ACM, New York, NY, USA*, 13 page
- Eyal, N. (2014). *Hooked: How to build habit-forming products*. Penguin.
- Gallula, D., Frank, A. J. (2014, September). User Empowering Design. In *Proceedings of the 2014 European Conference on Cognitive Ergonomics* (pp. 1-3).

- Gaver, B., Martin, H. (2000, April). Alternatives: exploring information appliances through conceptual design proposals. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 209-216).
- Goodman, M. (2021). *Broken attention: How to heal a world fractured by technology*. Kinneret Zmora.
- Goodwin, K. (2011). *Designing for the digital age: How to create human-centred products and services*. John Wiley & Sons.
- Hanaysha, J. (2016). Examining the effects of employee empowerment, teamwork, and employee training on organizational commitment. *Procedia-Social and Behavioral Sciences*, 229, 298-306.
- Hassenzahl, M., Tractinsky, N. (2006). User experience-a research agenda. *Behavior & information technology*, 25(2), 91-97.
- Jansen, B. J., Jung, S. G., Nielsen, L., Guan, K. W., Salminen, J. (2022). How to create personas: three persona creation methodologies with Implications for practical employment. *Pacific Asia Journal of the Association for Information Systems*, 14(3), 1.
- Katz, A., Sophia, Y. (2021). A contextualization feature to overcome intergenerational language barriers in communication apps. In *Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications - CHIRA* (pp. 166-173).
- Kim, H. W., Gupta, S. (2014). A user empowerment approach to information systems infusion. *IEEE Transactions on Engineering Management*, 61(4), 656-668.
- Leong, J., Pataranutaporn, P., Mao, Y., Perteneder, F., Hoque, E., Baker, J. M., Maes, P. (2021, May). Exploring the use of real-time camera filters on embodiment and creativity. In *extended abstracts of the 2021 CHI conference on Human Factors in Computing Systems* (pp. 1-7).
- Ng, A. (2004). User friendliness? User empowerment? How to make a choice? *2004 Spring LIS450III*.
- Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann.
- Norman, D.A. (1986) User-Centered System Design: New Perspectives on Human-computer Interaction. In: Norman, D.A. and Draper, S.W., Eds., *Cognitive Engineering*, Lawrence Erlbaum Associates, Hillsdale, 31-61.
- Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Civitas Books.
- Osborne, S., Hammoud, M. S. (2017). Effective employee engagement in the workplace. *International Journal of Applied Management and Technology*, 16(1), 4.
- Paay, J., Nielsen, H., Larsen, H. and Kjeldskov, J. (2018). Happy bits: interactive technologies helping young adults with low self-esteem. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction (NordiCHI '18)*. Association for Computing Machinery, New York, NY, USA, 584-596.
- Prinsloo, P., Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. In *Big data and learning analytics in higher education* (pp. 109-124). Springer, Cham.
- Rosello, O., Exposito, M., Maes, P. (2016, October). NeverMind: using augmented reality for memorization. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (pp. 215-216).
- Sagioglou, C., & Greitemeyer, T. (2014). Facebook's emotional consequences: Why Facebook causes a decrease in mood and why people still use it. *Computers in Human Behavior*, 35, 359-363. doi:10.1016/j.chb.2014.03.003
- Selwyn, N. (2011). *Education and technology: key issues and debates*. Continuum International Pub. Group.
- Seshadri, P., Joslyn, C., Hynes, M., & Reid, T. (2019). Compassionate design: Considerations that impact the users' dignity, empowerment and sense of security. *Design Science*, 5, E21.
- Schneider, H., Eiband, M., Ullrich, D., Butz, A. (2018, April). Empowerment in HCI-A survey and framework. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).
- Shakina, E., Barajas, A. (2020). 'Innovate or Perish?': Companies under crisis. *European Research on Management and Business Economics*, 26(3), 145-154.
- Shneiderman, B. (1990). Human values and the future of technology: A declaration of empowerment. *Acm Sigcas Computers and Society*, 20(3), 1-6.
- Subramaniam, V., Razak, N. A. (2014). Examining language usage and patterns in online conversation: communication gap among generation Y and baby boomers. *Procedia-Social and Behavioral Sciences*.
- Turkle, S. (2011). *Alone together : why we expect more from technology and less from each other*. New York :Basic Books
- Vorderer, P., Hefner, D., Reinecke, L., & Klimmt, C. (Eds.). (2017). *Permanently online, permanently connected: Living and communicating in a POPC World*. New York, NY: Routledge.

Observing the Uncanny Valley: Gender Differences in Perceptions of Avatar Uncanniness

Jacqueline D. Bailey^a, Karen L. Blackmore^b and Robert King

College of Engineering, Science and Environment, University of Newcastle, Ring Road, Callaghan, NSW, Australia
jacqueline.d.bailey@newcastle.edu.au, karen.blackmore@newcastle.edu.au, robert.king@newcastle.edu.au

Keywords: Uncanny Valley, Uncanniness, Avatar, Human Computer Interaction, Gender(Sex).

Abstract: The creation of avatars is a two-sided coin; on one side we see developers creating avatars with the skills, time, and resources available to them. However, these resources (or lack thereof) may lead to avatars falling into the uncanny valley. On the other side are the end-users who engage with the avatar, who ultimately are the focus for these designers and developers. However, many factors can influence the perception of any avatar created beyond the level of realism, including the physical appearance of the avatar or something more fundamental like its gender(sex). Currently, there is a gap in understanding of the influence of gender(sex) in avatar uncanniness perceptions, and this is mostly missing in design decisions for avatar systems. Bridging this gap has been a source of research focus spanning the development of new technologies for avatar development to measuring end-user perceptions of those avatars. Here we add to this discussion through an experiment involving a set of avatars presented to participants (n = 2065) who were asked to rank them from least to most uncanny based on their perceptions. This representative set of avatars were sourced from publicly available methods and have different levels of realism. Our findings indicate that perceptions of avatar uncanniness based on gender(sex) affects the overall perception of the avatar.

1 INTRODUCTION

‘Avatar’ is a common term that often refers to virtual humans in virtual environments. Originally derived from the Sanskrit word ‘Avatara’, referring to the descent of a god down into the mortal world (Adams, 2014), the term also appears in novels and literature (Stephenson, 1993). Avatars are often associated with virtual humans in serious gaming and simulation training scenarios.

Many factors influence the perception avatars, including aspects of their aesthetic characteristics such as hair colour, clothing, and gender(sex) (Fox et al., 2013). This can also include perceived avatar realism and uncanniness, which particularly affect to avatar faces. The differentiation between aesthetic and perceived characteristics may blur, for example, where a characteristic such as realism may impact on the look of an avatar and be the result of the tools used to create an avatar. This can also be influenced by how the avatar is presented (fully body or head and shoulders). These complex interactions present challenges for designers.

The focus of this work is an examination of gender(sex) in end-user uncanniness perceptions of avatar faces. Similar to the work of Stumpf et al. (2020), we use the social construct perspective of gender, here referred to as gender(sex), whereby gender identification, gender expression and performance might not necessarily align with sex.

There are numerous techniques available to create avatars, however, issues such as available time, funds, and other resource constraints impact the quality of the avatar’s appearance and animation. This is true for avatar face animation where subtle communication cues may be required. The quality of avatar face animation is dependent on several factors, first determined by the fidelity of the 3Dimensional models and textures to represent a human face. Creating highly realistic 3Dimensional models to represent human faces can be achieved using techniques such as linear transformations (Tewari et al., 2017; Thies et al., 2015) and higher-order tensor generalizations (Brunton et al., 2014). Data-driven approaches can also be used to create highly realistic textures of human faces (Saito et al., 2017). With the

^a  <https://orcid.org/0000-0002-8550-902X>

^b  <https://orcid.org/0000-0002-9111-0293>

emergence of approaches, the realism of human faces for avatars may increase, however, this is not yet the case in many practical applications.

The level of realism of an avatar ideally meets end-users' expectations without the feeling of uncanniness affecting their experience (MacDorman et al., 2009). When designing avatars, some explicit consideration of these issues is valuable, including the possible interrelationship with gender(sex) and uncanniness. Various perceptions can be affected by gender(sex), such as physical attractiveness and authority (Patel & MacDorman, 2015). Accordingly, this research seeks to further explore the interrelationship of observer gender(sex) on the perception of avatar uncanniness levels. We achieve this through a large-scale avatar ranking exercise considering the differences in avatar rankings based on the gender(sex) of participants. The goal is to evaluate whether gender(sex) has an impact on the perception of avatar uncanniness, and to generate a robust set of ranked avatars.

Ultimately, this information is important to inform design considerations for avatars, and particularly for avatars that may be used in serious games and training applications. Through this research, we propose that avatar creators adopt a more purposeful consideration of gender(sex) in design considerations for avatar-based systems. This may lead to enhanced affective outcomes for avatars in serious purposes.

The following section provides a review of relevant literature followed by the methodology for our study. Next, the results of the research detail the overall uncanniness ranks as well as potential impacts of participant gender(sex) on these ranks. We conclude with a discussion of the key findings, together with recommendations for future work.

2 BACKGROUNDS

Uncanniness can apply to an avatar when a human-like avatar fails to meet the expected visual, kinetic, or behavioural fidelity of a human. This negative emotional response can be attributed to the Uncanny Valley theory proposed by Mori in 1970 (Figure 1) and is demonstrated through a sharp dip in the linear progression of the perception of human likeness.

Mori suggests that the sense of eeriness is likely a form of instinct that protects us from proximal, rather than distal sources of danger, such as corpses or members of different species. This theory has now also been applied to human-like avatars that have been created using computer graphic capabilities with

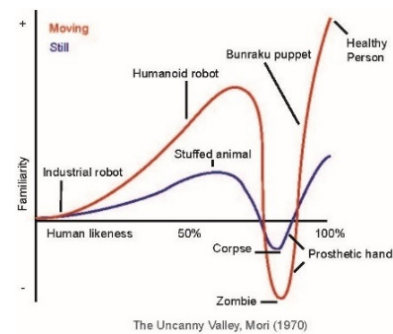


Figure 1: The Uncanny Valley (Mori et al., 2012).

a focus on boundary-pushing realism (Tinwell, 2014). The theory states that the human-like entities, such as avatars, that fail to meet expectations of human behaviour fall into the Uncanny Valley, which may stem from the consequences of realism levels seen in avatar faces.

2.1 The Consequences of Uncanniness Perceptions for Avatar Faces

The potential impact of uncanniness is an important issue in the perception of avatars that should be considered carefully. In particular, as the human visual system excels in detecting falsehoods that may be found in human-like facial features based on prior knowledge of human facial features (Seyama & Nagayama, 2007), avatars are particularly susceptible to uncanny valley effects. These issues extend to more than just the visual representation though.

Another key factor of avatar face animation is the approach used to make the avatar face move and express emotions in realistic ways. There are several methods used to create facial animation, including manual, motion capture, and data-driven approaches. Traditional manual methods can include a frame-by-frame approach or rotoscoping of facial movements. While these manual animation methods can effectively create facial animation, it is a highly time-consuming process. Alternatively, techniques such as motion capture provide a quicker means of capturing an actor's facial movement. These movements can be captured as a live performance through real-time motion capture or as a set of motions mapped post-capture (Davison et al., 2001). Little is understood though of the impacts of these different techniques on perceptions of uncanniness.

Thus, based on this existing research, we undertake a ranking exercise with a set of relatively homogenous avatars to better understand perceptual variability, and to create a robustly ranked set of avatars for future research. With these avatars

quantifiability rated in terms of uncanniness perceptions, further analysis can be carried out from additional data collected on this avatar set. While the set of avatars used in our study are of similar age and ethnicity, they are distinctly different with varying levels of realism as used by Seymour, Riemer Seymour et al. (2018) and Seyama and Nagayama (2007). In summary, we investigate the relationship between a set of avatar faces uncanniness perceptions and how gender(sex) may affect these perceptions.

2.2 Impact of Avatar Gender(Sex)

One fundamental aspect of avatar creation is the perceived gender(sex) of the avatar. There has been some research in this area, with issues such as existing gender(sex) stereotypes, the proteus effect, and gender(sex) swapping being explored. Some gender(sex) based stereotypes were observed by Fox and Bailenson (2009) who suggested that female avatars are more likely to appear either as hyper sexualised as an ornament within games or as a support role. The level of sexualisation of an avatar can also affect their perceived abilities (Wang & Yeh, 2013). When considering the Proteus Effect, Yee and Bailenson (2007) suggest that the appearance of an avatar will lead users to conform to behavioural expectations they have associated with a specific gender(sex). For example, Beltrán (2013) examined this and suggested that this conformity exists in simulation and training contexts, finding that using a male avatar to train professional women can negatively affect her and her peers' achievements. This is important to consider, as Beltrán (2013) argues that most simulation tools show a generic male avatar during training.

Additionally, gender(sex) swapping and exploring an end-user's gender identity or identities is an active area of research. Lehdonvirta et al. (2012) suggest that male participants are more likely to seek and receive help when disguised as a female avatar. Also, Hussain and Griffiths (2008) suggest that there are many social benefits to gender(sex)-swapping in online gaming, such as male players engaging as a female avatar to be more favourably treated by other male players.

Another aspect of avatar gender(sex) is the ability for end-users to explore and express their own gender identity (Baldwin, 2018). This is important as it allows users to express their gender identities which may or may not reflect the gender identity they express in their everyday lives. In video/computer gaming settings, players may be able to remain anonymous online while they explore their gender

identity or identities in a safe platform. To examine the potential influence of gender(sex) in the perception of avatar uncanniness, we conducted a mass-scale survey, described next.

3 STUDY METHODOLOGY

Using a set of 10 homogeneous avatars we consider the impact of participant gender(sex) in the perception of uncanniness. This data was gathered via an online Mturk (amazon.com, 2017) study which took participants 15-20minutes to complete. The participants were asked to rank the avatars from least to most uncanny or eerie. This study was approved by the University of Newcastle's Human Research Ethics Committee (H2015-0163).

In total, (n=2065) participants, with a mean age of 34.82 years completed the survey. Of these participants, 1050 self-identified as female, 1003 as male, three people self-identified as transgender people, two people selected 'other' for their gender(sex), and seven preferred not to say.

3.1 Avatar Set

A set of 10 avatar faces (Table 1) was used in this study to broadly capture the varying levels of realism obtained through different creation methods. As shown, they are of different but homogenous faces with assumed binary sexes (five females and five males). These avatars are considered representative of those found in simulation and training contexts, and are derived from various sources (Alexander et al., 2017; AppleInc., 2015; Battocchi, 2005; Metrics, 2018; Nao4288, 2013; von der Pahlen et al., 2014).

Table 1: Sample Images of the avatar set.

Female Avatars (Real to Low Levels of Realism)				
				
Rose	Emily	Ilana	Liliwen	Baillie
Male Avatars (Real to Low Levels of Realism)				
				
Rycroft	Ira	Victor	Macaw	Leo

3.2 Study Procedure

To begin the ranking survey, participants first completed a set of demographic questions on gender(sex), age, ethnicity, English language proficiency, current residential country, and highest

education level achieved. Additional questions asked participants to nominate whether they identified themselves as a video/computer gamer, on which platforms and media they interacted with avatars, and how many hours a week they spent interacting with avatars. Participants then ranked the avatar set from most uncanny to least uncanny. The approach adopted here follows Lange (2001), who used a similar rating exercise for their study into realism perceptions of virtual landscapes. In both cases, a ranking rather than rating approach is adopted as these have been shown to have higher reliability (Mantiuk et al., 2012; Winkler, 2009).

3.3 Statistical Analysis

The avatar uncanniness ranks were first analysed using Friedman tests. Following a significant Friedman test, post hoc Wilcoxon signed-rank tests using a Bonferroni adjustment for multiple tests (significance level reduced to .0001) were conducted to examine differences in rankings.

The variable ‘observer gender(sex)’ was investigated to determine whether the gender(sex) of a participant was related to the realism and uncanniness rankings of the avatar faces. To measure the impact of gender(sex), we used the approach suggested by (Conover & Iman, 1981) who suggest that rank transformation procedures allow the use of parametric methods. This would normally be done with observed scores being converted to ranks first. However, in this study, the ranks were provided by the participants, so the transformation step was not needed.

The impact of gender(sex) was tested using the General Linear Model (GLM) as the parametric procedure. For the effect of observer gender(sex), a full factorial two-way ANOVA model was fit with avatar and gender(sex) being the two factors. Follow up testing with pairwise comparisons were used to determine significant differences with a Bonferroni adjustment to control the familywise error rate due to multiple testing.

4 RESULTS

We first present results of the Friedman tests to compare the ranks for our avatar set’s uncanniness perceptions, followed by the impact of participant gender(sex) on these rankings.

4.1 Uncanniness Ranking Scores

Based on the Friedman test, there was a statistically significant difference in the perception of the avatars’ uncanniness $\chi^2(9) = 156.254, p < .001$. Interestingly, we see that the means scores for the uncanniness ranks are all clustered between 4.9-5.8. This suggests that there is some variation in the uncanniness perceptions, but it may be very subtle.

The rankings here show that the Mid1 realism avatars frame the two extremes of these ranks, with Victor (Mid1 realism male) being rated as the uncanniest ($M = 4.99, SD = 2.505$), whereas his female counterpart Ilana ($M = 5.87, SD = 2.324$) is considered the least uncanny of the avatars. However, Ira (high realism male) ($M = 5.14, SD = 2.805$) was considered uncannier than his counterpart Emily ($M = 5.37, SD = 3.006$). Leo (Mid2-Low realism male) was ranked fourth most uncanny by participants ($M = 5.47, SD = 2.739$), while Rycroft (real human male) is ranked as the fifth most uncanny avatar ($M = 5.51, SD = 3.114$). Interestingly, Rose (real human female) ($M = 5.65, SD = 3.360$) was considered less uncanny than her counterpart. Despite being considered the least realistic, the Bailie ($M = 5.66, SD = 3.214$) avatar was not considered to be the uncanniest.

A Wilcoxon signed ranked test using a Bonferroni correction, determined where the differences occurred for the uncanniness ranks. Although there were no large differences between the scores for the avatars, there were some significant differences in the ranks of the following avatars. When examining the scores for the high realism male and the real humans, both comparisons were statistically significant; Ira and Rycroft ($Z = -4.527, p < 0.001$) and Ira and Rose the real human female ($Z = -6.229, p < 0.001$). However, the scores for Emily (high realism female) were not statistically significantly different when compared to real humans, which may suggest that the gender(sex) of the avatar may influence uncanniness when compared to real humans. Another pattern occurs in the comparison of the uncanniness rank for Victor (Mid1 realism male); it is statistically significantly different to other avatars, including the female avatar from the same source. This again may suggest some gender(sex)-based perceptual differences.

4.2 The Impact of End-user Gender(Sex) on the Rank Scores

Following the analysis of differences in the overall uncanniness rankings, we consider gender(sex)-based variability in uncanniness ranks. Using a GLM with a

Bonferroni correction, we note a statistically significant interaction of our avatar set uncanniness ranks and the participants' gender(sex) $F(9, 19910) = 10.453, p < .001$.

The Posthoc analysis shows some statistically significant differences between the responses of female and male participants. Specifically, the female participants rated Rycroft (real human male) as less uncanny than the male participants (Rycroft $p = .001$, (Female ($M = 5.68, SD = .093$), Male ($M = 5.26, SD = .088$))). This pattern continues with scores for Rose (real human female) where the female participant scores are higher than the male participants (Rose $p < .001$, (Female ($M = 5.85, SD = .093$), Male ($M = 5.33, SD = .088$))). The high realism female is less uncanny by female participants when compared to the male participants (Emily $p = .001$, (Female ($M = 5.54, SD = .093$), Male ($M = 5.12, SD = .088$))).

In contrast, the high realism male avatar (Ira) was ranked uncannier by male participants (Ira $p = .005$, (Female ($M = 5.27, SD = .093$), Male ($M = 4.90, SD = .088$))). The independently sourced avatar (Baillie) was deemed less uncanny by male participants than the female participants (Baillie $p < .001$, (Female ($M = 5.47, SD = .093$), Male ($M = 5.98, SD = .088$))). Finally, Macaw (Mid2-Low male avatar) was regarded as less uncanny by male participants when compared to scores from female participants (Macaw $p = .001$, (Female ($M = 5.30, SD = .093$), Male ($M = 5.75, SD = .088$))). Generally, male participants are more likely to rank high realism avatars as uncanny.

Further, we note limited variation in the responses as the scores only vary in a range between 4.5 and 6.5 when compared by participant gender(sex). However, when examining the distribution of uncanniness scores for each avatar and grouped by participant gender(sex), several patterns emerge as shown in Figure 2.

The initial comparison considers differences in the distributions of the uncanniness scores for each avatar for female and male participant responses. From these histograms (Figure 2), we observe the Mid1 avatars (Ilana and Victor) and the Mid2-Low male avatar (Macaw) scores resemble a normal distribution. In contrast, Rose and Rycroft (real-world humans) and the Mid2-Low female avatar (Baillie) scores have a distinctly bi-modal distribution indicating participants have a divided opinion of these avatars. The remaining avatars present a somewhat flat, uniform like distribution of scores.

Using a 2 sample Kolmogorov-Smirnov Test to compare the shapes of the distributions, and with a Bonferroni correction for testing multiple pairs, we compared the female and male participant

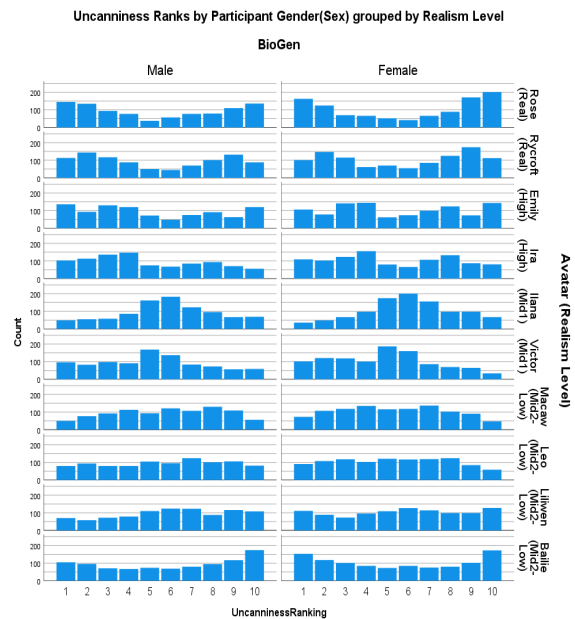


Figure 2: Rank scores comparison by avatar and participant gender(sex).

uncanniness score distributions. We see statistically significant differences in the distribution of the scores for some avatars when comparing the female and male distributions, including Macaw $D(1993) = 1.796, p = .003$, Rose $D(1993) = 1.785, p < .001$, Baillie $D(1993) = 1.785, p = .003$ and Rycroft $D(1993) = 1.819, p = .003$.

The second key comparison is an examination of the differences in the distribution of scores when comparing avatars between female and male participants. Using a series of Kolmogorov-Smirnov Tests, with a Bonferroni correction to examine these differences, reveals some statistically significant differences. When comparing Leo – Liliwen, the distribution of the female scores were statistically different while the male participants scores were not (Female($D(2090) = 1.947, p < .001$), Male ($D(1896) = 1.355, p = .051$)). We also note that the female participant scores when comparing Rose (real human female) and Emily (high realism female) are statistically different when to the male participant scores (Female ($D(2090) = 3.412, p < .001$), Male ($D(1896) = 1.447, p = .030$)). This pattern continues with Rycroft (real human male) and Rose (real human female) with female participants' scores statistically significant while the male scores are not statistically different (Female ($D(2090) = 1.312, p = .064$), Male ($D(1896) = 3.834, p < .001$)). When comparing the both the Mid2-Low female avatars (Liliwen – Baillie), the female participant scores are statistically significant, and the male participants' are not (Female

($D(2090) = 2.166$, $p < .001$., Male($D(1896) = 1.699$, $p = .051$)).

The pattern identified in the preceding paragraph is reversed for some of the avatars. For male avatars (Macaw – Ira), the high realism male comparison shows the male scores being statistically significant and the females do not (Male($D(1896) = 3.834$, $p < .001$), Female ($D(2090) = 1.312$, $p = .064$)). This continues with one of the Mid2-Low female avatars (Bailie) to the real human male (Ira) comparison where the distribution of the male scores is statistically different and the female participant scores are not (Male ($D(1896) = 2.825$, $p < .001$), Female ($D(2090) = 1.444$, $p = .031$)). Again, the comparison of the Mid2-Low male avatar (Leo) to the high realism male (Ira) distributions are statistically significant for male participants but not the female participants (Male ($D(1896) = 3.789$, $p < .001$), Female ($D(2090) = 1.553$, $p = .016$)). Lastly, we see a statistically significant difference between the distribution of scores for the Mid2-Low realism male avatar (Leo) to the high realism female avatar (Emily) for male participants but not for the females (Male ($D(1896) = 3.330$, $p < .001$), Female ($D(2090) = 1.837$, $p = .016$)). The remaining comparisons were not statistically significant for either gender(sex).

5 DISCUSSION AND CONCLUSION

Our investigation into the impact of gender(sex) considered if a participants' gender(sex) affects the perception of an avatar's uncanniness ratings. Male participants rated the higher realism avatars as uncannier than all other avatars except for the Mid1 realism male avatar. This aligns to previous findings where male participants were more sensitive to uncanniness in human-like agents (Tinwell, 2014).

Of interest, the top four ranks for female participants were populated by male avatar faces, suggesting that these avatars may have triggered negative responses for participants who identified as female. Existing research on avatar faces features and uncanniness suggest that avatars failing to display empathy or reacting appropriately may lead to assumptions of psychopathy in an avatar (Tinwell, 2014). As such, the critical dimensions of uncanniness, such as threat avoidance and alignment to terror management theory (MacDorman et al., 2009) may be more enhanced in female participants. It is also interesting that while there was some consistency in the avatars ranked as most uncanny,

the same consensus was not present in those ranked least uncanny, with female and male participants responding differently. For both genders(sexes), the avatars ranked as most uncanny were female, although they were different avatars. These differences led to further investigation of the distribution of uncanniness ranks by participant gender(sex) to reveal further insights.

From examination of the distribution of scores, we see some interesting differences in the female and male participant scores for each avatar. The real humans and one of the Mid2-Low realism female avatars scores exhibit a bi-modal distribution, suggesting that the general opinion of these avatars is divided. Additionally, despite both real humans achieving high realism ranks, these avatars fall around the middle of the uncanniness rankings. These distributions were found to be statistically significantly different when comparing the scores when grouped by the participants' gender(sex). These differences may suggest that some participants were convinced that these avatars were computer-generated images as opposed to real humans, or perhaps factors other than avatar realism contribute to perceptions of uncanniness. Further, we see that the one of the Mid2-Low realism female avatars (Bailie) was ranked as the least realistic but ranked only 8th on the uncanny scale, indicating that the avatar was considered unrealistic but not uncanny. This supports existing literature suggesting lower realism levels can lead to avatars being perceived as less uncanny (Tinwell, 2014).

Our findings highlight the importance of gender(sex) in avatar design decisions due to the impact of this variable on the perceptions of avatar uncanniness. As previously identified, a detailed consideration of the influence of gender(sex) in avatar uncanniness perceptions is mostly missing in the design decisions for avatar systems. Further, it is evident from the literature that the unavoidable design choice of gender(sex) for an avatar may have underlying cues and expectations placed on them based solely on their perceived gender(sex). Together, the findings of the research provide key insights into gender(sex) based perception of avatars.

Despite the significant contributions of this work, it is not without some limitations. Firstly, there is a lack of ethnic diversity in our surveys' sample. The primary ethnicities that completed this survey were of White/Caucasian and Asian background, which may lead to some bias in the interpretation of the data. Secondly, the results cannot be extended to nonbinary genders(sexes) due to small sample sizes. However, this does present an avenue for future work. Lastly,

while the participants in our study come from a wide range of ages (18-87 years old ($M = 34.82$, $SD = 11.52$)), a restriction put in place by Mturk is that participants must be at least 18 years old. Therefore, the results presented in our study may not be applicable to those under 18.

Another area of limitation relates to the avatars themselves. There is a lack of diversity in the ten avatars used in this study. The sample of avatars is primarily of an Anglo-Saxon appearance with little subjective difference in the facial features. Lastly, the current study asks participants to consider the avatars without the context of how they are used. We note that perceptions might differ with context as previously identified (Rosen, 2008). The complexity of the ranking task necessitated the use of a minimal avatar set. However, given that avatars, as virtual representations of humans, could potentially reflect the full diversity of the human form, it is arguable how large a set would be required to be representative. Thus, future work may seek to expand the avatars evaluated to examine the differences in a more diverse set.

Despite these limitations, the work presented here has produced interesting insights into gender(sex) differences in the perception of avatars and generates several avenues for future research. First, the work presented here has focused on perceptual effects of the gender(sex) of both participants and avatars. Future analysis may extend this to explore the differences in the rankings associated with perceptions of avatar-participant self-similarity and avatar sex. An area for further analysis considers the individual attributes of each of the ten avatars through a gender(sex)-swapped lens to further explore gender(sex) as a contributor to perceptions of uncanniness. In summary, the work presented here provides the basis for extending current understanding of gender(sex) differences in the perceptions of avatars.

ACKNOWLEDGEMENTS

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. The authors would like to thank Mr. Kim Colyvas from the Statistical Support Services, University of Newcastle, for his assistance with the analysis of data for this manuscript.

REFERENCES

- Adams, E. (2014). *Fundamentals of game design*. Pearson Education.
- Alexander, O., Rogers, M., Lambeth, W., Chiang, J., Ma, W., Wang, C., & Debevec, P. (2017). *The Digital Emily Project: Achieving a Photoreal Digital Actor*. USC Institute for Creative Technologies.
- amazon.com. (2017). *amazon mechanical turk*.
- AppleInc. (2015). *FaceShift*. In (Version 2014.2.01) FaceShift AG.
- Baldwin, K. (2018). Virtual Avatars: Trans Experiences of Ideal Selves Through Gaming. *Markets, Globalization & Development Review (MGDR)*, 3(3).
- Battocchi, A., Pianesi, F., Goren-Bar, D. (2005). *A first evaluation study of a database of kinetic facial expressions (dafex)* Proceedings of the 7th international conference on Multimodal interfaces, Trento, Italy.
- Beltrán, M. (2013). The importance of the avatar gender in training simulators based on virtual reality. Proceedings of the 18th ACM conference on Innovation and technology in computer science education.
- Brunton, A., Bolkart, T., & Wuhler, S. (2014). Multilinear wavelets: A statistical shape space for human faces. European Conference on Computer Vision.
- Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3), 124-129.
- Davison, A. J., Deutscher, J., & Reid, I. D. (2001). Markerless motion capture of complex full-body movement for character animation. In *Computer Animation and Simulation 2001* (pp. 3-14). Springer.
- Fox, & Bailenson. (2009). Virtual virgins and vamps: The effects of exposure to female characters' sexualized appearance and gaze in an immersive virtual environment. *Sex Roles*, 61(3-4), 147-157.
- Fox, J., Bailenson, J. N., & Tricase, L. (2013). The embodiment of sexualized virtual selves: The Proteus effect and experiences of self-objectification via avatars. *Comput Human Behav*, 29(3), 930-938.
- Lehdonvirta, M., Nagashima, Y., Lehdonvirta, V., & Baba, A. (2012). The stoic male: How avatar gender affects help-seeking behavior in an online game. *Games and Culture*, 7(1), 29-47.
- MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Comput Human Behav*, 25(3), 695-710.
- Mantiuk, R. K., Tomaszewska, A., & Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8), 2478-2491.
- Metrics, I. (2018). *Faceware*. In (Version 3.1) Faceware Tech.
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98-100.

- Nao4288, n. (2013). *Female Facial Animation ver 1.0 (No Sound)* [YouTube Video]. <https://tinyurl.com/yc6k6e3k>
- Patel, H., & MacDorman, K. F. (2015). Sending an avatar to do a human's job: Compliance with authority persists despite the uncanny valley. *Presence*, 24(1), 1-23.
- Rosen, K. R. (2008). The history of medical simulation. *Journal of critical care*, 23(2), 157-166.
- Saito, S., Wei, L., Hu, L., Nagano, K., & Li, H. (2017). Photorealistic facial texture inference using deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, United States.
- Seyama, J. i., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and Virtual Environments*, 16(4), 337-351.
- Seymour, M., Riemer, K., & Kay, J. (2018). Actors, avatars and agents: potentials and implications of natural face technology for the creation of realistic visual presence. *Journal of the Association for Information Systems*, 19(10), 4.
- Stephenson, N. (1993). *Snow Crash*. Bantam Books.
- Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., & Churchill, E. (2020). Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human-Computer Interaction*, 13(1), 1-69.
- Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., & Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. Proceedings of the IEEE International Conference on Computer Vision, Italy.
- Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), 183:181-183:114.
- Tinwell, A. (2014). *The uncanny valley in games and animation*. AK Peters/CRC Press.
- von der Pahlen, J., Jimenez, J., Danvoye, E., Debevec, P., Fyffe, G., & Alexander, O. (2014). Digital ira and beyond: creating real-time photoreal digital actors. ACM SIGGRAPH 2014 Courses,
- Wang, C.-C., & Yeh, W.-J. (2013). Avatars with sex appeal as pedagogical agents: Attractiveness, trustworthiness, expertise, and gender differences. *Journal of Educational Computing Research*, 48(4), 403-429.
- Winkler, S. (2009). On the properties of subjective ratings in video quality experiments. Quality of Multimedia Experience, 2009. QoMEX 2009. International Workshop on Quality of Multimedia Experience, USA.
- Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3), 271-290.

A Comparison of Date Selection Elements on Mobile Touch Devices in eCommerce Sites

Asta Romikaitytė¹, Stelian Adrian Stanci², Javier De Andrés³^a, Daniel Fernández-Lanvin²^b
and Martín González-Rodríguez²^c

¹*Department of Computer Science, Kaunas University of Technology, Lithuania*

²*Department of Computer Science, University of Oviedo, Spain*

³*Department of Computing, University of Oviedo, Spain*

asta.romikaityte@ktu.edu, {UO277653, jdandres, dflanvin, martin}@uniovi.es

Keywords: Date Input, Mobile Device, GOMS, Forms, Usability.

Abstract: There are several works exploring the different ways in which a user can input a date, as it is a very common operation on many websites, but the number of papers that cover this topic on mobile devices is very limited. In this paper several alternatives are considered, studied using Goals, Operators, Methods, and Selection Rules (GOMS), and tested in an online experiment with hundreds of users using these kinds of devices, to see how each one of them performs. The results show that the drop-down menu outperforms the others, which were more novel to the users, in completion time, user satisfaction and the number of errors committed.

1 INTRODUCTION

Nearly two-thirds (64.14%) of all Web traffic is calculated to come from mobile phones (“Mobile vs. Desktop Traffic Market Share [May 2022] | Similarweb,” n.d.). They allow users to conduct daily tasks: order a purchase, make an appointment, plan a travel, etc. In e-Commerce sites users are frequently required to introduce their birthdate during the registration process. A very well-known example is the insurance companies, which ask users to enter their birth, vehicle registration, vehicle manufacturing date, accident date, etc. (for example, *mutua.es*). If the user faces obstacles while filling in the form, they may decide to opt for another, more user-friendly application, thus and rival company. It leads to the loss of users as well as the profit (Wroblewski, 2008). As the form consists of a variety of data input elements and all of them contribute to the whole performance of the form, it is worth in-depth analyzing every single one of them.

In fact, a number of research has already been done analyzing the usability of form elements: some of them provide guidelines for the interaction of the

elements (Linderman, Fried, & 37signals (Firm), 2004; Martin, 2013, 2014; Seckler, Heinz, Bargas-Avila, Opwis, & Tuch, 2014); others study elements' performance for specific tasks (Healey, 2007; Jensen, Hansen, Eika, & Sandnes, 2020). A few studies have been executed in order to compare several date entries (Javier A. Bargas-Avila, Brenzikofer, Tuch, Roth, & Opwis, 2011; Brown, Jay, & Harper, 2010). However, a minority of them investigate the topic in the mobile context (Türkcan & Onay Durdu, 2018).

This research aims to compare three different date input methods on mobile applications that were not found to be compared before. These methods include spinners, drop-down menus and radio-buttons. The study seeks to provide guidelines for date entries' usage discussing their completion time, user satisfaction and errors' probability.

2 RELATED WORK

In previous studies, many aspects of Web forms usability have been explored. This section introduces results found about radio-buttons, drop-down menus,

^a <https://orcid.org/0000-0001-6887-4087>

^b <https://orcid.org/0000-0002-5666-9809>

^c <https://orcid.org/0000-0002-9695-3919>

overall form structure and previous studies focusing on date entries.

Regarding web form improvement, Bargas-Avila et al. study (J.A. Bargas-Avila et al., 2010) summarizes 20 guidelines for usable web form design. The list is divided into five sections: form content, form layout, input types, error handling and form submission. Several suggestions include: place the label above the input field ("Label Placement in Forms :: UXmatters," n.d.); ask one question per row; use radio buttons or drop-down menus for entries that can easily be mistyped (Linderman et al., 2004); for up to four options, use radio buttons (Healey, 2007); order options in an intuitive sequence (Beaumont, 2002); for date entries use a drop-down menu when it is crucial to avoid format errors (Christian, Dillman, & Smyth, 2007). Seckler et al.'s empirical study (Seckler et al., 2014) aims to challenge the research described and applies holistic approach in order to evaluate guidelines' effect on efficiency, effectiveness and user satisfaction. The results revealed that improved web forms resulted in faster completion times, fewer submission trials, and fewer eye movements.

Jensen et al.'s research (Jensen et al., 2020) compared different country entry elements, such as drop-down menus, radio-buttons and text fields with autocomplete. What concerns task completion time, the radio-button interface was found to be the slowest, while text fields were proved to be significantly faster. Even though, no significant difference between the drop-down menu and text-field could be found.

Desktop date entries were analyzed by Bargas-Avila et al. (Javier A. Bargas-Avila et al., 2011). They compared six date input methods: (1) three separate text fields; (2) drop-down menu; (3) text field with the label on the left; (4) text field with the permanent label inside the box; (5) text field with a temporary label inside the box; (6) calendar view. Wrong format and wrong date errors were counted; completion time was measured, and user satisfaction questioned. The fastest completion time was noticed when using versions 3 and 5. Drop-down menu and calendar view showed no formatting errors, but they also had longer input times. Also, more incorrect dates were captured for the calendar view.

Methods for specifying dates in mobile contexts were investigated by Turkcan et al (Türkcan & Onay Durdu, 2018). The study was conducted in order to evaluate text box, divided text box, date picker and calendar view for date entry. As in the previous study, this research tested task completion time, number of errors, and satisfaction, too. In terms of completion

times, even though the text box was found to be the fastest, no significant difference between the text box and the divided-text box was being found. Calendar view proved to be significantly slower. Also, participants made no mistakes when interacting with text boxes, whilst calendar view was found to be the most error-prone. Finally, the greatest satisfaction rate was shown by divided-text box followed by text box.

3 DATE SELECTION IN THE WEB

An analysis of the 10 most visited websites in retail (*amazon.com*, *ebay.com*, *rakuten.com*, etc.), social media (*facebook.com*, *twitter.com*, *instagram.com*, etc.) and information technology (*google.com*, *office.com*, *zoom.com*, etc.) sectors was performed ("Most Visited Websites - Top Websites Ranking for May 2022 | Similarweb," n.d.). Results showed that a drop-down menu for date entry was the most popular (retail – 83%, social media – 100%, IT – 100%). Spanish insurance companies with the heaviest website traffic (*mutua.es*, *general.es*, etc.) are using more diverse date entry elements, such as radio-buttons, text-fields, calendar views ("Most Visited Websites - Top Websites Ranking for May 2022 | Similarweb," n.d.). Some of those input methods have already been analyzed by previous studies (Türkcan & Onay Durdu, 2018), whilst radio-button analysis for date entry was not noticed, thus made in this paper.

4 METHOD

This study uses similar methodologies to the ones used by Bargas-Avila et al. (Javier A. Bargas-Avila et al., 2011) and Turkcan et al. (Türkcan & Onay Durdu, 2018).

4.1 Experimental Design

As a first filter, different date entry methods were evaluated using Goals, Operators, Methods, and Selection Rules (GOMS) method. GOMS allows estimation of the time required to complete different tasks on the GUI (Card, Moran, & Newell, 1983), while the Keystroke-Level (KLM) extension minimizes the effort needed to accomplish the calculations (Setthawong & Setthawong, 2019). The theoretical concept is widely known and used in

Human-Computer Interaction (HCI) research (John, Kieras, & Kieras, 1996), mainly due the GOMS models' ability to make exceptionally accurate predictions (Gray, John, & Atwood, 1993).

The calendars that were examined: (1) mathematical; (2) mathematical-aligned; (3) spinner; (4) date picker; (5) date picker with arrows; (6) slider; (7) radio-buttons. The designs are illustrated in Fig. 1-7.

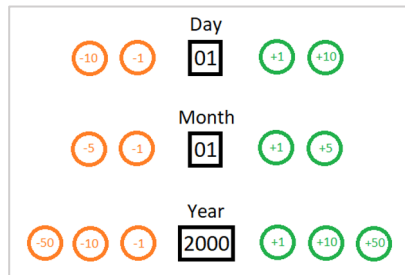


Figure 1: Calendar No. 1 – Mathematical.

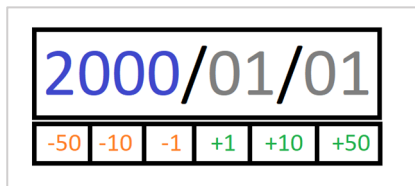


Figure 2: Calendar No. 2 – Mathematical-aligned.



Figure 3: Calendar No. 3 – Spinner.

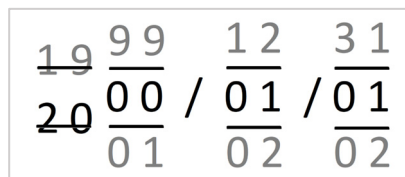


Figure 4: Calendar No. 4 – Date picker.

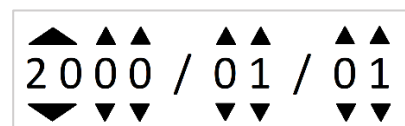


Figure 5: Calendar No. 5 – Date picker with arrows.

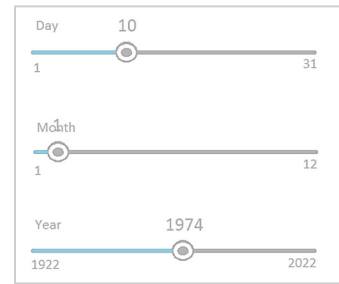


Figure 6: Calendar No. 6 – Slider.

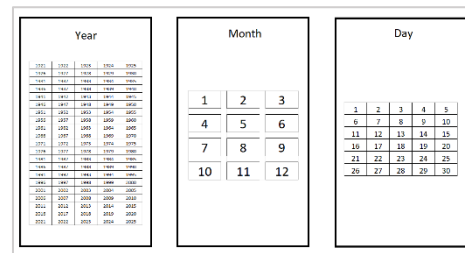


Figure 7: Calendar No. 7 – Radio-button.

As the study focuses on mobile applications' usability, adapted GOMS operators were being used (see Table 1). Adaptations were made by Setthawong et al. (Setthawong & Setthawong, 2019) so that element completion time could be accurately calculated for touchscreen interactions.

Table 1: Updated GOMS operators for interactions with a touchscreen.

Code	Description	Time(s)
E	Prepare fingers	0.5
T	Touch screen with finger	0.2
TT	Touch screen twice with finger	0.4
D	Move finger over surface	0.5
M	Move finger to a direct part of screen	0.7
F	Move finger over surface rapidly	0.4
S	Pinch, squeeze, spread, or splay gesture	0.7
P	Touch screen with finger for a long time	1.1
R	Touch screen with 2 fingers and rotate	0.8
L	Release fingers	0.1

Evaluated results revealed that Radio-buttons ($M = 6.75$ s) and Spinners ($M = 8.48$ s) are the most promising in terms of date entry completion times (see Table 2). Therefore, these two as well as the

drop-down menu (the most popular in the industry) date entries are being investigated further.

Table 2: Average completion times estimated using GOMS.

Calendar	M (s)
Mathematical	12.53
Mathematical-aligned	12.53
Spinner	8.48
Date picker	11.03
Date picker with arrows	13.47
Slider	10.95
Radio-buttons	6.75

4.2 Procedure

For carrying out the experiment, an application was developed in which the users had to input the same 3 different dates using the calendars mentioned previously (drop-down menu, radio-buttons and spinners). The technologies used for the development were HTML, CSS and JavaScript. The IDEs used were Visual Studio Code and IntelliJ; in order to host the website Firebase was used, and it can be accessed by visiting <https://researchprototype.web.app/>. The Firestore service provided by Firebase was used for storing the data.

At first, a page with a small explanation of the experiment was shown, then the user was asked about the age and gender, for classification purposes. Afterwards, a page with videos regarding how to use each calendar was presented, in order to make the users more familiar with them and gain some expertise. Before starting the experiment the user was asked not to navigate back through the pages, as it would interfere with the validity of the data.

The experiment consisted in introducing the following dates: 5th of March of 2022, 16th of May of 1998 and 20th of June of 1964. These dates were hand-picked having in mind the objective of not repeating numbers, and also varying the closeness to the current date (one near, one far and one in between them).

The drop-down menu was implemented with a combo box for each field, and the other two calendars were implemented following the design shown previously, that is, for radio buttons (7) and for spinner (3). Technical settings, set for the spinner, where: $Accuracy = 0.001$, $MinStepInterval = 5$, $MaxStepInterval = 500$, $MaxSpeedAtSteps = 30$, $NoActionDistance = 0.01$, $OneStepDistance = 0.02$.

For storing the times, the time span between the selection of the first element of the calendar and the completion of the last one was calculated (the operation of pressing the button for proceeding to the

next date was not taken into account), which would be the time it elapses from the beginning to the end of the task.

After the user finishes with all the inputs, a questionnaire was shown, with the aim of measuring user satisfaction. It was composed by two questions that were also used in the study of Bargas et al. (Javier A. Bargas-Avila et al., 2011), and it followed a 5-point Likert scale (scale: 1 = Strongly disagree; 5 = Strongly agree): (1) Filling in the date was comfortable; (2) I could fill in the date quickly and efficiently. When the user completed this questionnaire, the data (which comprises the user age and gender, times, dates and answers to the questionnaire) was sent to the database and there also existed the possibility of sharing the prototype through several social media pages, to provide the experiment with as much reach as possible.

5 RESULTS

In order to provide more accurate results, the data was processed removing entries that did not fulfil the necessary requirements for being considered as valid. The data was required to have all the questions in the questionnaire answered, all the dates completed (but not necessarily correct, as the mistakes were later counted for further analysis), and be done on a mobile device, which was verified using the User Agent from the browser of each participant. Entries that did not fulfil all these requirements were deleted.

Data records were considered to be outliers if the date entry completion time differed from the mean value by three standard deviations (Javier A. Bargas-Avila et al., 2011). Outliers were erased from the data sample.

After data cleaning, the date entries' completion times were checked for normality. Shapiro-Wilk test was being executed. As normality was not noticed ($p < .05$), calendars' completion times were log-transformed. Comparisons of the calendars were done by analyzing the mean values of figures for three dates entered by the recipients. For all statistical methods, .05 alpha value was used.

5.1 Recipients

After deleting invalid data or outliers, 277 data records in total were gathered from the participants for further analysis. 183 of them were male and were 94 female, with ages ranging from 14 to 57 years old ($M = 23.25$, $SD = 0.26$). Recipients were invited to participate in the experiment via Social Media

platforms or by e-mailing current universities' students.

5.2 Errors: The Wrong Dates Entered

In order to test for significant differences between the errors of the three calendar versions, a linear regression model was used. Results showed no significant differences between the data entry elements, as $p > .05$. Even though, the drop-down menu showed the best results, with $M = 1.28\%$, $SD = 0.24\%$, while the radio-buttons performance proved to be the worst ($M = 3.16\%$, $SD = 2.06\%$) (see Table 3).

Table 3: Average and standard deviation of errors made.

Calendar	M (%)	SD (%)
Drop-down	1.28	0.25
Radio-buttons	3.16	2.06
Spinner	2.49	0.63

5.3 Completion Times

Data on date entries' completion times did not follow a normal distribution ($p < .05$), thus it was log-transformed. A Multiple Linear Regression (MLR) model was built. Equation:

$$\text{CompletionTime} \sim \text{CalendarTypeBinaryValue} + \text{userAge} + \text{userGender} \quad (1)$$

Three models were tested, comparing two different calendar versions each time. The results indicate significant differences between all calendars' completion times (see Table 5). The best performance was shown by Drop-down menu ($M = 6.31$ s, $SD = 1.76$ s), followed by Radio-buttons ($M = 8.83$ s, $SD = 2.88$ s). Spinner proved to be the slowest date entry element ($M = 13.25$ s, $SD = 3.81$ s), noting that the test was being performed on Spinner with the settings described in the Procedure section (see Table 4). Compared to the completion time results estimated using GOMS, experimental analysis proved to have higher time needs for entering the dates. Although, as in GOMS estimation, Radio-buttons calendar proved to show faster performance than Spinner.

Table 4: Statistics for calendars' completion times.

Calendar	M (s)	SD (s)
Drop-down menu	6.31	1.76
Radio-buttons	8.83	2.88
Spinner	13.25	3.81

5.4 User Satisfaction Rating

The first statement users had to evaluate was about date entry elements' comfortability. As shown in Table 6, significant differences were noticed between every calendar pair. Table 7 indicates mean and standard deviation values for user satisfaction. Drop-down menu demonstrated highest comfortability ratings ($M = 4.25$, $SD = 0.98$), whilst Spinner did not prove to be a preferred option ($M = 2.34$, $SD = 1.54$).

Accordingly, the drop-down menu's results showcase that according to the participants, the calendar is significantly faster and more efficient than the other versions ($M = 4.24$, $SD = 0.91$). Spinner is significantly slower and the least efficient ($M = 2.36$, $SD = 1.55$). Radio-buttons remain having a neutral evaluation ($M = 3.59$, $SD = 1.48$).

Table 5: Statistics for user satisfaction.

Calendar type	Comfortability		Speed/ Efficiency	
	M	SD	M	SD
Drop-down menu	4.25	0.98	4.24	0.91
Radio-buttons	3.67	1.38	3.59	1.48
Spinner	2.34	1.54	2.36	1.55

6 DISCUSSION

6.1 Discussion of the Findings

First, an overview of the results of each metric can be made to provide a perspective of the performance of each calendar.

Regarding completion time, the best option was the drop-down menu and the worst, by a considerable margin, was the spinner, which took more than double the time to complete compared to drop-down menu (the mean of drop-down menu was 6.31 seconds, whilst spinner had one of 13.25 seconds), while radio-buttons was almost 2 seconds slower to the drop-down menu.

User satisfaction was another important measure in the analysis, and the ranking in the methods follows the same order as with completion time, the drop-down menu was the preferred option, and the spinner the least preferred one, with radio-buttons being in the middle of them, closer to the drop-down menu than to the spinner (this order is maintained in both comfortability and efficiency, as evaluated by the users).

As for the errors made, it is worth noting that, despite not being significant differences between the three calendars in this regard (as shown in section 7.B), the most error-prone method was radio-buttons, and the least one was the drop-down menu.

The conclusions at which we can arrive using these results is that the drop-down menu is the fastest and most effective option. But this result may be conditioned by several factors, such as the other two calendars being more novel to the user, and also their implementations, which can vary more than the one of the drop-down menu, and the optimal implementation for the best user experience remains to be researched. Many users had operated with drop-down menus before and were already familiar with it, but for others they had to gain some more expertise, which could be done using the explanations provided in the webpage (see section 4.B). One benefit that all of the methods share is that no formatting errors can be made, as it is implicit in the design of the calendar.

6.2 Limitations

The results suggest that spinner is an option that should be avoided, due to it having very poor results, but it has to be taken into account that there are many variables that affect its usage, such as its sensitivity or speed. The configuration used for this experiment has proven not to be very effective, but using a more user-friendly one could give results closer to the ones predicted by GOMS (see section 4.A), by taking advantage of the possibility of making a swiping movement in mobile devices more effectively. The experiment consisted in filling several dates, with no other field to be completed, but this is not a very realistic scenario, as usually there are other components around, so the results could change if the calendar was integrated in a more conventional environment, such as a registration form in a website. The role of using the same dates for each calendar has to be considered as well, this could have had an effect on the expectations of the user over the course of the

experiment. As mentioned in the previous section, drop-down menu is a very popular option, and users were more familiar with it than with the other two alternatives, in order to deal with this, more dates could be used, so that the user can achieve the same level of training for each option.

6.3 Feedback

Several users provided some feedback after doing the experiment, with the goal of describing how their experience was and what they struggled with the most. Many users felt that the sensitivity in the spinner was too high, which suggests that for future research it should be lowered so that users can be faster and more accurate. Another user noted that the drop-down menu was more aesthetic compared to the radio-buttons, as it required much less space in the webpage. One common suggestion was to change the order of the years in the radio-buttons method, and start from the current year instead of the lowest one, as many users had to scan the page for more time and to scroll down in order to find the requested year.

6.4 Future Work

As for future related research, several things could be further tested in order to clarify even more which type of calendar is more suitable for mobile devices. Many of the calendars presented were not included in the experiment, as the GOMS estimation was used as a criterion for deciding which ones to include, but as this research has shown, the numbers are not exactly the ones predicted (see section 4.A), so it remains to be seen how they would perform in another experiment. Another line to be extended would be testing different configurations of the spinner in order to find the most efficient one for the users, and see how it would perform.

Table 6: MLR models' statistics for completion times.

I compared calendar	II compared calendar	Intercept coefficient	Calendar coefficient	T value	F statistic	P value
Drop-down menu	Radio-buttons	9.08	-0.32	-12.22	50.01 > F(3, 550)	< .001
Drop-down menu	Spinner	9.42	-0.74	-33.04	365.06 > F(3, 550)	< .001
Spinner	Radio-buttons	9.44	-0.42	-16.41	89.94 > F(3, 550)	< .001

Table 7: MLR models' statistics for user satisfaction.

		Compared calendars		
		Drop down menu and Radio-buttons	Drop-down menu and spinner	Spinner and Radio-buttons
Comfortability	Intercept coefficient	3.67	2.43	3.84
	Calendar coefficient	0.57	1.9	-1.33
	T value	6.21	19.86	-12.88
	F statistic	13.55 > F(3, 550)	131.54 > F(3, 550)	55.52 > F(3, 550)
	P value	< .001	< .001	< .001
Speed / Efficiency	Intercept coefficient	3.56	2.32	3.69
	Calendar coefficient	0.65	1.88	-1.23
	T value	7.02	19.85	-11.67
	F statistic	16.68 > F(3, 550)	131.77 > F(3, 550)	45.47 > F(3, 550)
	P value	< .001	< .001	< .001

ACKNOWLEDGEMENTS

This work was funded by the Department of Science, Innovation and Universities (Spain) under the National Program for Research, Development and Innovation (Project RTI2018-099235-B-I00).

REFERENCES

- Bargas-Avila, J.A., Brenzikofer, O., Roth, S. P., Tuch, A. N., Orsini, S., & Opwis, K. (2010). Simple but Crucial User Interfaces in the World Wide Web: Introducing 20 Guidelines for Usable Web Form Design. User Interfaces. InTech.
- Bargas-Avila, Javier A., Brenzikofer, O., Tuch, A. N., Roth, S. P., & Opwis, K. (2011). Working towards usable forms on the world wide web: Optimizing date entry input fields. *Advances in Human-Computer Interaction*, 2011.
- Beaumont, Andy. (2002). Usable forms for the Web : includes fully usable code samples in HTML, Flash, JavaScript, ASP, PHP, and more : all code written for PC IE 4+, Netscape 4+, and Opera 5+, Mac IE 5+, Netscape 6 and Opera 5+. Glasshaus. Retrieved June 23, 2022, from <https://link.springer.com/book/9781590591604>
- Brown, A., Jay, C., & Harper, S. (2010). Audio Access to Calendars. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)-W4A '10*. New York, New York, USA: ACM Press. Retrieved June 2, 2022, from <http://www.w3.org/WAI/intro/aria.php>
- Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*.
- Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007). Helping respondents get it right the first time: The influence of words, symbols, and graphics in Web surveys. *Public Opinion Quarterly*, 71(1), 113–125.
- Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS Analysis for Predicting and Explaining Real-World Task Performance. *Human-Computer Interaction*, 8(3), 237–309.
- Healey, B. (2007). Drop Downs and Scroll Mice The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys. *Social Science Computer Review*, 25, 111–128. Retrieved June 17, 2022, from <http://ssc.sagepub.comhttp://online.sagepub.com>
- Jensen, E. T., Hansen, M., Eika, E., & Sandnes, F. E. (2020). Country Selection on Web Forms: A Comparison of Dropdown Menus, Radio Buttons and Text Field with Autocomplete. *Proceedings of the 2020 14th International Conference on Ubiquitous Information Management and Communication, IMCOM 2020*. Institute of Electrical and Electronics Engineers Inc.
- John, B. E., Kieras, D. E., & Kieras, ; D E. (1996). The GOMS family of user interface analysis techniques. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(4), 320–351. ACM PUB27 New York, NY, USA . Retrieved June 30, 2022, from <https://dl.acm.org/doi/abs/10.1145/235833.236054>
- Label Placement in Forms :: UXmatters. (n.d.). Retrieved June 23, 2022, from <https://www.uxmatters.com/mt/archives/2006/07/label-placement-in-forms.php>
- Linderman, Matthew., Fried, J., & 37signals (Firm). (2004). *Defensive design for the Web : how to improve error*

- messages, help, forms, and other crisis points, 246. New Riders.
- Martin, L. (2013). Estimating Time Requirements for Web Input Elements. *Proceedings of the 2nd International Workshop on Web Intelligence (WEBI-2013)*, 13–21.
- Martin, L. (2014). A RESTful web service to estimating time requirements for web forms. *International Journal of Knowledge and Web Intelligence*, 5(1), 62. Inderscience Publishers.
- Mobile vs. Desktop Traffic Market Share [May 2022] | Similarweb. (n.d.). . Retrieved June 23, 2022, from <https://www.similarweb.com/platforms/>
- Most Visited Websites - Top Websites Ranking for May 2022 | Similarweb. (n.d.). . Retrieved June 23, 2022, from <https://www.similarweb.com/top-websites/>
- Seckler, M., Heinz, S., Bargas-Avila, J. A., Opwis, K., & Tuch, A. N. (2014). Designing usable web forms- Empirical evaluation of web form improvement guidelines. *Conference on Human Factors in Computing Systems - Proceedings*, 1275–1284. Association for Computing Machinery.
- Setthawong, P., & Setthawong, R. (2019). Updated Goals, Operators, Methods, and Selection Rules (GOMS) with touch screen operations for quantitative analysis of user interfaces. *International Journal on Advanced Science, Engineering and Information Technology*, 9(1), 258–265. Insight Society.
- Türkcan, A. K., & Onay Durdu, P. (2018). Entry and selection methods for specifying dates in mobile context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10903 LNCS, 92–100. Springer Verlag.
- Wroblewski, L. (2008). *Web Form Design: Filling in the Blanks*. (M. Justak, Ed.). New York: Rosenfeld Media.

I Wandered Lonely in the Cloud: A Review of Loneliness, Social Isolation and Digital Footprint Data

Dominic Reedman-Flint¹, John Harvey¹^a, James Goulding¹^b and Gary Priestnall²

¹N/LAB, Nottingham University Business School, University of Nottingham, U.K.

²School of Geography, University of Nottingham, U.K.

Corresponding author: john.harvey@nottingham.ac.uk

Keywords: Social Isolation, Loneliness, Human-Computer Interaction, Digital Trace Data.

Abstract: The harm that social isolation and loneliness can have on physical, mental, and emotional well-being is now well evidenced. With social distancing and remote working now commonplace, the dangers of loneliness are ever more acute. Consequently, information technologies have taken on renewed importance to support healthy communication and reduce the negative impacts of social isolation. However, existing literature remains highly conflicted as to the relationship between technology use and its impact on loneliness. This is perhaps understandable: measures of loneliness have traditionally been examined within clinical settings, far removed from the everyday realities of computational interactions. Yet data logged about such interactions now offers potential to help identify isolation and loneliness and support those experiencing resulting health issues. We present a scoping review of this domain, focusing on detection of loneliness and social isolation through digital data. We interrogate a corpus of published articles from the HCI literature, identifying a series of methodological, epistemological, and ethical tensions therein, as well as emerging opportunities for future empirical study. We identify a need to examine such phenomena via actual behavioural data, rather than reliance on historical proxies such as age and gender, to help modernize our understanding of this growing social ill.

1 INTRODUCTION

"The world is the closed door. It is a barrier. And at the same time it is the way through. Two prisoners whose cells adjoin communicate with each other by knocking on the wall. The wall is the thing which separates them but it is also their means of communication. ... Every separation is a link."

The quote above by philosopher Simone Weil (1997) draws on Plato's concept of *metaxu*, meant here as something that both separates and connects simultaneously. It was originally intended to describe the challenge of communion with God, but its metaphor of a prison wall also serves as a useful analogy for how contemporary researchers describe technology and its effect on communication between people more generally. Recent works summarising loneliness and online social interaction have shown how information technologies, and in particular

interconnected digital devices, appear in the literature as both the cause of, and solution to, the growing issue of loneliness (Dunbar, 2021; Hertz, 2021). Much of the literature is therefore conflicted, illustrating that the causal relationship between the two categories is both complex and likely bi-directional; such that the aetiology of loneliness is at least partially context dependent. Technology transforms social relations, and in doing so, creates new opportunities for alienation and communion alike.

Despite the complexity involved in potential manifestations of loneliness, research conducted by social neuroscientists (Cacioppo and Patrick, 2008) provides scientific evidence that loneliness causes physiological events that wreak havoc on our health. Persistent loneliness leaves a mark via stress hormones, immune function and cardiovascular function (Knox and Uvnäs-Moberg, 1998) with a cumulative effect that brings health outcomes similar

^a <https://orcid.org/0000-0003-4188-1900>

^b <https://orcid.org/0000-0002-8892-6398>

to being a smoker (Holt-Lunstad *et al.*, 2010). Furthermore, loneliness exacerbates the risk of experiencing additional subsequent problems and altering behaviour insofar as it increases the likelihood ‘of indulging in risky habits such as drug taking and plays a role in mental disorders such as anxiety and paranoia’ (Griffin, 2010). Despite this, many people experiencing loneliness do not interface with medicalized settings; and few receive clinical diagnosis or support. Due to the hidden nature of the problem, estimates of loneliness are often absent from national statistics despite potential to inform social policy. As a result, researchers have often turned to demographic proxies for loneliness, with much work throughout the 20th century seeking to theorise the phenomena, and the nature of those people predisposed to it. In particular, the elderly (over 65s) and disabled, widely cited in academia as being most affected, underpin most analyses of loneliness prevalence. Yet the UK’s Office of National Statistics (ONS, 2018) and BBC, via the world’s largest empirical study of loneliness (BBC, 2019), recently showed that loneliness is far from constrained to these communities; and affects the population in far more ways than previously recognised

Recently, the issue has been further exacerbated by Covid-19, and our increasing exposure to conditions conducive to loneliness (Bu *et al.*, 2020). Dramatic societal change, lockdowns, and enforced social distancing has spawned a flurry of research to help identify those most at risk of loneliness and its myriad health consequences (Ernst *et al.*, 2022). Yet, the mass transformation of work and socialization into increasingly online settings has also created opportunity for researchers: digital identification of those at risk. Whether in cases where digital technology exacerbates loneliness or in situations it diminishes it, human-computer interactions leave behind a rich corpus of digital logs. Such digital footprint data, if handled responsibly, can offer new avenues to help identify, characterize, and intervene in this “wicked problem”. The development of big data and ubiquitous digital architectures to log everything from library records, transport movements, health records, social interactions and even shopping habits illustrates the wide range of siloed information available; data that might be harnessed to better understand vulnerability to loneliness and to direct social policy accordingly. With such potential, of course, comes risk: approaches that depend on indirect, closed, or proprietary data sources introduce new methodological, ethical, and privacy challenges, otherwise absent in clinical settings.

The overall aim of this work is therefore to establish the contexts in which digital data are being used when identifying people experiencing social isolation or loneliness, or where data are being inadvertently created by the interactions of those people experiencing loneliness (particularly those who are doing so in populations not typically recognised within clinical settings). Unlike review papers that have attempted to map the efficacy of digital technologies as loneliness interventions from a public health perspective (Shah *et al.*, 2019; Ibarra *et al.*, 2020), we provide a scoping review of how social isolation and loneliness interventions are already conceived, discussed, and enacted within the computing literatures. By examining epistemological and methodological bases of previous research we aim to expose some of the underlying intellectual commitments and assumptions currently being made in the field and contextualize ongoing debate. We first outline the methodology and scoping approach used to achieve this, before discussing emerging genres in the field and the key analytical differences that separate researchers. Finally, we discuss emerging issues and empirical gaps in the domain that, if addressed, can serve as the basis for addressing this phenomenon.

2 METHODOLOGY

While ‘loneliness’ and ‘social isolation’ are well accepted terms in everyday usage, definitions in academic contexts vary significantly by scientific discipline and empirical focus. Significant existing work has sought to characterise loneliness and social isolation through study of their respective aetiologies from a public health perspective (Holt-Lunstad, 2017; Elovainio *et al.*, 2017; Steptoe *et al.*, 2013; Lubben, 2017). Various constructs are described in the literature suggesting interconnectedness between the conditions; and in much work the terms ‘loneliness’ or ‘social isolation’ are used interchangeably - particularly in research which does not aim to conceptualise the respective concepts. Yet there is limited consensus here, with other research viewing the conditions as entirely distinct and to be considered independently (Matthews *et al.*, 2016); and in psychological and clinical research it is far more common to differentiate the constructs. Social isolation is typically described as the circumstance of being physically alone or otherwise detached from contact with friends, family, or society; Loneliness, in contrast, is described as a negative psychological response to such situations, commonly portrayed as

‘a subjective, unwelcome feeling of lack or loss of companionship... [occurring] when we have a mismatch between the quantity and quality of social relationships that we have, and those that we want’ (Perlman and Peplau, 1981). Academic definitions are contested for both terms, with some authors speaking of multiple sub-types of loneliness (Weiss, 1975), yet a separation is commonly recognised between them. As our intention is to synthesize literature around these concepts, we do not challenge such definitions - while the overlapping of language creates challenges for comprehensive summarisation of the field, it also highlights the need for consilience and transdisciplinarity to support further advances.

Archetypal definitions of loneliness and social isolation primarily came from diagnostic scales developed in the latter part of the 20th century. Weiss (1975) saw ‘social loneliness’ as a lack of or negative change in social connections below a desired level, whereas ‘emotional loneliness’ was a lack of deep, meaningful (i.e. romantic or familial) connection. This multidimensional approach to the study of loneliness is not, however, reflected in the widely used UCLA Loneliness Scale (Russell, 1996) which treats loneliness as a unidimensional construct. While this approach has been contested (Marangoni and Ickes, 1989), Russell et al. (1984) have argued that, despite different forms existing, that the UCLA scale adequately summarises these different loneliness states. UCLALS is the most widely used measurement for loneliness, but several alternative well-cited measures exist. Almost uniformly, however, these predate widespread adoption of the Internet and digital devices - whether de Jong Gierveld’s Scale (De Jong-Gierveld, 1987), the Social and Emotional Loneliness Scale for Adults (DiTommaso and Spinner, 1993), Differential Loneliness (Schmidt and Sermat, 1983) or the Loneliness Rating Scale (Scalise et al, 1984).

One of the key, outstanding questions for digital identification of loneliness, is therefore the extent to which present-day research should revisit such measures, given the rapid integration of technology into our daily lives over the past two decades (e.g., mobile private messaging, social networking, and livestreamed video). The impact the digital world has had on our experience of loneliness and social isolation remains unclear; do traditional metrics still hold; and to what extent have traditional proxies become anachronistic? Loneliness has predominantly been identified in clinical settings through surveys and self-diagnosis, or via direct medical assessment. Do such environments overlook people in need of support but who do not seek out practitioners, either

being unable or stigmatised from doing so? To consider questions of this nature, Munn et al. (2018) have advanced scoping reviews as useful tools to examine emerging evidence; especially when it is still unclear what other, more specific questions can be valuably posed and addressed by future empirical work. Such reviews aim to not only report on evidence that informs practice in the field, but to consider the way research has been conducted, and in particular ‘in contrast to traditional literature reviews scoping reviews are informed by an *a priori* protocol; Are systematic; Aim to be transparent and reproducible; and ensure data is extracted and presented in a structured way’ (Munn et al., 2018).

This scoping review focuses on two databases containing peer-reviewed papers in computing and its associated sub-disciplines, the Association for Computing Machinery (ACM) and the IEEE Xplore (IEEE) digital libraries respectively. Both libraries were searched, isolating abstracts containing the terms ‘loneliness’ or ‘social isolation’. This produced a total of 401 results, and the resulting corpus was screened for non-English, inaccessible, or non-peer reviewed articles or conference papers. Each article was then reviewed to ensure that social isolation or loneliness was not peripheral but a relevant empirical or conceptual focus of the work. Papers either: (1) identified social isolation or loneliness as part of their sampling procedure; (2) used social isolation or loneliness as a dependent or independent variable within analysis; or (3) social isolation or loneliness was specifically being explored or researched through an inductive or conceptual approach. This process is summarised via the PRISMA diagram in Figure 1. The final output corpus yielded 52 articles, which were taken forward to fine-grained analysis. Each article was decomposed, and re-summarised into a short commentary before then being analysed in relation to a series of methodological questions. Answers to these questions were tabulated, serving as the basis thematic identification discussions between the research team, who collaboratively identified genres and epistemological commitments made by the works.

The methodological research questions considered are: (1) Is the work empirical (data gathering) or conceptual? (2) Data provenance - if data are gathered what is collected and how? (3) Sampling - who is the focus of the research / who is thought to be lonely? (4) Does the study aim to identify people experiencing issues, intervene to help people, or to conceptualize what loneliness is? (5) Is loneliness or social isolation measured and if so, how; is it a pre-existing measure or newly developed? (6)

Is an expert identified within labelling (or not); who gets to identify those being labelled lonely?; (7) What methods were used to analyse loneliness data generated? (8) what metadata is likely to exist as part of the measures chosen e.g., social network data, spatial data, or time series data; and (9) What constitutes success and who are the beneficiaries? The answers to these questions were then used to (a) identify genres of work with clustered formulations of loneliness problems/challenges addressed via digital technology; (b) to examine contrasting methodology and epistemology of studies in the field, and (c) to indicate promising avenues for future research.

3 ANALYSIS

52 papers were included in the final corpus. Figure 1 shows the earliest three papers from the corpus (Zhang, 2009; Zhou et al., 2009; Waterworth and Ballesteros, 2009) appeared in 2009, with a gradual annual increase until 2020, which saw 12 relevant papers published (search conducted 04/2021). 37 articles used established loneliness measures, of which 16 used the UCLA loneliness scale. The vast majority of papers that measured loneliness or social isolation directly used survey questions using Likert Scales or specific protocols (e.g., UCLA or De Jong Gierveld scales). Those papers that did not use established measures did so primarily due to an inductive or exploratory qualitative focus, or because of emphasis on intervention development (e.g. use of robots) rather than intervention assessment.

A striking feature of the corpus is that two specific demographics dominate the empirical work (the elderly – 19 papers; and student populations – 9 papers). Such is their emphasis within the corpus that in the following section we examine each of these categories in turn as distinct genres. It is, however, worth noting that this bimodal split is based upon a priori demographic sampling decisions rather than selections made due to indicators of loneliness or social isolation. Most students graduate in their early twenties, whereas retirees (particularly those in nursing homes, a common setting of loneliness research) normally exceed 65 years. This >40-year gap reflects a large section of the populace absent from research. Within the corpus 15 articles explicitly sought to develop methods to identify lonely or socially isolated people. 12 papers focused explicitly on loneliness or social isolation interventions. The remainder contained a mixture of inductive, exploratory, and hypothetico-deductive

approaches in which loneliness or social isolation featured.

In the following section, we describe genres resulting from the analysis. Genres are structured from clusters of research that frame problems of loneliness, isolation, and digital technology in similar ways. Given the recent emergence of much of the work and the inevitable overlapping of some themes within papers the genres should not be thought of as internally consistent movements to which the authors are aligned, rather the aim here is to illustrate similarity of agendas and offer a lens through which some of the key debates can be seen. There are five identifiable genres within the corpus that describe all but one paper (related to testing a loneliness scale in Italy (Senese et al., 2020), albeit unrelated to digital footprints); these are now discussed in turn.

Genre 1: The Elderly

18 papers contained an explicit focus on loneliness or social isolation in the elderly, making it the largest genre in the corpus (Broadbent et al. 2018; Eldib et al., 2015; Yang and Bath, 2018; Pedell et al., 2010; Baecker et al., 2014; Light et al., 2017; Zadeh et al., 2020; Martinez et al., 2017; Austin et al., 2016; Ring et al., 2013; Ha and Hoang, 2017; Bacciu et al., 2016; Noguchi et al., 2018; Chang and Kalawsky, 2017; Yoshida et al., 2018; Mulvenna et al., 2017; Petersen et al., 2013; Waterworth et al., 2009). In these, older people are identified as an at-risk group due to the intersection of multiple life events - for example: leaving the workforce through retirement; losing family through bereavement; having children leave home; or being forced into care homes for health reasons. In each case, elderly people see transformations of their social networks, losing opportunities for meaningful social interaction. Such features of ageing are well known; and because elderly populations often tend to have limited geospatial mobility and higher chances of interfacing with medicalised settings, visibility of loneliness and isolation is increased, allowing for more easily targeted interventions. It is therefore understandable that this genre also features the highest relative share of interventions detailed within the corpus. In clinical practice, algorithmic risk-based estimates of loneliness in the elderly are likely to perform well, particularly where personal data is available and where trust in the robustness and explainability of the method can be secured. Indeed, recent work showcases machine-learning approaches already yielding reliable performance (Yang and Bath, 2018). However, current risk-based methods typically depend on presentation or referral within a clinical setting.

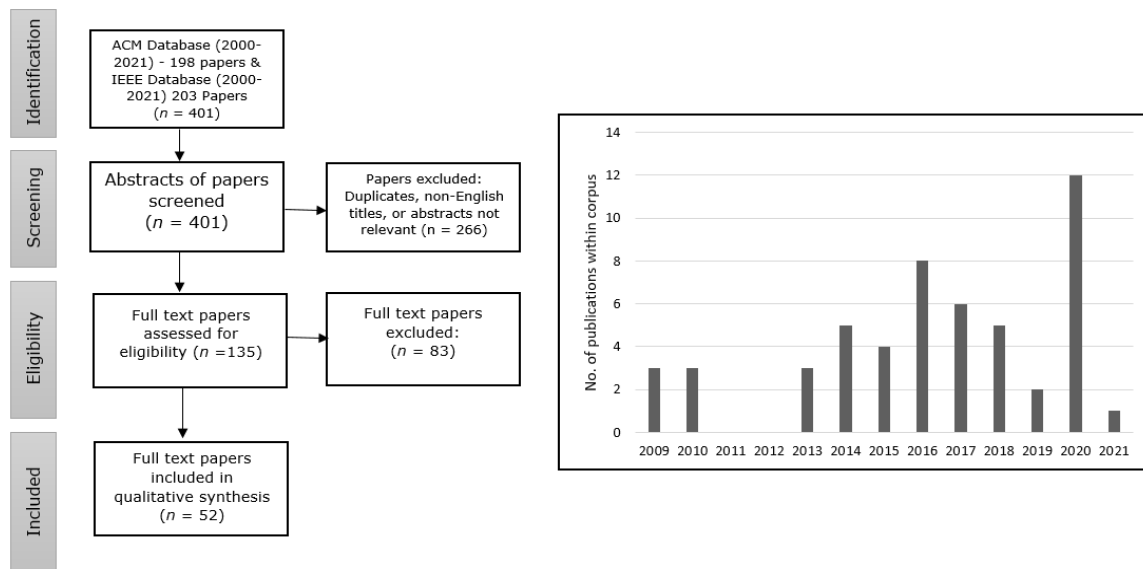


Figure 1: (Left) Review Screening Process; (Right) Yearly breakdown of publications within final corpus.

Sampling issues of this nature raise concerns about potentially under-served sub-populations (e.g., those living alone), not evident in clinical settings; and who may be more effectively identified via non-traditional means (e.g., via digital footprint data).

Genre 2: The Student Experience

Academics have historically been criticised for an overreliance on sampling students as part of psychological and clinical research, but in this instance, the need for research on the experience of loneliness within student populations is of clear relevance. Students often move cities to attend university, with a sizeable minority moving internationally to an unfamiliar place to live and work amongst unfamiliar people. A range of papers use students as an explicit empirical focus or do so implicitly by only sampling from student populations (Fuentes et al., 2016; Joyner et al., 2020; Zhou et al., 2020; Zhang, 2010; Kindness et al., 2013; Lu and Yao, 2010; Zhou et al., 2009; Xu et al., 2015; Ferrer et al., 2020). Migration is a key cause of loneliness for this cohort. Students can be made lonely and isolated via their own movement, which fragments prior social networks (in hope of the formation of new ones), while limiting familial support structures. The methodological tendency within this genre tends towards consideration of online data sources useful for identification of mental health problems, whereas for the elderly genre research methods tend to focus on interventions, and data generated by physical technology in situ. Most papers in the corpus leverage established survey measures to identify loneliness;

but only three papers seek to identify loneliness as a dependent variable with the remainder focusing more on interpersonal relationships and wellbeing. One article (Zhou et al., 2020) notes, for instance that ‘Previous studies on loneliness have mainly focused on using questionnaire-based loneliness scales, e.g., UCLA scale, for the measurement of loneliness. Nevertheless, the lonely may prevent reporting their real conditions since they are afraid of information disclosure, discrimination, unfair treatment, thus it makes the information accumulated by these questionnaires unreliable.’ The challenge of reliable sampling frames again raises the possibility to supplement traditional approaches using alternative health surveillance methods (particularly those that examine social network or communication data - e.g., classroom collaboration network data). Incorporating behavioural observation data with surveys is likely to help scale prevalence estimates of loneliness across broader populations (behavioural vs demographic proxies) although risk-based probabilistic approaches are better suited to population prevalence estimation than deterministic assessments used in clinical practice. There is likely a useful bifurcation in data and methods used for population and individual level assessment worthy of further inquiry, particularly due to the sensitive nature of social network data.

Genre 3: Online Services, the Web, and Apps as Windows into the Experience of Loneliness

A growing number of papers (Brueckner, 2020; De Choudhury et al., 2014; Joseph et al., 2014; Weinert et al., 2014; Jeong et al., 2015; Taylor et al., 2017;

Burke et al., 2010; Ananto and Young, 2021; Jeong et al., 2016; Galunder et al., 2018; Pulekar and Agu, 2016; Rabani et al., 2020; Kaur et al., 2020; Wu et al., 2016; Lu and Yao, 2010) draw on diverse data sources being used from popular Web and Smart Device applications such as (messaging, phone calls, web browser logs, gaming, teleworking applications, Facebook, Instagram, Twitter, Grindr, and Amazon Kindle) to explore loneliness within the respective user populations. 7 articles in the corpus used established survey measures, of which 5 applied the UCLA survey. This genre is characterised by big data and behavioural analytics approaches, with commercial Web services often being repurposed for prosocial reasons. For example Gao et al. (2019) correlate usage of a dating website with loneliness measures, regressing features engineered from profiles, postings, and check-ins with ‘surveys [of] psychological states by four professional questionnaires measuring different kinds of mental disorders: depression, loneliness, anxiety, and stress’. Mixed method approaches of this nature, that combine structured data (e.g. clinical surveys) with wide-ranging and often unstructured social network data, hold promise for prevalence risk estimates within their user population; though are likely to introduce potentially hidden forms of survivorship bias that require triangulation within a broader population before they could be validated. The linking of social network usage (both time spent, words/sentiments used, and people contacted) is a common area of study amongst this genre. An illustrative example can be seen in a study (Kaur et al., 2020) of over 140 million tweets to analyse personality insight, emotion, and sentiment analysis in relation to social isolation during the pandemic. The premise is a straightforward one: how we describe ourselves and who we interact with is liable to change throughout our lives, with such transitions being mirrored in our digital footprints online. If this is true, then as the authors note, aggregated results might be used to support ‘a public health indicator to anticipate the possibility of social isolation and design health policies accordingly.’ However, despite showing initial promise, further work is needed to evaluate the efficacy of such claims, particularly where approaches can be replicated alongside robust clinical measures.

Genre 4: Physical Technology, Robots, and Anthropomorphic Interactions

A series of papers (Lazányi, 2016; Eyssel and Reich, 2013; Lou et al., 2019; Li et al., 2020a; Li et al., 2020b) focus on the creation or delivery of physical

resources, often robotic interventions, to help alleviate social isolation and/or loneliness. The focus of these papers is primarily development of new technology rather than identification of loneliness. The genre emphasises intervention (or what has been called ‘prosocial interaction design’ – Harvey et al., 2014), and specifically focuses on transforming the physical environment around people likely to experience loneliness. However, unlike interaction design strategies that aim to pair people with other people to foster communion and thus remove loneliness, a far greater emphasis is placed on non-human subjects and mastery of anthropomorphism. The papers tend to focus on dyadic solutions to loneliness i.e., creating a surrogate partner (either high fidelity humanoids or non-humans such as pets with human-like features – Lou et al., 2019) with whom a person can form meaningful attachment. Though Eyssel and Reich (2013) found that people experiencing loneliness may be more likely to anthropomorphise robots, Li et al. (2020a; 2020b) find that the relationship is not so clear and that a more nuanced understanding of what loneliness is can better serve to predict efficacy of interventions, specifically through reference to ‘trait’ versus ‘state’ loneliness. Whilst this concept - effectively representing chronic and temporary loneliness - is not unique in clinical literature, it is within the papers identified within this corpus and suggests distinct data sources could be considered for more general measures. This genre points to the idea that multidimensional measures of loneliness in the digital world may help to develop subsequent interventions, it is therefore worthy of further research, particularly where new measures account for online experience and can be paired with longitudinal clinical outcomes.

Genre 5: Edge Communities

The smallest identifiable genre, covering 4 papers, considers differing edge communities - people living in more extreme physical or mental conditions, including refugees (Almohamed and Vyas, 2016), mental health patients (Bearse et al., 2020), cancer patients (Jacobs et al., 2015), ‘seafarers’, remote communities in Greenland, and welfare claimants in Northeast England (Jensen et al., 2020). Though these groups represent relatively small sub-sets of the broader population, they each nonetheless exhibit distinct behavioural characteristics which might prove useful in 1. generalized identification; and 2. understanding variance across sub-populations. All are characterised by transience i.e., people not expected to remain in a lonely state in the long term. They are, as Jensen et al. (2020) note, going through

‘digital liminal’ states and thus the data generated by these people is likely to experience sharp phase transitions. Given these demographics remain more likely to experience loneliness at some point than the general population they serve as excellent case studies for designing health surveillance to understand broader prevalence statistics. Further research is required to identify how digital footprints align across liminal populations and whether metadata can be ethically obtained, especially given the precarious lives these groups experience.

3.1 Epistemological and Methodological Tensions

Tension 1 - Sampling and Exclusion: As noted earlier, research into loneliness and social isolation has to date heavily emphasised student and elderly populations. Some exceptions exist, but children and those in the range 21-60 have seldom been considered as research subjects for digital footprints and loneliness – a clear gap in the domain, made all-the-more pressing due to changes in daily working environments since COVID-19. Such an omission is partly for good reason: students and the elderly are stable populations, relatively accessible, and with well reported challenges. Nonetheless, the corpus excludes a huge portion of the average human life course. Loneliness and social isolation do not, of course, act across such neat demographics in practice. A solution may be to develop encompassing prevalence estimates derived from digital footprint data.

Tension 2- Validity of Loneliness Measures: The use of traditional scales, for example the UCLA or De Jong Griefveld scales, may be anachronistic with the forms of loneliness and social isolation occurring in the modern world. Digital technology is enabling new forms of rich and multi-faceted communication that do not depend on being in the same place at the same time. While the number of “contacts” we maintain has increased due to technological innovations, has this impacted on the shared social experience that prevents loneliness? And is this divergence between isolation and loneliness represented in the metrics currently used? As most influential scales were created prior to widespread Internet and smartphone adoption they have little to say about the aspects of social life now integral to work, play, and socialisation. The notion of mixed modalities of loneliness is something many authors note, but it is rarely studied due to the absence of standardised measures.

Tension 3 – Clinical versus Non-clinical Populations and Non-overlapping Data Sets: This tension is best illustrated by the contrast between genre 1 (the elderly) and genre 3 (online services and their associated big data). In the former, researchers have excellent access to a population known to be more likely lonely and who also likely frequent clinical settings. This demographic is therefore easier to identify through formal methods, but they are likely to have sparser digital footprints when compared with the broader populace that could be used for prevalence estimates. In contrast, online services such as social media have rich behavioural data illustrative of the interactions of lonely people and often these data cut across multiple demographics but pairing clinical measures is harder from a methodological and ethical perspective. Here, again care must still be taken to avoid exclusionary bias. Machine learning models retain biases inherent to the datasets they are trained with - and the availability of digital footprint data (or lack thereof) in representative communities such as the elderly must be considered. As such hybrid approaches seem a sensible solution – there is no one size fits all.

4 CONCLUSIONS

Via this scoping review, several tensions, research gaps and opportunities have been evidenced. Unexpectedly, it is gaps in the current literature, that potentially offer the most insight, indicating the impact digital footprints might have in improving loneliness prevalence estimation and modernizing its characterization. If leveraged responsibly, behavioural datasets promise to advance understanding for public health and policy and help augment a domain that has historically been forced to rely on coarse proxy data such as age and general or limited clinical records. Social network, mobility, and behavioral data all hold potential to support more-encompassing prevalence statistics of loneliness. Yet our scoping analysis highlights several challenges and tensions therein, issues that require continued scholarly attention. The following opportunities, if invested in, may help attend to these challenges, consolidate knowledge and advance maturity in the field:

Opportunity 1 – Sampling - The Need for Longitudinal Study through the Life Course: To make use of digital footprints researchers need to ensure validity and efficacy, through ‘ground truths’ - labelling of loneliness in individuals, that can be

paired to rich, observed behavioural data to determine both indicators and antecedents. This is especially relevant for middle-aged populations, who occur far less in clinical contact records. Furthermore, inter- and intra- person reliability measurements for loneliness surveys are rarely conducted yet would be highly valuable when pairing longitudinal digital footprints for the purpose of identification and risk-based prevalence estimates.

Opportunity 2 – Validity - Studying Digital Loneliness as its Own Experience: distinct from the loneliness measured and identified via surveys such as the UCLA, the field should encourage research of extended multi-dimensional measures and identification tools to recognise new experiences across mixed modalities. Our contact lists are ever extending – however it is a lack of socially-shared interactions which may be at the root of modern experiences of loneliness.

Opportunity 3 – Make Data Useful, Open and Transparent: As evidenced in Genre 3, much of the data being studied by researchers is proprietary, closed, and often commercial in nature. Notwithstanding privacy concerns the need for open data is a pre-requisite if the field is to develop. In addition, only 3 papers include any spatial data, this is surprisingly low because mobility/movement tracking is often cited as a solution for loneliness monitoring (particularly in the elderly). Broader engagement of relevant communities, co-creation of research studies, and focus on initiatives that engage sufferers through open and transparent data sharing, are required not only if we are to model loneliness effectively – but if we are to generate practical interventions from model explanations, with real-world impact. Kurt Vonnegut once reportedly said *‘What should young people do with their lives today? Many things, obviously. But the most daring thing is to create stable communities in which the terrible disease of loneliness can be cured.’* To this we might add, identifying and supporting those in need is a first key step.

ACKNOWLEDGEMENTS

This article is dedicated to the memory of the primary author who sadly died near completion of the manuscript. Dr Reedman-Flint was supported by the Horizon Centre for Doctoral Training at the University of Nottingham (UKRI Grant No. EP/L015463/1)

REFERENCES

- Almohamed, A., & Vyas, D. (2016). Designing for the Marginalized: A step towards understanding the lives of refugees and asylum seekers. In Proceedings of the 2016 acm conference companion publication on designing interactive systems (pp. 165-168).
- Ananto, R. A., & Young, J. E. (2021). We can do better! an initial survey highlighting an opportunity for more HRI Work on loneliness. In Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (pp. 457-462).
- Austin, J., Dodge, H. H., Riley, T., Jacobs, P. G., Thielke, S., & Kaye, J. (2016). A smart-home system to unobtrusively and continuously assess loneliness in older adults. *IEEE journal of translational engineering in health and medicine*, 4, 1-11.
- Baecker, R., Sellen, K., Crosskey, S., Boscart, V., & Barbosa Neves, B. (2014). Technology to reduce social isolation and loneliness. In Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility (pp. 27-34).
- Bacciu, D., Chessa, S., Ferro, E., Fortunati, L., Gallicchio, C., La Rosa, D., ... & Vozzi, F. (2016). Detecting socialization events in ageing people: The experience of the doremi project. In 2016 12th International Conference on Intelligent Environments (IE) (pp. 132-135). IEEE
- BBC (2019) BBC Radio 4 - The Anatomy Of Loneliness - Who Feels Lonely? [Online] Available at: <https://bbc.in/3NOjIey> [Accessed 15 May 2021].
- Bearse, P., Manejwala, O., Mohammad, A. F., & Haque, I. R. I. (2020). An Initial Feasibility Study to Identify Loneliness Among Mental Health Patients from Clinical Notes. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 68-77). IEEE.
- Broadbent, E., Ahn, H. S., Kerse, N., Peri, K., Sutherland, C., Law, M., ... & Pandey, A. K. (2018). Can robots improve the quality of life in people with dementia?. In Proceedings of the Technology, Mind, and Society (pp. 1-3).
- Brueckner, S. (2020). Captured by an Algorithm. In ACM SIGGRAPH 2020 Art Gallery (pp. 458-459).
- Bu, F., Steptoe, A., & Fancourt, D. (2020). Who is lonely in lockdown? Cross-cohort analyses of predictors of loneliness before and during the COVID-19 pandemic. *Public Health*, 186, 31-34.
- Burke, M., Marlow, C., & Lento, T. (2010). Social network activity and social well-being. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1909-1912).
- Cacioppo, J. T., & Patrick, W. (2008). Loneliness: Human nature and the need for social connection. WW Norton & Company.
- Chang, A. S., & Kalawsky, R. S. (2017). Future configurable transport for the ageing population. In 2017 7th International Conference on Power Electronics Systems and Applications-Smart Mobility, Power Transfer & Security (PESA) (pp. 1-5). IEEE.

- De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 626-638).
- de Jong-Gierveld, J. (1987). Developing and testing a model of loneliness. *Journal of personality and social psychology*, 53(1), 119.
- DiTommaso, E., & Spinner, B. (1993). The development and initial validation of the Social and Emotional Loneliness Scale for Adults (SELSA). *Personality and individual differences*, 14(1), 127-134.
- Dunbar, R., 2021. *Friends: Understanding the power of our most important relationships*. Hachette UK.
- Eldib, M., Deboeverie, F., Haerenborgh, D. V., Philips, W., & Aghajan, H. (2015). Detection of visitors in elderly care using a low-resolution visual sensor network. In *Proceedings of the 9th International Conference on Distributed Smart Cameras* (pp. 56-61).
- Elovainio, M., Hakulinen, C., Pulkki-Råback, L., Virtanen, M., Josefsson, K., Jokela, M., ... & Kivimäki, M. (2017). Contribution of risk factors to excess mortality in isolated and lonely individuals: an analysis of data from the UK Biobank cohort study. *The Lancet Public Health*, 2(6), e260-e266.
- Ernst, M., Niederer, D., Werner, A. M., Czaja, S. J., Mikton, C., Ong, A. D., ... & Beutel, M. E. (2022). Loneliness before and during the COVID-19 pandemic: A systematic review with meta-analysis. *American Psychologist*.
- Eyssel, F., & Reich, N. (2013). Loneliness makes the heart grow fonder (of robots)—On the effects of loneliness on psychological anthropomorphism. In *2013 8th acm/ieee international conference on human-robot interaction (hri)* (pp. 121-122). IEEE.
- Ferrer, M., Mancha, V., Chumplitaz, M., Begazo, J., & Chauca, M. (2020). Characterization of the impact of SARS-CoV-2 pandemic social isolation on the psychosocial well-being of public university students, based on the GHQ-28 Scale. In *Proceedings of the 4th International Conference on Medical and Health Informatics* (pp. 295-301).
- Fuentes, C., Rodríguez, I., & Herskovic, V. (2016). Making Communication Frequency Tangible: How Green Is My Tree?. In *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction* (pp. 434-440).
- Galunder, S. S., Gottlieb, J. F., Ladwig, J., Hamell, J., Keller, P. K., & Wu, P. (2018). A VR ecosystem for telemedicine and non-intrusive cognitive and affective assessment. In *2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1-6). IEEE.
- Gao, X., Zhang, C., Ma, L., Wang, Y., Wang, J., & Zhang, D. (2019). Correlating msm's mental health with usage behaviors on msm-specific social applications. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation IEEE*.
- Griffin, J. (2010). The lonely society?. *Mental Health Foundation*.
- Ha, T. V., & Hoang, D. B. (2017). An assistive healthcare platform for both social and service networking for engaging elderly people. In *2017 23rd Asia-Pacific Conference on Communications (APCC)* (pp. 1-6). IEEE.
- Harvey, J., Golightly, D., & Smith, A. (2014). HCI as a means to prosociality in the economy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2955-2964).
- Hertz, N. (2021). The lonely century: how to restore human connection in a world that's pulling apart. *Currency*.
- Holt-Lunstad, J. (2017). The potential public health relevance of social isolation and loneliness: Prevalence, epidemiology, and risk factors. *Public Policy & Aging Report*, 27(4), 127-130.
- Holt-Lunstad, J., Smith, T. B., & Layton, J. B. (2010). Social relationships and mortality risk: a meta-analytic review. *PLoS medicine*, 7(7), e1000316.
- Ibarra, F., Baez, M., Cernuzzi, L., & Casati, F. (2020). A systematic review on technology-supported interventions to improve old-age social wellbeing: loneliness, social isolation, and connectedness. *Journal of healthcare engineering*, 2020.
- Jacobs, M. L., Clawson, J., & Mynatt, E. D. (2015). Comparing health information sharing preferences of cancer patients, doctors, and navigators. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 808-818).
- Jensen, R. B., Coles-Kemp, L., Wendt, N., & Lewis, M. (2020). Digital liminalities: Understanding isolated communities on the edge. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).
- Jeong, E. J., Kim, D. J., & Lee, D. M. (2015). Game addiction from psychosocial health perspective. In *Proceedings of the 17th International Conference on Electronic Commerce 2015* (pp. 1-9).
- Jeong, E. J., Kim, D. J., Lee, D. M., & Lee, H. R. (2016). A study of digital game addiction from aggression, loneliness and depression perspectives. In *2016 49th Hawaii International Conference On System Sciences (HICSS)* (pp. 3769-3780). IEEE.
- Joseph, K., Carley, K. M., & Hong, J. I. (2014). Check-ins in "Blau Space" Applying Blau's Macrosociological Theory to Foursquare Check-ins from New York City. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3), 1-22.
- Joyner, D. A., Wang, Q., Thakare, S., Jing, S., Goel, A., & MacIntyre, B. (2020). The synchronicity paradox in online education. In *Proceedings of the Seventh ACM Conference on Learning@ Scale* (pp. 15-24).
- Kaur, S., Kaul, P., & Zadeh, P. M. (2020). Study the impact of covid-19 on twitter users with respect to social isolation. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-6). IEEE.

- Kindness, P., Mellish, C., & Masthoff, J. (2013). How virtual teammate support types affect stress. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (pp. 300-305). IEEE.
- Knox, S. S., & Uvnäs-Moberg, K. (1998). Social isolation and cardiovascular disease: an atherosclerotic pathway?. *Psychoneuroendocrinology*, 23(8), 877-890.
- Lazányi, K. (2016). Investing in social support—Robots as perfect partners?. In 2016 IEEE 14th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 25-30). IEEE.
- Li, S., Ni, S., & Peng, K. (2020a). The Role of Good Human Uniqueness in Social Robot Anthropomorphism Influenced by Chronic Loneliness. In 2020 IEEE International Conference on Human-Machine Systems (ICHMS) (pp. 1-6). IEEE.
- Li, S., Xu, L., Yu, F., & Peng, K. (2020b). Does trait loneliness predict rejection of social robots? In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 271-280).
- Light, A., Howland, K., Hamilton, T., & Harley, D. A. (2017). The meaning of place in supporting sociality. In Proceedings of the 2017 Conference on Designing Interactive Systems (pp. 1141-1152).
- Lu, X., & Yao, J. (2010). The Influence of Internet Interpersonal Communication to Relationship and Loneliness of College Students. In 2010 International Conference on Web Information Systems and Mining (Vol. 2, pp. 386-390). IEEE.
- Lou, C., Zhao, J., Li, X., Wei, H., Zhang, Y., & Zhao, H. (2019). Pet Robot Emotional Interaction for Urban Autism. In Proceedings of the 2019 2nd International Conference on Intelligent Science and Technology (pp. 1-6).
- Lubben, J. (2017). Addressing social isolation as a potent killer!. *Public Policy & Aging Report*, 27(4), 136-138.
- Marangoni, C., & Ickes, W. (1989). Loneliness: A theoretical review with implications for measurement. *Journal of Social and Personal Relationships*, 6(1), 93-128.
- Martinez, A., Ortiz, V., Estrada, H., & Gonzalez, M. (2017). A predictive model for automatic detection of social isolation in older adults. In 2017 International Conference on Intelligent Environments (IE) (pp. 68-75). IEEE.
- Matthews, T., Danese, A., Wertz, J., Odgers, C. L., Ambler, A., Moffitt, T. E., & Arseneault, L. (2016). Social isolation, loneliness and depression in young adulthood: a behavioural genetic analysis. *Social psychiatry and psychiatric epidemiology*, 51(3), 339-348.
- Mulvenna, M., Zheng, H., Bond, R., McAllister, P., Wang, H., & Riestra, R. (2017). Participatory design-based requirements elicitation involving people living with dementia towards a home-based platform to monitor emotional wellbeing. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2026-2030). IEEE.
- Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology*, 18(1), 1-7.
- Noguchi, Y., Kamide, H., & Tanaka, F. (2018). Effects on the self-disclosure of elderly people by using a robot which intermediates remote communication. In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 612-617). IEEE.
- ONS (2018) Loneliness [Online] available at: <https://bit.ly/3zd10ZL> [Accessed 15 May 2021]
- Pedell, S., Vetere, F., Kulik, L., Ozanne, E., & Gruner, A. (2010). Social isolation of older people: the role of domestic technologies. In Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (pp. 164-167).
- Perlman, D., & Peplau, L. A. (1981). Toward a social psychology of loneliness. *Personal relationships*, 3, 31-56.
- Petersen, J., Austin, D., Kaye, J. A., Pavel, M., & Hayes, T. L. (2013). Unobtrusive in-home detection of time spent out-of-home with applications to loneliness and physical activity. *IEEE journal of biomedical and health informatics*, 18(5), 1590-1596.
- Pulekar, G., & Agu, E. (2016). Autonomously sensing loneliness and its interactions with personality traits using smartphones. In 2016 IEEE healthcare innovation point-of-care technologies conference (HI-POCT) (pp. 134-137). IEEE.
- Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2020). Multi-Class Suicide Risk Prediction on Twitter Using Machine Learning Techniques. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 128-134). IEEE.
- Ring, L., Barry, B., Totzke, K., & Bickmore, T. (2013). Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In 2013 Humaine Association conference on affective computing and intelligent interaction (pp. 61-66). IEEE.
- Russell, D. W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of personality assessment*, 66(1), 20-40.
- Russell, D., Cutrona, C. E., Rose, J., & Yurko, K. (1984). Social and emotional loneliness: an examination of Weiss's typology of loneliness. *Journal of personality and social psychology*, 46(6), 1313.
- Scalise, J. J., Ginter, E. J., & Gerstein, L. H. (1984). Multidimensional loneliness measure: the loneliness rating scale (LRS). *Journal of Personality Assessment*, 48(5), 525-530.
- Schmidt, N., & Sermat, V. (1983). Measuring loneliness in different relationships. *Journal of personality and social psychology*, 44(5), 1038.
- Senese, V. P., Nasti, C., Mottola, F., Sergi, I., & Gnisci, A. (2020, September). Validation and measurement

- invariance across gender and age of the Italian Interpersonal Acceptance-Rejection Loneliness Scale. In 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 000289-000294). IEEE
- Shah, S. G. S., Nogueras, D., Van Woerden, H., & Kiparoglou, V. (2019). Effectiveness of digital technology interventions to reduce loneliness in adults: a protocol for a systematic review and meta-analysis. *BMJ open*, 9(9), e032455.
- Steptoe, A., Shankar, A., Demakakos, P., & Wardle, J. (2013). Social isolation, loneliness, and all-cause mortality in older men and women. *Proceedings of the National Academy of Sciences*, 110(15), 5797-5801.
- Taylor, S. H., Hutson, J. A., & Alicea, T. R. (2017). Social consequences of Grindr use: Extending the internet-enhanced self-disclosure hypothesis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 6645-6657).
- Waterworth, J. A., Ballesteros, S., & Peter, C. (2009). User-sensitive home-based systems for successful ageing. In 2009 2nd Conference on Human System Interactions (pp. 542-545). IEEE.
- Weil, S. (1997). *Gravity and grace*. U of Nebraska Press
- Weinert, C., Maier, C., Laumer, S., & Weitzel, T. (2014). Does teleworking negatively influence IT professionals? An empirical analysis of IT personnel's telework-enabled stress. In *Proceedings of the 52nd ACM conference on Computers and people research* (pp. 139-147).
- Weiss, R. (1975). *Loneliness: The experience of emotional and social isolation*. MIT press.
- Wu, Y., Yuan, J., You, Q., & Luo, J. (2016). The effect of pets on happiness: A data-driven approach via large-scale social media. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 1889-1894). IEEE.
- Xu, D., Qian, L., Wang, Y., Wang, M., Shen, C., Zhang, T., & Zhang, J. (2015). Understanding the dynamic relationships among interpersonal personality characteristics, loneliness, and smart-phone use: evidence from experience sampling. In 2015 International Conference on Computer Science and Mechanical Automation (CSMA) (pp. 19-24). IEEE.
- Yang, H., & Bath, P. A. (2018). Prediction of Loneliness in Older People. In *Proceedings of the 2nd International Conference on Medical and Health Informatics* (pp. 165-172).
- Yoshida, M., Kleisarchaki, S., Gtirgen, L., & Nishi, H. (2018). Indoor occupancy estimation via location-aware HMM: An IoT approach. In 2018 IEEE 19th International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM) (pp. 14-19). IEEE.
- Zadeh, P. M., Khani, S., Pfaff, K., & Samet, S. (2020). A computational model and algorithm to identify social isolation in elderly population. In 2020 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-6). IEEE.
- Zhang, M. (2010). Exploring adolescent peer relationships online and offline: an empirical and social network analysis. In 2009 WRI International Conference on Communications and Mobile Computing (Vol. 3, pp. 268-272). IEEE.
- Zhou, Z., Sun, X., Chen, X., Fan, C., & Pan, Q. (2009). The Relation between Different Constructs in Middle Childhood Peer Interaction and Loneliness: A Mediation Model. In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (pp. 1-4). IEEE.
- Zhou, Q., Li, J., Tang, Y., & Wang, H. (2020). Discovering the Lonely Among the Students with Weighted Graph Neural Networks. In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 474-481). IEEE.

Low-power Machine Learning for Visitor Engagement in Museums

Marcus Winter¹^a, Lauren Sweeney¹^b, Katie Mason¹^c and Phil Blume²^d

¹*School of Architecture, Technology and Engineering, University of Brighton, Brighton, BN2 4GJ, U.K.*

²*The Regency Town House, Hove, BN3 1EH, U.K.*

marcus.winter@brighton.ac.uk, {l.sweeney3, k.mason6}@uni.brighton.ac.uk, phil.blume@rth.org.uk

Keywords: Machine Learning, Human Pose Estimation, Embodied Interaction, Visitor Engagement, Museums.

Abstract: Low-power Machine Learning (ML) technologies that process data locally on consumer-level hardware are well suited for interactive applications, however, their potential for audience engagement in museums is largely unexplored. This paper presents a case study using lightweight ML models for human pose estimation and gesture classification to enable visitors' engagement with interactive projections of interior designs. An empirical evaluation found the application is highly engaging and motivates visitors to learn more about the designs. Uncertainty in ML predictions, experienced as tracking inaccuracies, jitter, or gesture recognition problems, have little impact on their positive user experience. The findings warrant future research to explore the potential of low-power ML for visitor engagement in other use cases and heritage contexts.

1 INTRODUCTION

Machine Learning (ML) in cultural heritage is typically concerned with enhancing collections or supporting museum operations. Its potential for audience engagement is largely unexplored, despite the recent emergence of low-power ML technologies (Goel et al., 2020) and open ML model repositories, enabling low-budget development of interactive ML applications that run on consumer-level hardware and avoid privacy issues by processing data locally.

This paper describes a case study developing and evaluating an interactive ML application using these technologies in a heritage context to promote visitors' engagement and learning. Its main contributions are:

- a prototype application for visitor engagement, based on low-power ML technologies;
- an empirical evaluation focusing on visitors' reaction to uncertainty in ML predictions; user experience and visitor engagement.

The following sections contextualise the work, present the prototype application, describe the evaluation study and discuss its findings. The paper concludes by considering limitations of the work and setting out future research directions.

2 RELATED WORK

Museums have long been testbeds for novel technologies, however, their uptake of ML is difficult to assess. French and Villaespesa (2019:102) point out that "the broad language used to describe AI initiatives makes searching for use cases a daunting task". Based on a survey of 61 AI initiatives in museums, they focus their discussion on three application areas: computer vision to enhance collections data, ML for visitor data, and voice assistants for visitor engagement.

Responses to the MAIA survey (Hughes-Noehrer, Jay and Gilmore, 2022:Q8) allow for a similar classification, with 30% describing applications for collections, including enhancing metadata, image tagging, text extraction and OCR; 17% describing applications to support museum operations, including resource planning, ticketing and programming based on visitor data; and 20% describing applications that can be related to visitor engagement such as interactive or personalised experiences, or production of exhibits (other answers mention technologies rather than applications). Contrasting the current 20% of AI initiatives focusing on visitor engagement, 76%

^a <https://orcid.org/0000-0001-6603-325X>

^b <https://orcid.org/0000-0001-8312-1550>

^c <https://orcid.org/0000-0001-9532-9375>

^d <https://orcid.org/0000-0002-3211-3419>

of respondents think that AI can enhance the visitor experience on-site, or will be able to in the future (Hughes-Noehrer, Jay and Gilmore, 2022:Q15).

Clusters of past research using ML technologies for visitor engagement include:

- Chatbots offering conversational interfaces to explore collections (e.g., Boiano et al. 2003; Mollica, 2017; Anne Frank House, 2017)
- Robots engaging visitors in the gallery space (e.g., Burgard et al., 1999; Pitch et al. 2011; Del Vacchio, Laddaga and Bifulco, 2020)
- Interfaces to explore ML-enhanced collections (e.g., The Metropolitan Museum of Art, 2022; Harvard Art Museums, 2022a)
- Novel exhibits that would not be possible with conventional technologies (e.g., Tate, 2016; Mihailova, 2021).

While these efforts show that ML already plays an important role in museums, they also suggest a need for research exploring the potential of emergent low-power ML technologies for visitor engagement, which so far have received little attention.

3 HERITAGE CONTEXT

The Regency Town House, built in the 1820s as part of architect Charles Augustine Busby's *Brunswick Estate*, is a museum and heritage centre with a focus on the architecture and social history of Brighton & Hove between the 1780s and 1840s. Among its numerous collections it holds Busby's architectural plans, manuscripts, and aquatints, including an original copy of his 1834 publication, *Collection of Designs for Modern Embellishments*.

While Busby is mainly known as a designer of buildings, this book features his work on interior designs, which has so far received little attention. To celebrate his interior designs in an authentic setting, the brief was to develop an accessible, immersive, and engaging application that would enable visitors to interact with Busby's colourful and elaborate wall designs in a way that is both educational and fun.

4 PROTOTYPE APPLICATION

The developed prototype uses light-weight ML models for human pose estimation and gesture classification to support embodied interaction (Dourish, 2001) with interactive projections, which has been shown to facilitate meaningful experiences (Tan and Chow, 2017) and lead to high levels of

engagement (Lindgren et al., 2016). To satisfy a requirement for cross-platform support and enable other heritage organisations to access the application, the prototype is implemented as a web application. It runs in standards-compliant web browsers and does not require any specialist hardware.

Human pose estimation is used for embodied interaction with designs, turning users' hands into virtual paint brushes to successively reveal designs. A representation of the user's body is shown on screen to support coordination (Figure 1b). The prototype uses the MoveNet pose detection model (Voter and Li, 2021), which provides sufficient performance on most mid-range computers available today.

Gesture recognition is used for control operations. A single trigger gesture is used to start a session from the initial information screen (Figure 1a); progress from reveal action to showing the complete design; and progress from beholding a complete design to loading the next design to reveal. The prototype uses a pre-trained classifier based on body pose data in the COCO format (Lin et al., 2014) and a k-nearest neighbours (k-NN) algorithm. The classifier issues W3C standard Document Object Model (DOM) events for detected gestures, which are listened to by the web application to control the application state.

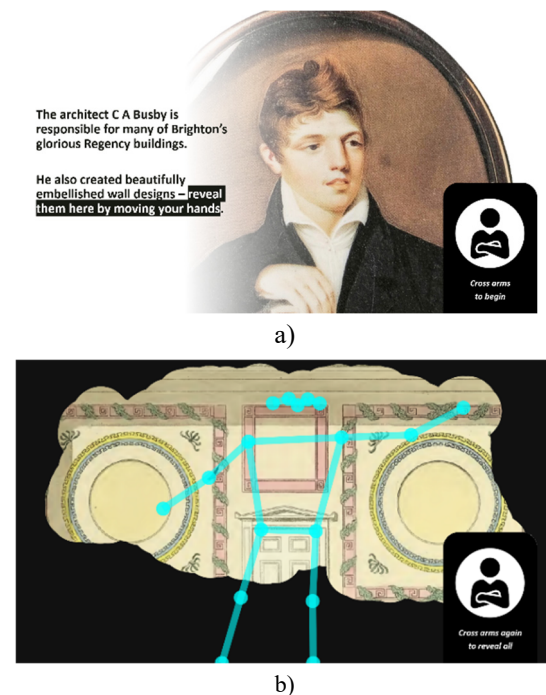


Figure 1: Interactive wall projection, including (a) initial information screen first encountered by visitors, (b) partly revealed design, with user's body representation and a hint explaining the control gesture in the bottom right corner.

Both MoveNet and the gesture classifier use Tensorflow.js (Smilkov et al., 2019), a JavaScript implementation of the Tensorflow open-source software library for ML (Abadi et al., 2016). The prototype has no other dependencies but uses plain HTML, CSS and JavaScript to ensure standards compliance and minimal overhead.

An admin interface provides functionality to control and customise the application. This includes loading content images, information screens, interaction hints and gesture classifier via URLs, enabling heritage organisations to customise the application for their specific use context and host critical resources in their own web space.

5 EVALUATION

The prototype application was empirically evaluated at the Town House over two days in June 2022, focusing on (i) usability and user experience, (ii) visitors' reaction to uncertainty in ML predictions, and (iii) visitor engagement and learning. The study design was scrutinised by the University of Brighton's ethics board and received a favourable opinion.

5.1 Participants

Participants were recruited through the Town House blog and social media channels to reach their regular audiences. The age distribution in the sample ($n=31$) is broadly comparable to the average age distribution among heritage audiences in England (DCMS, 2020) with the exception of 35-44 years (6.5%) and 45-54 years (32.3%), which are typically both around 16% (ibid). Participants' frequency of visiting museums and historic buildings (55% up to 5 times per year, 32% 6-15 times, 13% more than 15 times) is higher than the average among heritage audiences in England (ibid). Some participants (10%) say they usually visit on their own, while most visit with friends or family (42%), or say they do both (48%), suggesting that many participants understand museum visits as a social occasion.

5.2 Experimental Setup

The system was deployed in The Regency Town House first floor front dining room (Figure 2), using a notebook computer with Intel Core i7 processor, 16GB RAM and integrated webcam, and a projector with 1920 x 1080 pixels resolution and 6,000 lumen brightness.

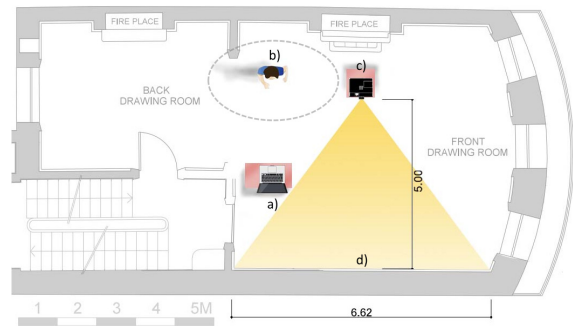


Figure 2: Room layout with positions of (a) computer and webcam, (b) visitor and approximate interaction space, (c) projector and (d) interactive projection.

The projector was fitted with a wide-angle lens to cover the whole sidewall, measuring 6.62m width x 3.72m height excluding bottom skirting, from a distance of 5m. The spatial layout was determined by practical constraints preventing the projector to be mounted at sufficient height for the projection beam to clear visitors interacting in the area in front of the wall. Instead, the projector was mounted on a pedestal at 0.5m height and visitors positioned themselves to the side of it to get a good view of the projection. The computer with integrated webcam was mounted on a small table at 0.65m height facing the user. It was positioned at maximum distance to the user while still clearing the projector beam to not throw a shadow. This resulted in an effective interaction space of approximately 3.4m x 2.2m in which users could operate and their full body was recognised.

5.3 Procedure and Data Collection

Visitors arriving at the Town House were welcomed by a researcher and informed about the context and purpose of the study, before ascending to the first floor dining room. Here Town House staff introduced them to the prototype application, providing information about Busby's interior designs and their relation to the Town House. Participants were also given some initial instructions to get them started with the embodied interaction. A researcher present in the room observed visitors' interaction with the prototype. After visitors were finished with their interaction, they were invited by Town House staff to take part in a short interview about their experience. Interviews were carried out by a researcher in the ground floor reception room. After the interview, participants were asked to complete a short user experience questionnaire.

Observations involved a researcher positioning themselves at the far end of the room to observe

visitors' interaction, which typically lasted 5-15 minutes, and take notes recording their initial reactions, level of engagement, learnability and usability issues, and any comments and feedback directed at Town House staff present in the room.

Interviews were designed to last 3-5 minutes and included a total of five questions, plus a section on demographic information. Following guidance in Valenzuela and Shrivastava (2008), the researchers took care to make participants feel comfortable and avoided or elucidated technical jargon during the interview. Answers were recorded in a bespoke coding sheet and notes were revised and clarified immediately after the interview while memory was still fresh.

The questionnaire was based on the short form of the UEQ User Experience Questionnaire developed by Schrepp, Hinderks and Thomaschewski (2017). It was filled in by participants after the interview and took 2-3 minutes to complete.

Observation notes, interview notes and completed questionnaires were digitised after the evaluation event to support the data analysis, while the original paper copies were destroyed.

5.4 Data Analysis

Qualitative observation notes and interview notes were transcribed and analysed independently by three researchers using an emergent coding process described in Miles and Huberman (1994). This involved first a data reduction step and then a data visualisation step to identify common themes. Themes identified in the three separate analyses were then discussed and synthesised by all three researchers together, using affinity diagrams as described in Courage and Baxter (2005). This resulted in a set of consolidated themes from these two datasets and allowed for triangulation between observed interaction (observation notes) and self-reported views about the experience (interview notes) to informing the findings.

Quantitative demographic data, as well as two quantitative sub-questions in the interview asking interviewees to rate on a fixed scale the accuracy of the pose estimation and the impact any inaccuracies had on their overall experience, were simply aggregated.

Quantitative data from the short version of the UEQ were analysed using the UEQ Data Analysis Tool developed by Schrepp (2017).

5.5 Findings

5.5.1 Usability and User Experiences

Observations show that most participants had little or no problems interacting with the system. While some were initially unsure how to start and asked staff for instructions, they quickly picked up on how the interaction works and confidently used both body poses and the cross-arms gesture to interact with the system. Overall, the observed behaviour suggests a good degree of usability and learnability in the current design, even though some aspects could be further improved, in particular, the gesture recognition (see discussion of uncertainty below). This is supported by interview data, with almost all respondents reporting that it was clear how the interaction works, even though some qualified this by attributing it to the brief introduction by staff, or observing others before interacting themselves, or saying it took them a while to work out what to do.

Some participants reported that they found it difficult to reach the corners and lower areas of designs, or remarked that the embodied interaction was physically challenging, confirming similar reports in the literature (Hincapié-Ramos et al., 2017; Jang et al., 2017). A few participants linked this to accessibility, pointing out that it was more difficult to use for people at their age, and for others with restricted mobility or dexterity. One suggestion to address these issues was to not only use users' hands as virtual paint brushes but also their feet. Alternative interaction modes were also suggested, for example offering people with mobility issues the opportunity to reveal the projected designs via a touch screen.

Several interview responses suggested alternative ways to provide initial instructions for situations where staff cannot be present (e.g., in writing; short animation). They also commented on the usefulness, design, or timing of the on-screen hints reminding users to cross their arms at certain points. Observations suggest these hints were often unnoticed or ignored by users, with some preferring to continue revealing a design with their hands rather than cutting short the process by crossing their arms. Some responses also questioned the need for an on-screen body representation or suggested alternative representations. All of these comments, however, were in the spirit of further improving the prototype rather than suggesting any fundamental problems in its interaction design.

Regarding their overall user experience, many participants were observed to exhibit behaviour or make comments indicating positive approval, using

terms like "amazing" or "extraordinary" or "really cool". A similar picture emerges from interviews, with many participants emphatic about the experience being fun and engaging. Several interviewees suggested that children would appreciate the embodied interaction and some suggested gamifying the experience to make it even more engaging.

Results from the UEQ further support these findings, with particularly positive scores for scales relating to ease of use and stimulation (Table 1). Mean scores of 1.716 (SD=0.928) for pragmatic quality, 1.500 (SD=1.397) for hedonic quality and 1.602 (SD=1.074) overall put the prototype in the top 10% of scores in the UEQ benchmark dataset of 468 studies involving 21,175 people (Schrepp, 2017), which indicates an excellent user experience overall (Figure 3).

Table 1: UEQ scores and confidence intervals ($p=0.05$).

Scale		Mean	Conf. Interval	
obstructive	supportive	1.21	0.74	1.69
complicated	easy	2.14	1.73	2.55
inefficient	efficient	1.31	0.89	1.74
confusing	clear	2.17	1.79	2.55
boring	exciting	1.14	0.56	1.73
not interesting	interesting	1.76	1.22	2.30
conventional	inventive	1.79	1.16	2.41
usual	leading edge	1.46	0.89	2.04

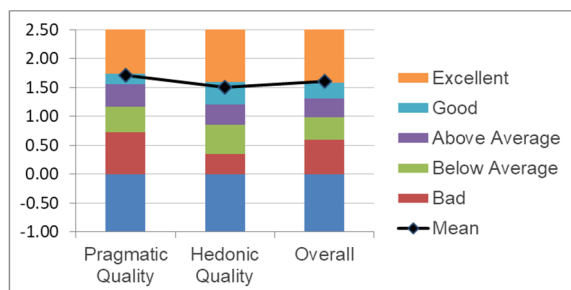


Figure 3: UEQ mean scores against benchmark dataset.

5.5.2 Uncertainty in ML Predictions

Uncertainty in ML predictions manifests itself in the prototype in various ways. In the context of human pose estimation, low confidence scores can lead to some body parts not being rendered on screen. Inaccurate estimations lead to a mismatch between on-screen representation and actual body pose, whereas slight differences in estimations between video frames lead to jitter in the on-screen body representation. In the context of gesture classification, uncertainty can lead to false positives

(i.e., a body pose is wrongly classified as a gesture) and false negatives (i.e., a correctly performed gesture is not recognised).

Observations show that the cross-arms gesture was often triggered accidentally (false positives) and, conversely, in some cases did not trigger as expected (false negatives). The former was typically caused by (a) visitors crossing their arms as part of their normal body posture; (b) wrong classification of other body poses involving crossing an arm over the torso; and (c) poor pose estimation due to lighting conditions, loose clothes or shoulder bags obscuring body parts. The latter was typically triggered by visitors performing the cross-arms gesture in a different manner than the classifier was trained to recognise (e.g., several visitors performed an "X" with lower arms crossing diagonally, while the classifier was trained mostly with lower arms crossed almost horizontally). Observations also show tracking inaccuracies and body parts not rendered on screen due to low confidence scores, for example when people performed extreme body poses to explore the limitations of the system. Overall, these aspects had little impact on visitors' engagement. While there were some reactions expressing surprise or even amusement, visitors generally worked through those situations unperturbed and seemed to accept them as part of the experience.

This is supported by interviews, which asked visitors to rate both the application's tracking accuracy and the impact of inaccuracies on their experience, and to expand on their ratings with open comments. Figure 4 shows that participants' ratings tend towards more positive assessments of both aspects. However, while more respondents rate the tracking accuracy as good (52%) rather than perfect (16%), this trend is reversed in how tracking inaccuracies impacted on the experience, with more respondents saying they had no impact (45%) rather than little impact (13%), suggesting a certain level of acceptance of the effects of uncertainty.

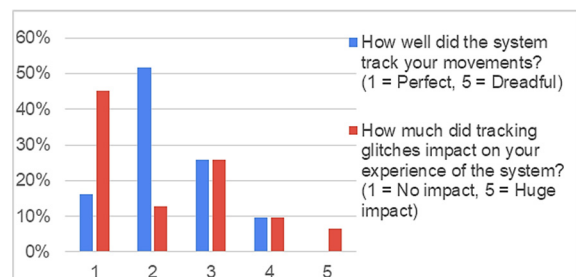


Figure 4: Visitors' ratings of how well the application tracked their movements and how much any tracking inaccuracies impacted on their experience.

Open answers describe issues as experienced by participants. While some of these relate to uncertainty in ML predictions, others are caused by ML inference latency, or are the result of interaction design decision aimed at mitigating some of the challenges of embodied interaction:

- Some participants remarked on the body representation sometimes being inaccurate or out of proportion or not at scale. This is only partly caused by uncertainty in ML predictions, as the prototype also makes purposeful adjustments that enable all participants to reach the top of the screen regardless of their body height or distance from the camera.
- Some respondents remarked on jitters in their on-screen body representation, which is caused by slight variations in key point predictions between video frames.
- Some respondents remarked on lag or delay, which is caused by latency in ML inference, and is inversely proportional to the processing power of the machine it runs on.

Overall, the data shows that several aspects of the user experience are affected by ML uncertainty or latency, however, it also suggests a high level of acceptance and willingness to work around any issues. Some interviewees even suggested that glitches made the experience more interesting as they added an element of unpredictability.

5.5.3 Visitor Engagement and Learning

Observations show how several visitors methodically reveal designs, often over prolonged periods of time, engaging with one design after another, directing enquiries about specific design features and colours to staff during their interaction, and talking about designs and the differences between them with staff and other visitors after their interaction. While most participants are clearly excited about the technology, they also show their appreciation for the designs they reveal, suggesting that the prototype overall manages to "preserve the primacy of the object and aesthetic encounter" (vom Lehn and Heath, 2003, p.3) and raises visitors' interest in the designs, motivating them to enquire and learn about them.

Interview data strongly supports this, with overwhelmingly positive answers when asked whether the prototype was an engaging way to learn about Busby's interior designs. Several participants expressed their excitement at the vivid colours of the designs and at seeing them projected at scale in an authentic environment. While there were two responses saying they would prefer static images or

written materials, most expressed their satisfaction at the interactive and immersive nature of the prototype, with some explicitly stating that it was more fun to reveal the designs rather than just observing them.

Several participants pointed out the value of additional narrative delivered by staff, either as part of the introduction, or commenting about specific designs, or answering questions by participants during or after their interaction. This reinforces the notion that the interaction awakens interest and motivates visitors to learn about the designs by asking staff for more information. It also indicates that the prototype can be part of a wider engagement and learning strategy in museums involving staff or other experts who can provide additional information and support a more conversational form of learning.

6 DISCUSSION

The empirical evaluation suggests good usability and learnability of the developed prototype. This was supported by participants receiving initial instructions by staff, the direct mapping between users' body pose and on-screen representation, and on-screen hints when to use the cross-arms gesture. Feedback suggests a need to improve the gesture recognition and explore the possibility of users customising or hiding their body representation.

While the presence of staff clearly added value to the experience, this might not always be feasible, especially in smaller museums with resource constraints. Future work should explore how the prototype can be further developed for use cases with no staff present. Feedback suggesting to gamify the experience, and observations of visitors' persistence in trying to reveal - or "complete" - all areas of a design, indicate a natural challenge to build on.

Uncertainty in ML predictions, perceived as tracking inaccuracies, jitter in the on-screen body representation, and gesture recognition problems, had little impact on participants' overall user experience, who generally showed a high level of tolerance, simply trying again when something did not work as expected. As this behaviour is likely to be influenced by staff readily providing hints and explanations, future work should explore if this tolerance holds without staff being present.

Despite these issues, the UEQ results suggest an excellent overall user experience against benchmark data (Schrepp, 2017; Schrepp, Hinderks and Thomaschewski, 2017). The high score for pragmatic quality supports findings indicating good usability, while the high score for hedonic quality supports

findings that participants enjoyed the experience. They add further support to literature on the engaging qualities of gesture-based interfaces (van Beurden, Ijsselsteijn and de Kort, 2012) and show that ML uncertainty does not diminish their appeal.

From a museum perspective, key questions include whether the application supports visitors' engagement and learning, and whether its use of ML is a feasible alternative to specialist hardware. The results show that the application is highly engaging and motivates visitors to ask questions about designs and learn more about Busby's work. Participants had no privacy concerns about being observed by a camera, and while they noticed the effects of uncertainty in ML predictions, these had little impact on their positive experience.

The findings provide useful insights informing future development and research. While not allowing for extrapolation to other use cases or contexts, they give an indication of the potential of low-power ML as an enabling technology for visitor engagement and provide a snapshot of related user experience issues.

7 LIMITATIONS

Aiming for high ecological validity, the evaluation study took place in the intended target environment and involved participants recruited via The Regency Town House's social media channels. The sample size and composition, choice of methods and rigorous data analysis ensure high internal validity, however, the bespoke nature of the prototype and the evaluation environment make it problematic to generalise findings to other applications and environments. As such, no recommendations or design guidelines are offered for low-power ML applications for visitor engagement.

8 CONCLUSIONS

This paper describes a prototype application using human pose estimation and gesture recognition for visitor engagement with interior designs in a heritage setting. Unlike other applications involving embodied interaction, it does not require specialist hardware but uses pre-trained ML models and runs on a mid-range computer with a webcam, putting it into reach for smaller museums with limited budgets and development capabilities.

The prototype uses low-power ML technologies, which are particularly well suited for interactive

applications as they process data locally rather than transmitting to a server, reducing latency and preserving visitors' privacy.

An empirical evaluation in the intended target environment found it usable, learnable and offering an excellent overall user experience. Besides engaging visitors of all ages, it motivated them to ask questions about the interior designs they revealed and to learn more about them in informal conversations with staff. Uncertainty in ML predictions, perceived by visitors as tracking inaccuracies, jitter in their on-screen representation and gesture recognition issues, had little impact on their positive experience.

The findings indicate that low-power ML holds great promise for visitor engagement in heritage contexts and warrant future research to explore this potential. This includes developing designs that can run unsupervised in the gallery space, without staff being present to provide information and assistance, and exploring how other ML capabilities can support visitor engagement and learning in museums.

ACKNOWLEDGEMENTS

This research was supported by the Ignite funding scheme of the Community University Partnership Programme (CUPP) at the University of Brighton. We would like to thank visitors to The Regency Town House for taking part in the evaluation and sharing their valuable views and feedback.

REFERENCES

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16). arXiv:1605.08695.
- Anne Frank House (2017). Anne Frank House launches bot for Messenger. Available <https://www.annefrank.org/en/about-us/news-and-press/news/2017/3/21/anne-frank-house-launches-bot-messenger/>. Retrieved 6 August 2022.
- Boiano S, Gaia G, Caldarini M (2003) Make your museum talk: natural language interfaces for cultural institutions. Museums and the Web 2003. Available <https://www.museumsandtheweb.com/mw2003/papers/gaia/gaia.html>. Retrieved 6 August 2022.

- Burgard, W., Cremers, A. B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., and Thrun, S. (1999). Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1–2), 3–55.
- Courage, C. and Baxter, K., 2005. Understanding your users: A practical guide to user requirements methods, tools, and techniques. Gulf Professional Publishing.
- DCMS (2020). Taking Part 2019/20: Cross-sectional survey. Technical Report. Available https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/916246/Taking_Part_Technical_Report_2019_20.pdf. Retrieved 27 July 2022.
- Del Vacchio, E., Laddaga, C., & Bifulco, F. (2020). Social robots as a tool to involve student in museum edutainment programs. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 476–481.
- Dourish, P. (2001). *Where the action is: the foundations of embodied interaction*. MIT Press, Cambridge, Mass.
- French, A. and Villaespesa, E. (2019). AI, visitor experience, and museum operations: a closer look at the possible. In *Humanizing the Digital: Un-proceedings of the MCN 2018 Conference*, 101–113.
- Gaia, G., Boiano, S., and Borda, A. (2019). Engaging museum visitors with AI: The case of chatbots. *Museums and digital culture*, 309–329. Springer.
- Goel, A., Tung, C., Lu, Y. H., and Thiruvathukal, G. K. (2020). A survey of methods for low-power deep learning and computer vision. In *6th World Forum on Internet of Things (WF-IoT)*, 1–6. IEEE.
- Harvard Art Museums (2022). AI Explorer: Explore how a computer sees art. Available <https://ai.harvardartmuseums.org/about>. Retrieved 11 August 2022.
- Hincapié-Ramos, J. D., Guo, X., Moghadasian, P., and Irani, P. (2014). Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1063–1072. ACM.
- Hughes-Noehrer, L., Jay, C. and Gilmore, A. (2022). *Museums and AI Applications (MAIA) Survey*. University of Manchester. Dataset. <https://doi.org/10.48420/19298588.v1>
- Jang, S., Stuerzlinger, W., Ambike, S., and Ramani, K. (2017). Modeling cumulative arm fatigue in mid-air interaction based on perceived exertion and kinetics of arm motion. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3328–3339. ACM.
- Lee, L., Okerlund, J., Maher, M. L., & Farina, T. (2020, July). Embodied interaction design to promote creative social engagement for older adults. In *International Conference on Human-Computer Interaction*, 164–183. Springer, Cham.
- Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV)*, 740–755. Springer.
- Lindgren, R., Tscholl, M., Wang, S., and Johnson, E. (2016). Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education*, 95, 174–187.
- Mihailova, M. (2021). To dally with Dalí: Deepfake (Inter) faces in the art museum. *Convergence*, 27(4), 882–898.
- Miles, M.B., and Huberman, A.M. 1984. *Qualitative Data Analysis*. Newbury Park, CA: Sage.
- Mollica, J. (2017). Send Me SFMOMA. Available <https://www.sfmoma.org/read/send-me-sfmoma/>. Retrieved 6 August 2022.
- Pitsch, K., Wrede, S., Seele, J. C., and Süssenbach, L. (2011). Attitude of german museum visitors towards an interactive art guide robot. In *Proceedings of the 6th international conference on Human-robot interaction*, 227–228. ACM.
- Schrepp (2017). UEQ Data Analysis Tool. Available https://www.ueq-online.org/Material/Short_UEQ_Data_Analysis_Tool.xlsx. Retrieved 26 July 2022.
- Schrepp, M., Hinderks, A., and Thomasczewski, J. (2017): Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *IJIMAI*, 4 (6), 103–108.
- Smilkov, D., Thorat, N., Assogba, Y., Nicholson, C., Kreeger, N., Yu, P., Cai, S., Nielsen, E., Soegel, D., Bileschi, S. and Terry, M. (2019). Tensorflow.js: Machine learning for the web and beyond. *Proc. of Machine Learning and Systems*, 1, 309–321.
- Tan, L., and Chow, K. K. (2017). Facilitating meaningful experience with ambient media: an embodied engagement model. In *Proceedings of the 5th International Symposium of Chinese CHI*, 36–46.
- Tate (2016). Can a machine make us look afresh at great art through the lens of today’s world? IK Prize 2016: Recognition. Available <https://www.tate.org.uk/whats-on/tate-britain/exhibition/ik-prize-2016-recognition>. Retrieved 11 August 2022.
- The Metropolitan Museum of Art (2022) The Met Art Explorer. Available <https://art-explorer.azureweb sites.net/search>. Retrieved 11 August 2022.
- van Beurden, M.H., Ijsselsteijn, W.A., de Kort, Y.A. (2012). User Experience of Gesture Based Interfaces: A Comparison with Traditional Interaction Methods on Pragmatic and Hedonic Qualities. *LNCS*, vol 7206, 36–47. Springer, Berlin.
- Voter, R. and Li, N. (2021). Next-Generation Pose Detection with MoveNet and TensorFlow.js. Available <https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html>. Retrieved 26 July 2022.
- Winter M., Jackson P. (2020) Flatpack ML: How to Support Designers in Creating a New Generation of Customizable Machine Learning Applications. *LNCS* vol 12201, 175–193. Springer Nature.

Student-centered Development of an Online Software Tool to Provide Learning Support Feedback: A Design-study Approach

Gilbert Drzyzga^a and Thorleif Harder^b

Institute of Interactive Systems, Technische Hochschule Lübeck, Germany
{gilbert.drzyzga, thorleif.harder}@th-luebeck.de

Keywords: Student-centered Design, Design Process, Gestalt Laws, Interaction Principles, Usability, User Experience.

Abstract: Students in online degree programs have a higher risk of dropping out (Diaz, 2002; Beard and Harper, 2002; Baker et al., 2015). The use of learning support tools such as learner dashboards (LD) can promote self-regulated learning, which can have a positive impact on student learning (Jivet et al., 2018; Konert et al., 2016). This paper presents a three-stage design study and shows how the layout of the proposed LD was implemented from the initial digital design to a low-fidelity prototype. First, the developed wireframes were checked for consistency with respect to the Gestalt laws (Wertheimer, 1922). From the resulting wireframe design, a clickable low-fidelity prototype was developed. In the second step, this interactive prototype was reviewed by students (n=24) respect to the seven interaction principles (DIN EN ISO 9241-110, 2020). In the third step, the revised prototype was subjected to an eye-tracking procedure using the Thinking Aloud technique (n=10). The results so far show that the LD should be presented at a reduced information level during initial access, but that this level can be supplemented by additional elements if necessary. The navigation hierarchy should be kept flat and the information should be easy to understand.

1 INTRODUCTION

In order to be able to offer students of digital study programs support that promotes learning during their studies and to reduce dropout in this form of study overall, a research project of a university network for digital study programs with over 4,000 enrolled students is investigating how students could be supported in their learning process (Diaz, 2002; Beard and Harper, 2002; Baker et al., 2015). To this end, a learner dashboard (LD) is being developed to stimulate student self-regulation. The LD is to be integrated as a plug-in into the university network's Moodle learning management system (LMS) and, once completed, will be made available to the community as open source. Within the framework of a user-centered design (UCD) approach, it is to be investigated which indicators, representation, information and intervention possibilities could be useful in this process and in which way. To this end, a design-study was conducted that employed various usability methods in three steps to obtain qualitative feedback for the implementation of the user interface from experts

(see Figure 1).

The design-study is based on findings from an initial literature review with the focus on self-regulated learning in digital learning environments and expert interviews (n=9). Based on these findings, a first wireframe was implemented in the winter semester 2021 / 2022 and subsequently evaluated (Figure 1, Level 1). During the summer semester of 2022, a low-fidelity prototype was implemented and is to be used in the Moodle LMS of the university network in the near future as the first development level of an LD (Figure 1, Level 2 and 3). This first technical deployment level provides descriptive evaluations for students. In this paper, we present the research questions based on qualitatively conducted studies presented in the current version of the LD as a work in progress. This approach has proven successful, as it has enabled successive optimization of the design of the user interface with the involvement of experts.

Research Questions (RQ):

- **RQ1:** How must the user interface elements of an online support tool be designed to effectively support learning?
- **RQ2:** How must the user interface and individual

^a  <https://orcid.org/0000-0003-4983-9862>

^b  <https://orcid.org/0000-0002-9099-2351>



Figure 1: Representation of the three-level design study.

content elements of the online support tool be designed to promote usability and user experience?

2 METHODOLOGICAL APPROACH

The used investigations in the design-study were conducted on the basis of three usability tests with the focus group of students.

1. Evaluation of the wireframe by students (n=24) using the Gestalt laws (students of a module on human-computer interaction)
2. Evaluation of the low-fidelity prototype using seven interaction principles. (Students of a human-computer interaction module)
3. Student evaluation of the low-fidelity prototype, including examination of the user experience (UX) using the User Experience Questionnaire-Short (UEQ-S) (Students of a usability module)

3 DEVELOPMENT AND IMPLEMENTATION OF THE LD

The first graphic design in the form of a wireframe of the LD (Figure 2) formed the basis for the investigations. The wireframe was developed based on the literature review and expert interviews (n=9) as part of the UCD process following the paper prototype method.

3.1 Evaluation based on Gestalt Laws and Fact and Interaction Problems (Step 1)

The wireframe shown in Figure 2 was examined in two parts with an additional assignment over a period of four weeks as part of a module in the online course

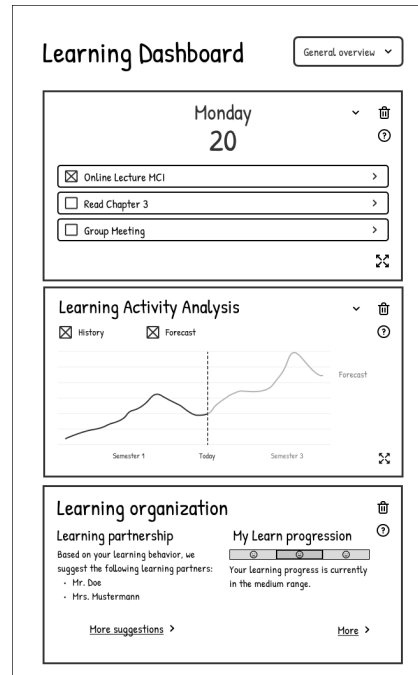


Figure 2: First version of the LD wireframe.

Media Informatics in the summer semester of 2022: In the first part, the online students (n=24) were asked to evaluate the wireframe for the five Gestalt laws: "Law of Proximity", "Law of Similarity", "Law of Proximity", "Law of (Good) Continuation", "Law of Good Gestalt" (Wertheimer, 1922) with at least one positive or one negative example each. The findings of Gestalt psychology are intended to reduce the cognitive effort required to assimilate information in LD in the first place (Wagemans et al., 2012). The second part was to investigate whether factual and interactional problems are present in the design. Both were justified with examples in the elaboration. In the section "Created alternative solutions of the wireframe by students" the additional task is considered separately with an example. The result of the investigation was a total of 75 DIN A4 pages of student findings. Figure 3 shows an example of an identified problem on the "Law of Proximity".

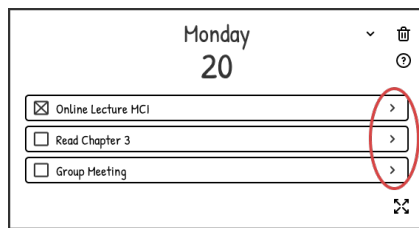


Figure 3: Example from a student of an identified problem with the Gestalt "Law of Proximity".

Alternative Solution of the Wireframe Created by Students

Those students were given the additional task of creating an alternative solution for a map of their choice based on the "Law of Proximity". 11 of the 24 students solved this task. An example of such a design can be seen in Figure 4, which illustrates how students perceive such functionalities and which aspects are important to them. Moreover, this demonstrates the benefits of the approach taken in this design-study, where the involvement of students results in high-quality, student-centered outcomes.

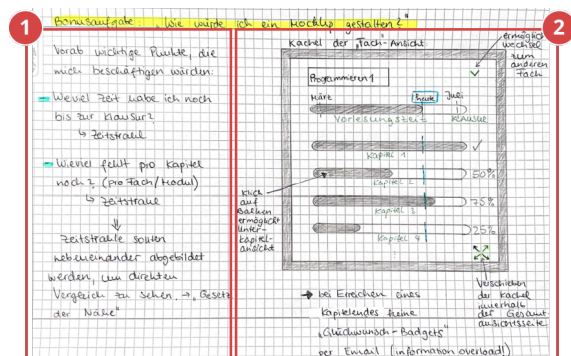


Figure 4: Example of a wireframe of a card developed by a student.

The following text describes the German text shown in Figure 4 in English.

Bonus Task "How would I design a mockup?" The area outlined in red with the number one describes the issues that would concern the student when viewing the card:

- *How much time do I have until the exam? → timeline*
- *How much is still missing per chapter? (per subject / module of study) → timeline*
- *Timelines should be displayed side by side to see direct comparison → "Law of proximity"*

The area outlined with the number two represents the card of the "Course" view. The course "Program-

ming I" is displayed as the heading. Next to the heading, there is a possibility to go to another course using the downward arrow. Below that, bars for different information are displayed. The top one shows the complete lecture period, including the start of the exam period. The following four progress bars indicate chapters to be completed within the course. Progress bars show a subchapter view when clicked. Click on the icon in the lower right corner to move the card anywhere within the overall view page. A "congratulations-badge" should not be emailed to the student as this could result in information overload.

The navigation concept was revised because it was described as contra-intuitive by those students. In order to take the aspect of individualizability into account, an editing functionality was also integrated. For the calendar, the interaction options were revised (marker element deleted, information expanded or summarized, navigation elements for changing the days of the week). For the learning progress, the graphical representation was adapted by popular request of the students. Due to the variety of information within the LD, an additional information element in the form of a question mark was inserted in a prominent position (in the first line of the grid, right-aligned). This can be used to quickly obtain information about dashboard elements from any view. Figure 5 shows the revision in detail.

3.2 Evaluation based on the Seven Interaction Principles (Step 2)

The revised wireframe served as the basis for creating a low-fidelity prototype. There were 14 different views with different modal dialog boxes in the final version of the LD prototype. The individual interaction elements of the LD (question mark icon, zoom icon, pencil icon, and menu) were designed to be clickable.

The prototype was made available to students of a human-computer interaction module (n=24) for evaluation during a four-hour online workshop via an Internet browser. The experts tested and evaluated the seven interaction principles using a two-part task (DIN EN ISO 9241-110, 2020). The workshop was accompanied by 3 persons from the university sector. The result of the study was a total of 54 DIN A4 pages with the students' findings.

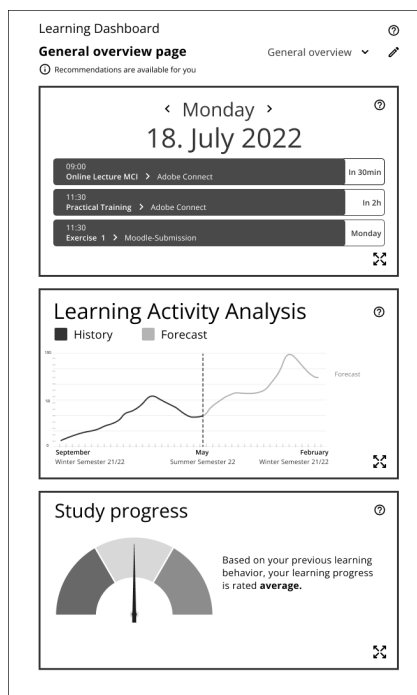


Figure 5: Revised variant of the LD's wireframe.

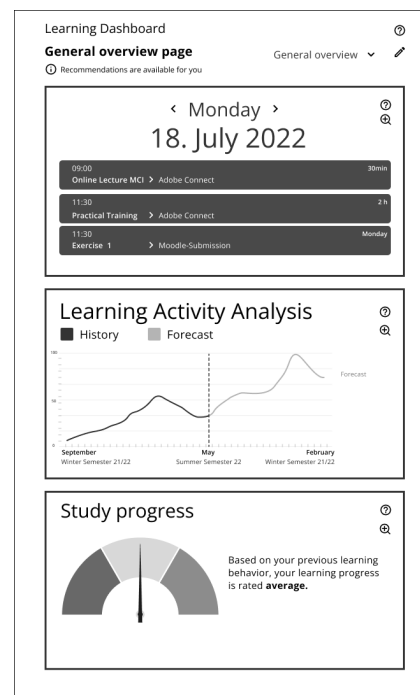


Figure 6: Low-fidelity prototype of the LD.

3.3 Performing the Eye-tracking Procedure with Thinking Aloud (Step 3)

The entire low-fidelity prototype was revised with the new findings (Figure 5 shows part of the revised LD). Afterwards, it was evaluated by a group of students from a usability module ($n=10$). The entire study was conducted in one day. Each session lasted a total of 30 minutes (including a welcome, introduction of the project, and explanation of the procedure). Students were presented with several subtasks to complete during the 20-minute testing period and were guided by moderators. After the eye-tracking study, students performed a rating of the LD using the short version of the User Experience Questionnaire (UEQ-S) (Schrepp et al., 2017). This was used to evaluate the overall usability of the prototype. Since eye-tracking study was conducted only recently, only the already available results are presented below.

First Results of the Third Step of the Design Study

The students confirmed the implementation of the navigation concept and the general structure of the LD. Specifically, they stated that the descriptions should be precise and brief and can be expanded by

clicking on further or more detailed information as needed. This applies, for example, to the recommendations on learning progress, which should initially be presented only in condensed form when the detailed view is called up. The reason for this is that the LD is accepted as an additional element to support learning, but this should not interfere too much with the actual learning time of the students. It should be possible to add additional user-defined information to the assessments (e.g., current exam grades or progress in reading the learning content). However, this needs further research as students expressed that the functionality might be too complicated or time-consuming. For the top-level help page, students would like to see a walkthrough or short slideshow for the LD instead of textual information (also time-consuming). In the usability test, it was found that the navigation structure is not always intuitive: for example, the icon for more information about the displayed content was not always recognized as such. One solution would be to make the entire map clickable. The individual views (overall view, semester view and module view) could not be distinguished in some cases. Here, the way the content is displayed should be reconsidered in order to better distinguish between the different views.

3.4 Student Opinions from the Evaluations

The evaluations clearly showed that the students got a good idea of the future purpose and benefit that the shown wireframes depicted. For example, the students commented in particular on the interaction principle "User Engagement" that the interest was aroused "[...] whether the learning behavior has improved (or worsened) after a few days - the ambition is aroused, game character". Further feedback showed that the students were pleased with the LD in their studies and could imagine "[...] that the available information would lead to structuring my studies further or better and thus increase the chances of success. Which in turn leads to using the system permanently."

4 OUTLOOK & DISCUSSION

The next steps include the evaluation of the eye-tracking study, the revision of the LD and subsequent transfer as a plug-in into the LMS. The results so far show that the direct exchange with the students is crucial in order to be able to offer a LD that is used and accepted. Important aspects and impulses for the information transfer, the design as well as the structure and conception of a learning-supporting LD could be gained. It is important that no additional (time and cognitive) effort is required when using the LD, e.g. for familiarization and understanding of its use, as it bears the risk of not being used by online students. The design study brought substantial results for further research and implementation of the LD in the LMS.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education, grant No. 01PX21001A / 01PX21001B.

REFERENCES

- Baker, R. S., Lindrum, D., Lindrum, M. J., and Perkowski, D. (2015). Analyzing early at-risk factors in higher education e-learning courses. *International Educational Data Mining Society*.
- Beard, L. A. and Harper, C. (2002). Student perceptions of online versus on campus instruction. *EDUCATION-INDIANAPOLIS THEN CHULA VISTA*-, 122(4):658–663.
- Diaz, D. P. (2002). Online drop rates revisited. *The technology source*, 3(3):35–51.
- DIN EN ISO 9241-110 (2020). DIN EN ISO 9241-110:2020-10, ergonomie der mensch-system-interaktion - teil 110: Grundsätze der dialoggestaltung (ISO 9241-110:2020); deutsche fassung EN ISO 9241-110:2020.
- Jivet, I., Scheffel, M., Specht, M., and Drachsler, H. (2018). License to evaluate: Preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 31–40.
- Konert, J., Bohr, C., Bellhäuser, H., and Rensing, C. (2016). Peerla-assistant for individual learning goals and self-regulation competency improvement in online learning scenarios. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 52–56. IEEE.
- Schrepp, M., Hinderks, A., and Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(6), 103-108.
- Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., and von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological bulletin*, 138(6):1172.
- Wertheimer, M. (1922). Untersuchungen zur lehre von der gestalt. *Psychologische forschung*, 1(1):47–58.

EETAS: A Process for Examining Ethical Trade-Offs in Autonomous Systems

Catherine Menon¹^a, Silvio Carta²^b and Frank Foerster¹^c

¹*Department of Computer Science, University of Hertfordshire, College Lane, Hatfield, U.K.*

²*School of Creative Arts, University of Hertfordshire, College Lane, Hatfield, U.K.*

{c.menon, s.carta, f.foerster}@herts.ac.uk

Keywords: Ethics, Design, Ethical Trade-Offs.

Abstract: Public-facing autonomous systems present society with significant ethical challenges, not least of which is the need for stakeholder understanding and discussion of how these systems balance competing ethical principles. In this paper we present EETAS: a structured, gamified process for obtaining stakeholder input into the ethical balances and trade-offs which they consider it acceptable for a proposed autonomous system to make. We describe how outcomes from the EETAS process can be used to inform the design of specified autonomous systems, as well as how the process itself can improve stakeholder engagement and public understanding of ethics in AI and autonomous systems. In support of this we present the findings from an initial EETAS pilot study workshop, which shows an indicative trend of improvement in public understanding and engagement with AI following participation.

1 INTRODUCTION

One of the most complex obstacles to public acceptance of autonomous systems (AS) and AI is the understanding of how competing ethical requirements in these systems may be managed. Standards such as BSI 8611 (British Standards Institute, 2016) on the ethical design and applications of AS, the IEEE guidance on ethically-aligned design (IEEE, 2018) and the Turing Institute guidance on understanding artificial intelligence ethics and safety (Leslie, 2019) provide information to developers on ethical imperatives to be satisfied by the system, but there is very little existing guidance for either stakeholders or developers on managing and understanding ethical complexities and balances.

Furthermore, conversation around AI has traditionally focused on the technology and its capabilities, rather than the diverse ethical concerns of stakeholders and wider society. Nonetheless, autonomous systems cannot exist in an ethical vacuum; rather, they will be expected to conform to the social, legal and ethical norms of the community in which they operate (IEEE, 2018). This is a non-

trivial task, as different societies – and different stakeholders within those societies – may prioritise ethical principles differently. For example, societies with a relatively greater regard for governmental authority may be comfortable with autonomous systems which prioritise public safety over data privacy, as demonstrated in the adoption of the TraceTogether app in Singapore (Lee, 2020). Similarly, stakeholders within other societies which prioritise individual choice over public cohesion may have an ethical preference for AS which obey user commands even where this could compromise public safety, such as allowing customisation, or “individuation” into autonomous vehicle technology (Hancock, 2019). In addition to this, societies are of course not homogenous, and individual stakeholders within any given society may also prioritise ethical principles differently depending on age, class, gender and perceived technical competence (Park, 2021).

In this paper we present a structured process for obtaining stakeholder input into the acceptability of ethical trade-offs in a proposed autonomous system. This structured process enables stakeholders to identify potential trade-offs between different ethical

^a <https://orcid.org/0000-0003-2072-5845>

^b <https://orcid.org/0000-0002-7586-3121>

^c <https://orcid.org/0000-0002-9263-3897>

principles and to work collaboratively to identify constraints, or limits, within which they consider these trade-offs to be ethically acceptable. We hypothesise that the benefits of this structured process are two-fold: firstly that participation will improve public understanding of, and willingness to engage with, ethical complexities of AS and secondly that AS designers will gain insight into potential design choices which may be made to render the autonomous system more acceptable to the public.

We also present an interactive tool (Figure 1) which we have created to provide a visual representation of the outcomes of the EETAS process. This tool serves as a record of the public discussion, which can be retained by end-users or stakeholder organisations and used to illuminate diverse public perspectives on AS ethics. In addition, the tool can be used later in the lifecycle to communicate the autonomous system's ethical prioritisations and to increase end users' understanding of it.

In Section 2 we present a discussion of existing literature which considers questions of ethical prioritisation in autonomous systems. Section 3 contains our description of the EETAS process, while Section 4 provides a description of an initial pilot study workshop which has demonstrated an indicative trend between participation in EETAS and enhanced public understanding of AS. Section 5 identifies our conclusions and some steps for further work.

2 BACKGROUND

The concept of trade-offs, or risk balancing, between two desirable properties is well-established as a research area. Expected utility theory (von Neumann, 1947) describes how an individual's general attitude to risk and benefits can change their willingness to accept particular specified risks. Similarly, prospect theory (Kahneman & Tversky, 1979) also allows a more complex framing of risk perception and risk appetite. The trolley problem (Foot, 1967) is of course perhaps the seminal example of risk balancing, and has informed much of the public discourse around autonomous vehicle behaviour.

Beyond this, risk balancing as a concept is well-explored in autonomous system development. (IET, 2019) describes trade-offs between safety and security of cyber-physical systems, while (Akinsanmi, 2021) considers the balancing of public health, privacy and digital security. Within specific autonomous domains the concept of prioritising

certain safety or ethical properties has also been discussed: (Thornton, 2018) describes the tension between the desire for personal autonomy on the part of an autonomous vehicle user, and the more general desire for fairness and public safety while (Lin, 2015) also considers how specific actions on the part of an autonomous vehicle – e.g. driving closer to another car in order to give more room to a pedestrian – transfer the risk from one segment of the population (pedestrians) to another (other drivers). In the field of healthcare, ethical trade-offs between privacy and well-being are also common (Lee, 2020), (Martinez-Martin, 2020).

Other existing work focuses specifically on trade-offs which affect the design process. (Dobrica, 2002) presents a comprehensive survey of trade-offs in complex systems design, while (Goodrich, 2000) discusses these trade-offs within an autonomous context, specifically that of collision avoidance systems. Similarly (Bate, 2008) considers trade-offs more generally within safety-critical systems, while (Menon, 2019) proposes a methodology for developers of autonomous vehicles to justify and communicate the ways in which their system design has been informed by ethical trade-offs.

The benefit of using a tangible element such as the interactive tool in Figure 1 to test and visualise trade-offs in real-time is supported by a large body of literature, including (Schrier, 2019), (Rossi, 2019), (Larson 2020). More generally, games have been shown to be a successful vehicle for engagement with ethics principles, especially in industry testing. Examples include Judgment Call, (Ballard, 2019), a game developed to help AI developers to identify ethical questions using design fiction, as well as MiniCode, a design fiction toolkit developed for near-future technology designers and developers (Malizia, 2022).

Much of the existing work around autonomous systems and AI is focused on developers, intended either to provide them with insight into how a system can be designed or to be used as guidance on making ethically justifiable decisions. However, there is comparatively little work which provides stakeholders and end-users with an opportunity to express their concerns around AI ethics, or to inform the design of a proposed system by providing input into the perceived acceptability of ethical trade-offs. The process we describe here addresses this gap.

3 EETAS PROCESS

In this section we describe a process for obtaining stakeholder input into decisions about the ethical prioritisations embedded into autonomous systems. It is important that the EETAS process takes place relatively early in the design of such systems, to allow the outcomes to be fed back into the design lifecycle and consequently enable developers to integrate ethically acceptable behaviour into the system from the ground up. In consequence, the system may not be fully specified at the time the EETAS process takes place. This is expected, and the process allows for under-specification and early prototypes of an autonomous system to be used.

3.1 Step 1: Provide AS Description

A participant group is selected, including developers of the AS under consideration, proposed end-users, regulators and members of the public. The developers provide the group with a written, accessible description of the AS and its relevant functions. Appropriate descriptions may specify, for example, that this is “an assistive robot that reminds you when to take medication, alerts you when you have left the oven on and engages you in conversation”. It is likely that participants will have further questions around the functionality of the system – e.g. “does the assistive robot speak to me or do I access it via a screen?” – and these should be clarified with the developers as part of this step. As EETAS is ideally undertaken during the early stages of development (e.g. requirements gathering or design), it is likely that some questions around specific functionality cannot yet be answered. These, in turn, become the seeds for the scenarios that teams will identify in Step 3.

3.2 Step 2: Identify Relevant Ethical Principles

The participants are divided into teams of between 4 and 10. This follows (Curral, 2001), with the intent of ensuring diversity of perspectives while still enabling effective and equitable dialogue. Each team is provided with a set of pre-prepared cards listing ethical and ethically-informed functional properties which may be desirable for this type of autonomous system. The ethical properties in this set have been identified from a literature review of existing and developing standards, including (BSI, 2016), (IEEE, 2018), (Leslie, 2019), (National Cyber Security Centre,

2019). Some sample principles, which we used in the initial pilot validation workshop (Section 4) are:

- System promotes human physical safety
- System obeys human commands
- System promotes affinity with human user
- System maintains data privacy
- System is accurate
- System is fair
- System maintains human autonomy
- System promotes human long-term health

We note that not all the ethical or functional properties in the full set will be relevant for every system, and that for specific systems there may be additional ethical or functional properties. To address this teams are also provided with a set of blank cards and are encouraged to “tailor” the set of ethical properties to discard those they consider irrelevant, and identify any others considered relevant to this specific system, using techniques such as collaborative discussion, brainstorming and if-then thinking.

3.3 Step 3: Scenario Construction

Teams are then asked to generate scenarios in which two of the set of ethical properties are in conflict with each other during system operation. For example, a team may postulate a scenario where: “the user asks their assistive robot not to remind them about medication today, because they don’t want to take it”. In this scenario the ethical properties of “system promotes human long-term health” and “system obeys human commands” are in conflict. Similarly, allowing teams to tailor the set of ethical principles may give rise to a scenario for a robot doctor where the system attempts to “engender trust in the human user” by mimicking human appearance and gestures to make the patient feel at ease, thereby causing conflict with another ethical property: “system does not attempt to deceive”.

To assist in generating the scenarios, teams are provided with a ready-made checklist of guidewords, to be applied in turn to each of the ethical properties. These guidewords enable teams to work collaboratively to brainstorm scenarios, following the principles of Hazard and Operability Analysis (HAZOP) studies (BSI HAZOP, 2016). The guidewords are presented in Table 1.

Teams should remember that the intent is to identify scenarios in which two or more ethical principles are in conflict: it is not sufficient to identify scenarios which themselves simply represent ethical hazards.

Table 1: HAZOP guidewords for EETAS.

Guide word	Meaning
TOO MUCH	Ethical conflict arising from a scenario where the robot performs its functions in such a way that it grants this ethical property to too many people / in too many circumstances / to too high a degree
NOT ENOUGH	Ethical conflict arising from a scenario where the robot performs its functions in such a way that it grants this ethical property to too few people / in too few circumstances / to too restricted a degree
UNIFORMLY	Ethical conflict arising from a scenario where the robot performs its functions in such a way that the outcome is applied uniformly to everybody / is applied in exactly the same way to everybody
INCONSISTENTLY	Ethical conflict arising from a scenario where the robot performs its functions inconsistently / differently for different people / differently each time
UNEQUALLY	Ethical conflict arising from a scenario where the robot performs its functions such that the beneficial outcome applies only to some people

Participants can be encouraged to apply creativity when identifying scenarios, and may find it helpful to consider the following questions:

- Who is the user of the system?
- Who would be negatively affected in this scenario?
- Who would benefit in this scenario?

3.4 Step 4: Identify Constraints

Teams are then asked to swap scenarios with each other. Each team then works collaboratively to identify design, environmental or end-user constraints under which they would accept different ethical balances in each of the provided scenarios. To assist in this activity, we suggest participants should consider the following questions:

- Which outcome, or balance of outcomes, would you prefer in this scenario?
- Would you accept any alternate outcome in this scenario if users were told beforehand that this

is how the system operates? What about if the general public were told beforehand?

- Do you think the trade-off in this scenario is appropriate given the corporate goals and strategy of the design organisation?
- Do you think the person benefitting from different balances of outcomes in this scenario has the moral right to do so?
- Could some of the ethical trade-offs described in this scenario be acceptable in a different environment? With different users? If these did not impact the same people?
- Is there more information which you would need in order to accept some of the possible ethical trade-offs in this scenario?

To gamify this step, each team is allocated points for every scenario in which they identify constraints that render at least two different balances of ethical principles acceptable. Teams should be asked to vote on whether they think these constraints are feasible to implement, and additional points allocated accordingly.

3.5 Use of Design Tool

Steps 3 – 4 are to be performed with the aid of a pre-prepared design tool, EETAS-Trade-Offs-for-You (EETAS-TOY), which represents the AS by a solid block and the relevant ethical principles as sliding bars, as in Figure 1. Participants connect bars end-to-end to represent ethical trade-offs and to discuss how different principles may be prioritized in each scenario.



Figure 1: The EETAS-TOY gamified tool.

Bars may be connected to other bars further down the structure to represent where a single ethical property (e.g. “maintains privacy”) is implicated in multiple trade-offs (e.g. in balance with both “maintains security” and “explains decisions”). Should participants wish to decouple the two bars representing these two trade-offs, this can be done by identifying a design requirement which permits the decoupling of those aspects of the design which can provide privacy at the cost of security, and privacy at the cost of explainability.

3.6 Recording Outcomes

The outcomes of each step are to be recorded using techniques such as mind-mapping (Beel, 2011). These records can then be used by the AS developers to identify further design requirements which enable or implement the constraints identified by each team. The EETAS-TOY tool itself may be retained by user organizations to aid in explaining ethical trade-offs or to act as a public record of the conversation.

4 PILOT STUDY VALIDATION

We conducted an initial pilot study workshop to investigate participants' perception of the EETAS process and its effect on their own understanding of ethical complexities in autonomous systems. As a preliminary pilot study, this workshop aimed to provide a partial validation of the EETAS process by establishing a link between EETAS participation and understanding of, and willingness to engage with, ethical complexities of AI. A further, orthogonal, aim of the study was to investigate the extent to which the EETAS-TOY tool was perceived as helpful in facilitating discussion and communication amongst participants about the ethical complexities and balances in autonomous systems.

The experiment was approved by the University of Hertfordshire's Health, Science, Engineering and Technology Ethics Committee under protocol number SPECS/SF/UH04940.

4.1 Pilot Study Design and Methodology

The pilot study was carried out at the University of Hertfordshire, with participants recruited following self-selection into the study. After obtaining consent, participants were randomly divided into teams of 4 – 5. The random assignment was performed by the researchers in order to mitigate against the confounding effects of team members knowing each other, or sharing demographic characteristics. All participants were provided with an overview of the purpose of the EETAS process and the workshop, but were not introduced to each individual step of the process in advance, in order to avoid anticipation of some of the discussion points.

All teams were given a high-level written descriptive specification of the robot chosen for consideration throughout the workshop: an assistive robot for use in a domestic environment. Owing to time constraints, a real-world robot prototype and

specification could not be sourced for the workshop, and instead the specification was produced by the researchers and based on previous work carried out at the University of Hertfordshire Robot House (Menon, 2019), (Koay, 2020), (Saunders, 2016).

Some functionality ascribed to the assistive robot within this written specification included:

- Moving about the house in response to user commands or actions
- Reminding the user to take medication
- Engaging the user in social interaction or conversation
- Notifying the user of hazardous conditions such as the oven being switched on
- Communicating with other smart device sensors in the house, including camera, audio and personal computers
- Communicating warnings to external medical monitoring systems regarding the health and activities of the user

Participants were given a set of pre-printed cards containing the eight ethical principles described in Section 3.2. Owing to time constraints, all teams were instructed to consider only these ethical principles throughout the workshop and not to expand their selection. Each team was also given an EETAS-TOY tool (Figure 1) and shown how this could be used to represent ethical trade-offs and balances.

Participants were provided with an initial questionnaire and asked to provide information on age, gender and whether they had any background relating to either design or robotics. They were also asked to rank the eight ethical principles in order of how important they considered each of them to be for the assistive robot under consideration.

Following this, teams were introduced to each other and took part in a small ice-breaker. They were then instructed to complete Steps 3 and 4 of the EETAS process, Steps 1 and 2 having been completed by the researchers (owing to time restrictions, the HAZOP guidewords were not used). Teams were allocated 20 – 35 minutes for each step, with the researchers indicating when the time for each step was nearing completion. All teams were given structured worksheets to record their identified scenarios (Step 3), and constraints (Step 4, following swapping of team records).

Teams worked simultaneously in different parts of the workshop room, with each team being observed and monitored by one of the researchers. The researchers were able to answer questions and remind participants of the requirements of each step, but did not contribute to the discussions or guide them in any way.

4.2 Post-Study Questionnaires

Following the workshop, participants were asked to complete some further post-study questionnaires. These included the following questions:

- Participants were asked to rank the eight ethical principles in order of how important they now considered each of them to be for the assistive robot
- Participants were asked to give a numerical score of how well they understood ethical trade-offs before the EETAS process, and how well they understood these trade-offs following EETAS (0 = not at all, to 5 = very well)
- Participants were asked to give a numerical score of how helpful they found the EETAS process in understanding ethical trade-offs (0 = unhelpful, to 5 = very helpful)
- Participants were asked to give a numerical score of the EETAS-TOY tool in a) understanding and b) communicating about ethical trade-offs (0 = unhelpful, to 5 = very helpful)

4.3 Pilot Study Results

As this was a preliminary study, with correspondingly low participant numbers (<20), no statistical significance between conditions and questionnaire responses was expected. Nevertheless, there were indicative trends to support our hypothesis of a causal relationship between participation in the EETAS process and improved public understanding of AI ethical complexities.

4.3.1 Participant Demographics

Participant selection was strongly biased towards both design and robotics, with 93% of participants identifying as having a background in design, and 43% a background in robotics. This commonality in background is due to constraints around the recruitment and identification of participants, with most participants sourced via existing connections to the University of Hertfordshire. The age range of participants was 19 – 61 years old, with the average age being 37. The gender balance was roughly equal, with 57% male participants and 43% female.

4.3.2 Participant Responses to EETAS

As may be expected, prior to the workshop participants without a robotics background rated their existing understanding of ethical trade-offs in autonomous systems as lower (mean value 2.6) than

those with a robotics background (mean value 3.4). Post-workshop, the gap had narrowed, with those from a non-robotics background rating their understanding of ethical trade-offs as an average of 3.8, compared with 4.2 for those from a robotics background. This corresponds to an increase of 58% greater improvement in understanding ethical trade-offs for those without a robotics background, as compared to those with.

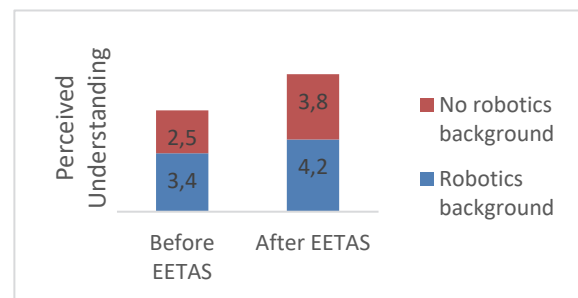


Figure 2: Change in perceived understanding of ethical trade-offs as a result of EETAS participation.

When asked to identify how helpful the process was in understanding ethical trade-offs (0 = unhelpful, to 5 = very helpful), 94% of participants ranked the helpfulness of the EETAS process at 3 or above, with the mean ranking being 3.7. Interestingly, there was no difference noted in this result between those with a robotics background and those without.

When asked about the helpfulness of the EETAS-TOY tool in identifying ethical trade-offs, 64% of participants ranked this as 3 or above (mean value 3.3) and when asked about the helpfulness of the tool in discussing ethical trade-offs, 71% of participants ranked this as 3 or above (mean value 3.7). In contrast to the scores for the perceived helpfulness of EETAS, which were independent of background, those without a robotics background considered the tool more helpful than those with.

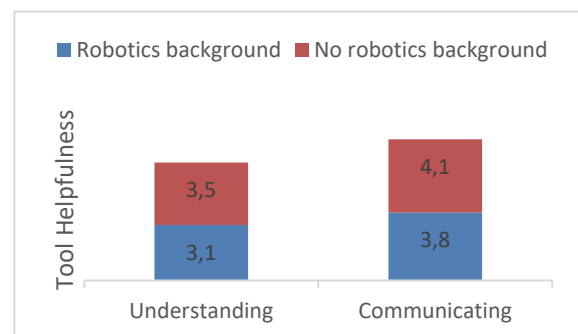


Figure 3: Perceived helpfulness of the EETAS-TOY tool in understanding and communicating.

4.3.3 Observations and Timing

Three teams were each monitored by a researcher, who recorded the total time taken for each step in the EETAS process, and the total time spent using the EETAS-TOY tool. Only time spent actively and purposely using the tool in discussion was recorded, and time spent “fidgeting” with the tool or learning how to use it was discarded.

On average, teams used the EETAS-TOY tool during 45% of the time they were engaged in Step 3 (scenario construction). Teams did not use the tool to any significant extent in Step 4 (identifying constraints). There was no correlation noted between background in either design or robotics and readiness to engage with the tool.

4.3.4 Discussion and Indicative Trends

Although no statistically significant conclusions can be drawn, the results demonstrate some potential indicative trends. Firstly, the EETAS process was considered by a large majority of the participants to be helpful in understanding and discussing ethical trade-offs. This was not correlated with prior experience: those with a robotics background found it to be as helpful as those without. This supports our initial hypothesis that EETAS can be used to improve public understanding of, and engagement with, ethical complexities in autonomous systems.

Moreover, all participants considered that their understanding of ethical trade-offs in autonomous systems had increased following participation in the EETAS process. In this case the extent of the effect could be seen to be correlated with prior experience: those without a robotics background considered that their increase in understanding was greater than those with. This indicates that the EETAS process may serve a useful purpose in raising understanding of autonomous system ethics amongst those who have traditionally been marginalised, or excluded from, existing conversations around AI.

Finally, participants considered the EETAS-TOY design tool to be useful in identifying ethical trade-offs, and discussing these within their teams. Observational monitoring supported an indication that the tool appears to stimulate positive interaction amongst participants by providing a physical aid to visualise trade-offs. We consider it likely that the tangible element of the tool is of value here in supporting participants in abstract reasoning and discussion of unfamiliar concepts.

5 CONCLUSIONS

We have presented a structured, collaborative process for improving public engagement and understanding of ethical complexities in autonomous systems, and for consequently informing the design of these systems. Our results from an initial pilot study workshop indicate that the process improves understanding of ethical trade-offs, most significantly for those who do not have prior experience in robotics or autonomous systems. Our results also indicate that the process is helpful in stimulating debate and discussion around ethical trade-offs, particularly amongst groups of people who have a varied prior understanding of the issues.

We have also considered the use of a physical design tool, the EETAS-TOY tool, in helping participants understand and communicate about ethical trade-offs in a specified system. Our results indicate that participants consider this a helpful physical aid to concretize some of the more abstract concepts, an effect which is more pronounced amongst those without a background in robotics or autonomous systems.

In terms of next steps, we plan to run a larger workshop involving developers and a more varied participant group (e.g. stakeholders, regulators, end-users). We intend to use a real-world robot prototype within this workshop, in order to assess the effectiveness of EETAS in producing outcomes which can be used by the developers to inform the design of the robot.

We also plan to explore the design space of the EETAS-TOY tool more fully. We anticipate developing multiple versions and configurations of the tool, in order to assess the beneficial effects of different tools on participant understanding. New versions will include tools with automated mechanisms to adapt to participant choices, tools with different shapes to assess the importance of visual balance and tools with simplified shapes and interaction methods to aid accessibility.

REFERENCES

- Akinsanmi, T., Salami, A. (2021). *Evaluating the trade-off between privacy, public health safety, and digital security in a pandemic*, In Data and Policy 3(27).
- Ballard, S., Chappell, K.M. and Kennedy, K. (2019). *Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology*. In Proceedings of the 2019 Designing Interactive Systems Conference, pp. 421-433.

- Beel J., Langer, S. (2011). *An Exploratory Analysis of Mind Maps*. In Proceedings of the 11th ACM Symposium on Document Engineering, pp. 81-84.
- British Standards Institute. (2016). *Guide to the ethical design and application of robots and robotic systems*, BSI 8611.
- British Standards Institute (2016). *Hazard and Operability Studies: Application Guide*, BSI 61882.
- Curral L., Forrester, R., Dawson, J., West, M. (2001). *It's What You Do And The Way You Do It: Team Task, Team Size and Innovation-related Group Processes*. In European Journal of Work and Organizational Psychology, 10(2), pp. 187-204.
- Foot, P. (1967). *The problem of abortion and the doctrine of double effect*, Oxford Review 5, 5—16.
- Hancock, P., Nourbakhsh, I., Steward, J. (2016). *On the future of transportation in an era of automated and autonomous vehicles*. In Proceedings of the National Academy of Sciences, 116(16), pp 7684 – 7691.
- Institute of Electrical and Electronic Engineers (2018). *Ethically Aligned Design*, IEEE, v2, https://standards.ieee.org/wpcontent/uploads/import/documents/other/ead_v2.pdf
- IET, (2019). *Code of Practice: Cyber Security and Safety*, Available at <https://electrical.theiet.org/guidance-codes-of-practice/publications-by-category/cyber-security/code-of-practice-cyber-security-and-safety/>
- Kahneman, D., & Tversky, A. (1979). *Prospect theory: An analysis of decision under risk*. *Econometrica: Journal of the Econometric Society*, pp. 263-291.
- Koay, K., Syrdal, D., Dautenhahn, K., Walters, M..(2020). *A narrative approach to human-robot interaction prototyping for companion robots*, in Paladyn, Journal of Behavioural Robotics, 11, pp. 66 – 85.
- Larson, K. (2020). *Serious games and gamification in the corporate training environment: A literature review*. *TechTrends*, 64(2), pp.319-328.
- Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*, The Alan Turing Institute. <https://doi.org/10.5281/zenodo.3240529>.
- Lee, L., Lee, H. (2020). *Tracing surveillance and auto-regulation in Singapore: 'smart' responses to COVID-19*. In *Media International Australia*, 177(1), pp. 47-60.
- Lin, P. (2015). *Why Ethics Matters for Autonomous Cars*. In *Autonomes Fahren*, Springer Vieweg, pp 69 –85.
- Malizia, A., Carta, S., Turchi, T., Crivellaro, C. (2022). *MiniCoDe Workshops: Minimise Algorithmic Bias in Collaborative Decision Making with Design Fiction*. In Proceedings of the Hybrid Human Artificial Intelligence Conference.
- National Cyber Security Centre (2019). *Intelligent Security Tools*, Available at <https://www.ncsc.gov.uk/collection/intelligentsecurity-tools>
- Park, J., Hong, E., Le, H. (2021). *Adopting autonomous vehicles: The moderating effects of demographic variables*. In *Journal of Retailing and Consumer Services*, 63.
- Rossi, F., Mattei, N. (2019). *Building ethically bounded AI*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 9785-9789.
- Saunders, J., Syrdal, D., Koay, K., Burke, N., Dautenhahn, K. (2016). *Teach Me - Show Me' - End-user personalisation of a smart home and companion robot*. In *IEEE Transactions on Human-Machine Systems*, 46(1), pp. 27–40.
- Schrier, K., (2019). *Designing games for moral learning and knowledge building*. *Games and Culture*, 14(4), pp.306-343.
- Thornton, S. (2018). *Autonomous vehicle motion planning with ethical considerations*, (Doctoral dissertation), Stanford University.
- Von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behaviour*. Princeton University Press.

AUTHOR INDEX

-
- | | | |
|--------------------------------------|---------------------------------|--------------------------------|
| Ademola, A. 89 | Goulding, J. 225 | Pupczyk, D. 97 |
| Almeida, M. 81 | Harder, T. 244 | Ramos, F. 81 |
| Antunes, M. 81 | Harvey, J. 225 | Reedman-Flint, D. 225 |
| Bach, M. 97 | Holtzblatt, K. 17 | Ripamonti, L. 59 |
| Bailey, J. 209 | Jensen, P. 97 | Rodríguez-Obregón, J. 179 |
| Baras, K. 49 | Jin, J. 187 | Romikaitytė, A. 217 |
| Baratè, A. 121 | Jonstrup, A. 32 | Ronen, H. 201 |
| Blackmore, K. 209 | Katz, A. 201 | Sacramento, E. 81 |
| Blume, P. 236 | Kim, H. 89 | Salamanca-Sánchez, D. ... 179 |
| Boccignone, G. 59 | Kim, Y. 89 | Sandkvist, E. 32 |
| Borghese, N. 59 | King, R. 209 | Sarabia, J. 171 |
| Bracke, S. 41 | Klein, S. 137 | Schwarz, F. 70 |
| Brambilla, S. 59 | Korsten, H. 121 | Sellen, A. 5 |
| Branco, A. 81 | Kowol, K. 41 | Serrano, S. 23 |
| Carrera-Rivera, A. 113 | Larrinaga, F. 113 | Shichel, I. 201 |
| Carta, S. 249 | Lasa, G. 113 | Smirnov, A. 129 |
| Carvalho, D. 81 | Levashova, T. 129 | Sol, R. 49 |
| Chessum, K. 145 | Liu, H. 145 | Sotelo, M. 23 |
| Daza, I. 23 | Llorca, D. 23 | Stanci, S. 217 |
| De Andrés, J. 163, 217 | Lucaci, A. 32, 97 | Steinø, N. 32 |
| Diaz, S. 171 | Ludovico, L. 121 | Strugeon, E. 163 |
| Dix, A. 7 | Madsen, C. 32, 97 | Sweeney, L. 236 |
| Drzyzga, G. 244 | Marçal de Oliveira, K. 163 | Tymoshchuk, O. 81 |
| Feijóo-García, M. ... 179, 193 | Martinez-Arellano, G. 113 | Valoyes-Porras, D. 179 |
| Feinleb, M. 70 | Mason, K. 236 | Vinnikov, M. 155 |
| Fernández-Lanvin, D. 163, 217 | Meixner, G. 70 | Wasinger, R. 137 |
| Fetscher, D. 137 | Menon, C. 249 | Watanabe, K. 105 |
| Foerster, F. 249 | Motahari, K. 155 | Wenzel, T. 137 |
| Frommholz, I. 145 | Oliveira, E. 81 | Winter, M. 236 |
| Gallula, D. 201 | Omata, M. 105 | Yigit, H. 163 |
| Golubski, W. 137 | Pedro, L. 81 | Zheng, M. 187 |
| González-Daza, B. 193 | Perez, J. 171 | Zubizarreta, A. 171 |
| González-Rodríguez, M. . 163,
217 | Priestnall, G. 225 | |
| Gottschalk, H. 41 | | |
-



Proceedings of CHIRA 2022

6th International Conference on Computer-Human Interaction Research and Applications

<https://chira.scitevents.org>

LOGISTICS:



EVENT MANAGEMENT SYSTEM:



IN COOPERATION WITH:



PROCEEDINGS WILL BE SUBMITTED FOR INDEXATION BY:



Copyright © 2022 by SCITEPRESS - Science and Technology Publications, Lda. All Rights Reserved

ISSN: 2184-3244

ISBN: 978-989-758-609-5