

Towards an Evaluation of Graphical User Interfaces Aesthetics based on Metrics

Mathieu Zen, Jean Vanderdonckt

Louvain School of Management

Université catholique de Louvain

Louvain-la-Neuve, Belgium

mathieu.zen@uclouvain.be, jean.vanderdonckt@uclouvain.be

Abstract—The graphical user interface (GUI) of an interactive system is nowadays the most frequently used interaction modality. While the contents are of high importance, the Look and Feel is an equally essential factor determining the GUI quality that is impacted by several determinants such as but not limited to aesthetics, pleasurability, fun, etc. Therefore, GUIs aesthetics is a potential element to focus on in order to facilitate communication between device and user. On that basis, one question that comes up is: “Is it possible to evaluate the quality of a GUI by estimating its aesthetics through a series of measurable geometric metrics?”. This paper suggests possible directions to address the previous question by, first, introducing a simplifying model of GUIs aesthetics that captures aesthetics aspects and regions-related metrics. In a second phase, a methodology for the evaluation of GUIs aesthetics is defined based on the underlying model. The paper finally puts forwards a model-based implementation of the aforementioned methodology in the form of a web service tool for metrics-based evaluation of GUIs and discuss the results of a survey on users aesthetics perceptions.

Keywords - user interface aesthetics, metrics-based evaluation

I. INTRODUCTION

Graphical User Interfaces (GUIs) have nowadays a large presence in humans everyday life and this is not likely to decrease. According to Cisco’s white paper about mobile data traffic forecast, the number of mobile devices per capita will amount to 1.4 by 2018, therefore exceeding the number of people on earth [1]. Whether it is for working on a computer at the office, or for entertaining on a video game console, or even for communicating with others on a smartphone and simultaneously entering an address in a G.P.S. device while driving a car, people are constantly confronted with GUIs for providing electronic information.

Considered as a medium, GUIs are intended to convey a message. In the first place they are designed to attract users eyes and secondly to make them consult it easily and efficiently. While the contents are of high importance, the form is equally essential. Indeed, design appreciation is the initial interaction people have with an interface when they first face it – some works report that this interaction takes less than half a second to be completed [2, 3]. Moreover, there is a strong indication that perceived quality of a UI has a positive impact on the idea users have about the system’s usefulness and usability [4]. Other works in human-computer interaction have shown that UI conception represents a substantial part of software development [5]. Therefore, the *Look and Feel* of GUIs, or in other words *GUIs aesthetics* is a potential target

for optimization when it comes to improve the communication between device and user.

By focusing on the improvement of this process, designers would be able to provide a more optimized output that comes closer to the final UI. Therefore, this paper brings a reflection on “How to model GUIs aesthetics evaluation?” and on the underlying concepts that have to be taken into account to set up such a model. Our hypothesis is that a systematic approach to GUI aesthetics is possible, by using a set of objective criteria (i.e. metrics) to generate recommendations. These criteria would be derived from and be validated by a series of experiments, such as the one presented in this work.

Concretely, this could take the form of an evaluation of GUIs aesthetics based on metrics. This all seems fairly straightforward as the only questions to address are the generation of recommendations based on UI aesthetics evaluation metrics and, as a continuation, the implementation of these recommendations in a new design. However, these issues conceal a more complex one: aesthetics modeling. As suggested by Fig. 1, it is possible to fragment a GUI in several regions and compute design arrangement measures [6]. Yet, proceeding in such a manner inevitably induces a lack of validity as the results do not rely on a proper model of aesthetics.

The remainder of this paper is as follows: In the next section we review the scientific literature regarding different concepts of aesthetics measurement such as visual techniques and aesthetics measures. In Section 3 we propose a model for GUI aesthetics evaluation and in Section 4 an associated

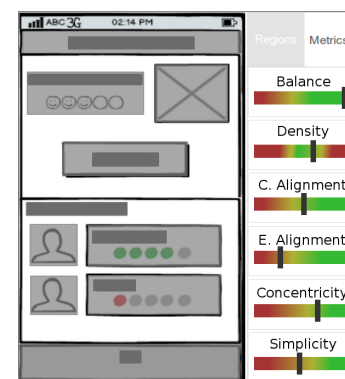


Fig. 1: Metrics-based evaluation of a wireframe prototype.

method for evaluating different types of interfaces. In Section 5 we present a brief description of an implementation of the method in the form of a webservice. In Section 6 we present the results of a preliminary survey carried out to investigate users aesthetics perceptions, and in Section 7 we conclude with some final remarks and future work.

II. STATE OF THE ART AND RELATED LITERATURE

A. What is aesthetics?

Before considering a problem, it is of good note to first describe the concepts that are questioned. One may ask what is actually aesthetics? And why is it so important to be considered in computer science?

Aesthetics, or beauty, is a complex subject and its definition tend to vary among authors, certainly due to its subjective nature. One can find different definitions in the literature:

- *Baumgarten 1750* [7] : The first definition of the term in modern philosophy comes from his book “Aesthetica”. “The end of aesthetics is the perfection of sense cognition as such”. It is strictly a quest for perfection, implying that all kind of imperfection must be avoided.
- *Clay 1908* [8] : Aesthetics is the criterion by which beautiful things are to be judged. It can be expressed as a degree with which things approximate an ideal standard.
- *van Damme 1996* [9] : There exists a concept of universal aesthetics seen as the existence of a standard of beauty that is shared between cultures, races and genders. If it cannot be proved for all characteristics of forms, some aspects such as symmetry, smoothness, brightness and balance can be regarded as widespread aesthetics principles. Forms manifesting these principles can be appreciated both interculturality and transculturally.
- *Lavie and Tractinsky 2004* [10] : Aesthetics user’s perceptions encompass two main dimensions, “classical aesthetics” and “expressive aesthetics”. While “classical aesthetics” refers to standard principles of beauty such as sequence, tidiness and symmetry, “expressive aesthetics” represents qualities that go beyond the classical principles and that stress the designer’s creativity and expressive power. “Originality”, “fascinating design” and “special effects” are non-standard design aspects that are part of this category.

Through several studies in multiple fields, researchers have shown that aesthetics has an impact on preferences [11], usability perception [12, 13, 14], credibility [15] and performance [16]. Not only interface design itself has an impact on its perceived usability [13] but the design of the device displaying the interface as well [12]. Furthermore, interface design must be chosen adequately with respect to the nature of a targeted audience – some works indicate that, in website design, the use of vertical symmetry is often preferred by designers when targeting a male audience [11] while for targeting women there is a preference for rounded lines, various font colors and for the use of “sweeter” colors such as pink and yellow [17].

Obviously, GUI aesthetics matters; hence, it is clearly subject to evaluation. In this respect, a qualitative approach can be used to evaluate UI visual quality. This is the topic we are investigating in the following subsection.

B. A qualitative approach

Aesthetics in Computer Science refers mainly to UI design, where requirements can typically concern the overall usability of the system as well as specific aesthetic considerations. Yet, a distinction can be drawn between aesthetics and usability [18]. If usability requirements should be focused on how to design UIs (leading to more effective and productive tools), aesthetics considerations, on their side, should serve to improve satisfaction, User eXperience (UX) and social acceptance.

Ideally, these requirements need to be discussed in the light of popular usability and aesthetics recommendations – resulting from guides and various software related to a usability and involving specific guidelines [19, 20] or related to a more aesthetic based approach focusing on the needs to respect a set of visual techniques [21, 22, 23, 24]. The latter approach implies not only a qualitative evaluation considering how a design fulfill a set of visual criteria [25] but also a quantitative evaluation with metrics, all of which specific to a design aesthetics property [26, 27].

One concept we use in this paper and that was first introduced by Vanderdonckt [25] is that of *visual technique* and is defined as follows:

A *visual technique* relies on a commonly accepted visual principle to suggest the arrangement of the layout frame components.

One property that contributes mainly to design aesthetics and can be considered as a visual technique in the strict meaning of the term is the **visual balance** – a widely accepted principle in art techniques. To better understand the principle, let’s consider the following illustration [28] shown in Fig. 2, a plain black disk placed on a white rectangle.

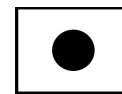


Fig. 2: This disk is centered. We can “see it at a glance”.

Using a ruler, we could easily and quite accurately determine the distance separating the center of the disk from the edges of the square. And, therefore conclude that the disk is right in the center. But, do we really need to do all this? Of course not, thanks to our capacity to feel balance in our environment, we can “see it at a glance” and conversely, we do not need to measure exactly to see that a figure is slightly instable. Another analogy is the man that stands on his legs [29]. It is this perception of balance that allows us to keep firmly attached to the ground and remain upright. Even if it looks stupid and so automatic, there is no measure as fast and as accurate as the intuitive sense of balance. In a majority of discipline related to design thinking, visual balance is considered as the designer’s compass to measure what “looks

right” and what “looks bad” [30]. Therefore, a qualitative definition for the visual technique of balance is:

Visual balance is a search for equilibrium along a vertical or horizontal axis in the layout objects weights.

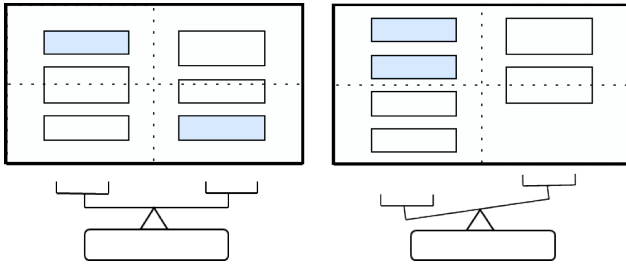


Fig. 3: Example of horizontally balanced and unbalanced interfaces

This definition is illustrated in Fig. 3 showing an example of balance and unbalanced interface. It is clear that the screen on the right can be characterized as unbalanced since it presents more objects in the left part of the screen. Additionally, two of its rectangles are emphasized adding even more weight to this part of the screen. On the opposite, the screen on the left is quite balanced since all of its quadrant have similar weights. The screen is thus at equilibrium both vertically and horizontally.

In the context of a metrics-based UI evaluation approach, the goal is to obtain comparable properties in order to confront different UIs together and assess which one could be the most attractive or preferred one from a user’s perspective. This is why the issue of aesthetics quantification, subject covered in the next subsection, is a logical sequel in our attempt to identify literature surrounding UI aesthetics modelling.

C. Quantifying aesthetics

The quantification of aesthetics through mathematical formulae is not new as a goal, as it is demonstrated by the existence of the Golden Section [10], a proportion considered as the number that would ideally represent beauty through order. This number has been presumably considered, often erroneously [31], as widely used by artists, mathematicians, sculptors and architects in their works. Fig. 4 presents how the designers of the Parthenon were presumably influenced by the Golden Section. This antique example is clearly not conducive to give credit to aesthetics quantification. Indeed, while the aesthetic property of the golden ratio has always been speculated, it has never been suitably demonstrated [32, 33, 34]. Hence the need to rely on a model that captures specific aesthetics aspects in a comprehensive way – one that express them in quantitative results representing the degree of users perceptions.

For the quantification of UI aesthetics, the method remains quite similar in the sense that it is necessary to define a set of visual properties that can be quantified with a fair degree of accuracy. Of course, the presence of these visual properties have to positively affect aesthetics. In the following lines, we introduce the concept of *aesthetics metric*. It is first Ngo [26]

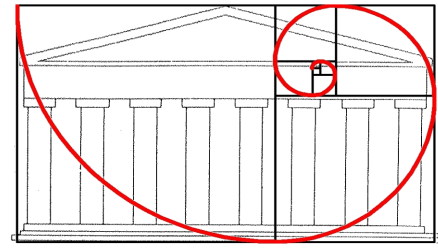


Fig. 4: The parthenon is presumed built according to the golden section.

who uses this concept with the name of *aesthetics measure* but without defining it properly. A definition can be found by confronting this *aesthetics measure* to a close concept we previously defined, the *visual technique* [25].

An *aesthetics metric* is the application of a mathematical formula based on a visual principle resulting in a quantitatively descriptive measure.

While a *visual technique* is considered as a GUI specific visual principle and refers mainly to a purely theoretical and qualitative concept, an *aesthetics metric* evaluates the arrangement of objects in a layout via a mathematical formula that provides when computed a quantitative measure [6]. The measure that is computed through the formula is an element that is interesting to analyze whether it is to attempt to quantify usability, aesthetics or even another GUIs property. The interest of manipulating metrics is that it brings a more rigorous and coherent approach to evaluate aesthetics.

Visual balance is a suitable example to support the aforementioned concepts of visual technique and aesthetics metric. So, as it has been proven that balanced interfaces have a positive impact on user’s aesthetics perception [35], the visual technique of balance is deemed to be a vital element in interface design. Moreover, it is possible to quantify the balance property of an interface. This is what Ngo did by providing a balance metric and 13 other metrics each linked to an interface visual property [26]. Afterwards, a number of studies have been conducted to add empirical validation and support the previous metrics [36, 37, 38, 39].

Over the years, several researchers have studied the possibility to use metrics to assist design process; in the form of objective tools which roles are to guide the designers through a semi-automatic evaluation process (AIDE) [40], allow them to specify design constraints among interaction objects (DON) [41] or even provide a family of consistency checking features (SHERLOCK) [23], or in the form of a suite of aesthetics metrics and associated formulae such as balance, linearity, sequence and orthogonality [26, 27].

Very first approaches were primarily focused on alphanumeric displays [42] and proposed already basic quantitative measures such as balance, symmetry and alignment. Others have tried to establish an objective metrics-based model including various aesthetics measures [36, 37]. On the one hand, a large number of measures provides a more robust model but generates a more complex practical application,

raising a problem of model conciseness [36]. On the other hand, selecting only few metrics may lead to statistically not significant results due to an uncomplete model. [37] This is why it may be appropriate to explore the effect of individual measures - or a small set of measures at the most - on global aesthetics as suggested by [43].

Some authors draw a distinction between the visual metrics - relying on the visual arrangement of the GUIs objects - and the structural metrics [44, 45, 46, 47] - relying on the basic architecture and structural code of a GUI and often related to a usability purpose. The latter includes particularly a subset of UI elements measurements [44, 45, 47] such as the number of controls, words, fonts, colors, etc. and a subset of UI performance measurements [46] such as task efficiency, readability, load time, etc. Regarding the visual metrics, some measures besides Ngo's formulas are also often discussed in litterature such as layout complexity [13, 48, 49, 50, 51] and elements proximity [52, 53].

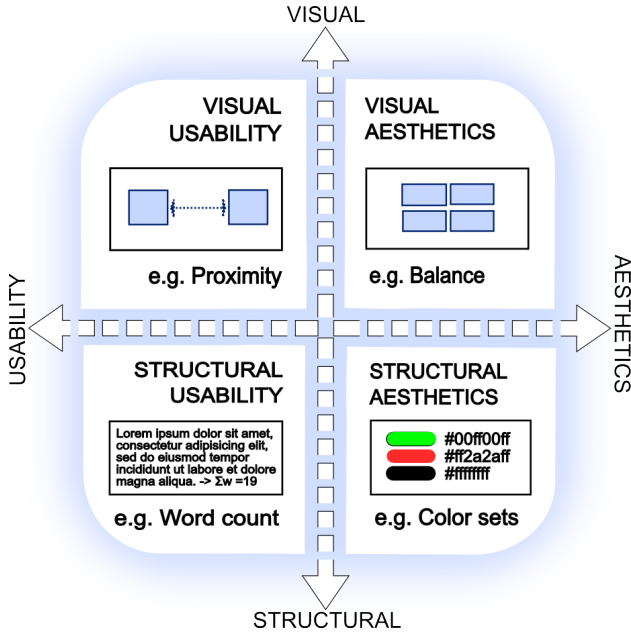


Fig. 5: A metrics map based on two main axes: aesthetics/usability and visual/structural.

As presented in Fig. 5, a UI metric can be spotted on a map along two main axes. The vertical axis determines the extent to which a metric is built upon structural or visual UI elements while the horizontal axis determines the metrics main contribution to UI, that is either usability-oriented or aesthetics-oriented. For example, a metric of visual balance as stated is a visual metric that uses elements that are part of the interface view (position, area, etc.) and the aim of which is to give feedback on UI aesthetics, this makes it a *visual aesthetics* metric. However, the categorization remains disputable since the boundaries are porous and a balance metric could obviously well be computed using structural elements (e.g. interface objects attributes). Similarly, the number of color sets - classified as a structural aesthetics metric because it can be computed based on the UI structural code - could also be considered as a visual aesthetics metric.

The difficulty to provide an acceptable model of aesthetics measurements resides in the fact that most of the metrics are highly correlated because they are often the resulting transformation of the same inputs attributes. Indeed, we can easily assume that a symmetric interface has to be balanced but we cannot conclude the opposite. Moreover, some metrics may also be combined to various formulas changing the final result and inducing another interpretation. It is the case for balance which can be computed through a comparison of quadrants weights [26], a pixel-based weightmap [30] or even a comparison between the UI objects center of mass and the screen center [42] (Fig. 6). If a model integrates this metric, it must take into account the formula that has been used for its measurement.

Balance computed using	
Ngo Formula	Center of Mass Formula
$BM = 1 - \frac{ BM_{vertical} + BM_{horizontal} }{2}$	$BM = 1 - \frac{2 * \sqrt{(X_{CS} - X_{CM})^2 + (Y_{CS} - Y_{CM})^2}}{\sqrt{W^2 + H^2}}$

Fig. 6: Several possible formulas for several metrics of balance

Our ambition is to gather different metrics already defined by authors and introduce new metrics in order to propose a multi-metric solution that relies on a corpus of different formulas. Therefore, we propose for instance a concentricity metric as exposed in Fig. 7 which measures how the interface objects are gathered in the center of the interface or rather kept in the corners.

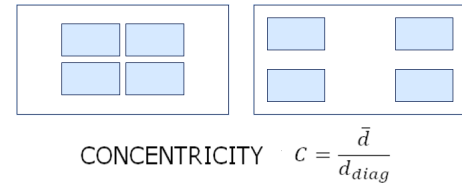


Fig. 7: Illustration of concentricity measurement

It could be measured using the average of distances between the interface center and each object center. Inventing new metrics is not difficult per se since it is a simple combination of a visual perception and a set of graphical attributes. The most challenging part concerns its validity, a problem dealing with issues like how to objectify a concept which is the fruit of one - or few - person's labor. Indeed, it is necessary to confront the metric with human's perception and adjust it in order to reach an acceptable level of objectivity. If this is not completed appropriately, the metric will remain subjective, inducing a lack of generalizability. In addition, if a metric is kept founded upon personal intuitions, it will for sure arouse a variety of possible interpretations, then deviating from its original purpose which is to provide designers with clear and normative UI aesthetics evaluation items.

III. MODEL

In order to tackle an evaluation approach based on metrics, we established a class diagram model illustrated in Fig. 8

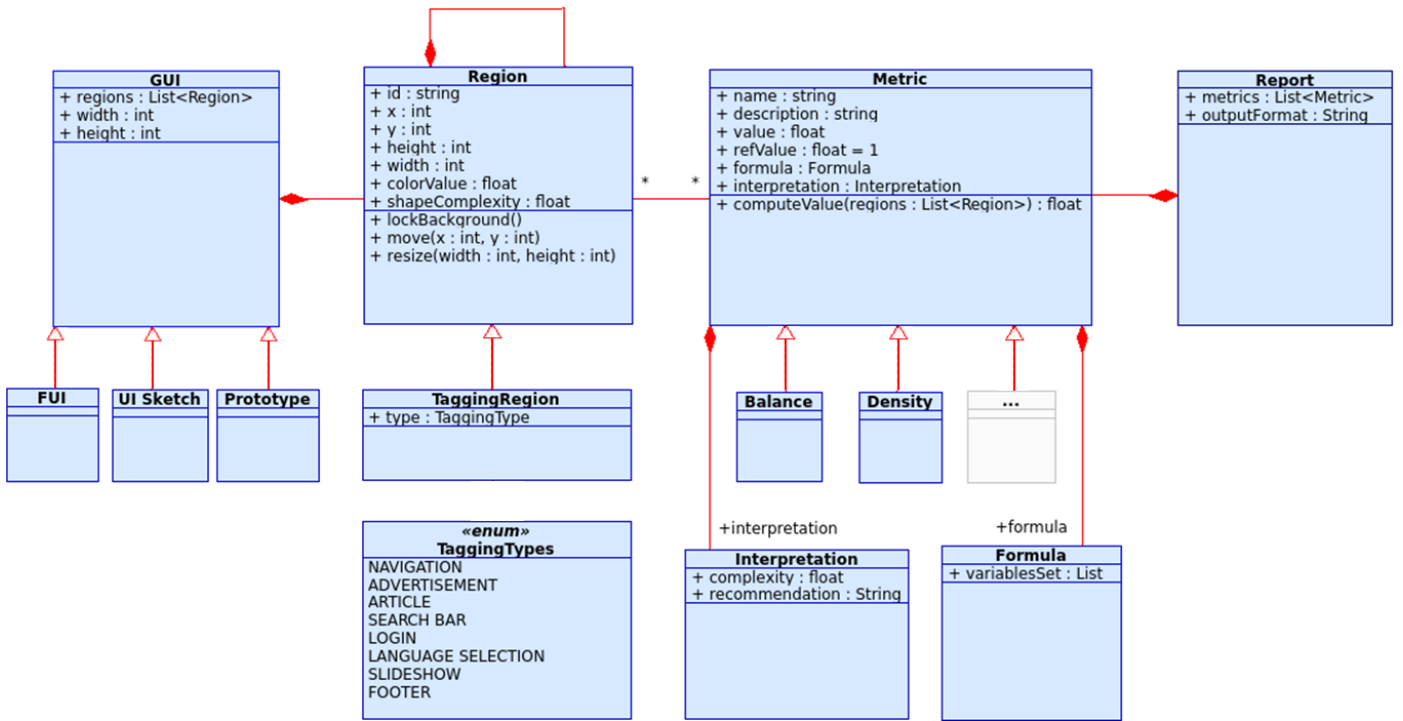


Fig. 8: Modeling of user interface aesthetics evaluation

encompassing all relevant concepts. By this term, we mean all object abstractions that are somehow related to the process of UI evaluation and for which it makes sense to be integrated in our model. The concepts are the following:

- A *Graphical User Interface (GUI)* is a visual platform representing a way for humans - i.e. users - to interact with a device. It consists generally on a set of graphical items (windows, buttons, menus, etc.) with which users interact through the use of one or several modalities such as mouse, keyboard, touchscreen, voice recognition system, etc.
- A *Final User Interface (FUI)* is a concept resulting from the Cameleon Reference Framework [54]. It consists of source code, in any programming language or mark-up language (e.g. HTML5) interpreted and compiled in order to represent a graphical design. This resulting graphical design is highly dependant on the environment (e.g. browser). The FUI is therefore split into two sublevels: the source code and the running interface. As an evaluation approach intends to be the most independant possible regarding the source code used to implement an interface, it is mainly the running interface that is used in our approach
- A *UI Sketch* is a low fidelity hand-drawn UI prototype commonly used by designers in the early stage of interface creation process due to its rapidity and flexibility of use. Typically, UI sketching activities are performed on paper.
- A *Wireframe* is a prototype that represents the skeletal framework of a UI with a higher degree of fidelity than a simple sketch. It includes accurate elements that give a better visibility of the expected UI. A series of tools allow to design wireframe providing the user with common design patterns to achieve a prototype that is the closest one to the actual UI.
- A *Region* represents a UI specific area. This area is defined using different two-dimensions attributes which are X, Y, width and height expressed as integers representing the pixels. Moreover, a *Region* can be characterized with a color and a shape complexity value and can be part of another region; hence the recursive association. A *TaggingRegion* is one of its specification including a tagging type that describes its intrinsic goal (e.g. Advertisement).
- A *Metric* is an attempt to quantify a visual technique characterizing the arrangement of the layout frame components in a UI (e.g. Balance). It basically takes a set of regions as an input, runs a computation through a pre-conceived formula (each one uses a different set of region attributes) and returns a numeric value between 0 and 1. This value is then assumed to quantitatively represent the UI score for a visual property.
- A *Report* as its name suggests, encompass a set of metrics and presents them in a end user perspective. In addition to the numeric values, a *Report* propose a description for each metric involved and some recommendations on how to improve the evaluated UI based on the latter values.
- *Formula* and *Interpretation* are strongly linked with the concept of aesthetics metric. Indeed, a metric always consists of at least one formula, the aim of

which is to achieve on one hand a computation of GUI arrangement figures, and at least one interpretation, whose purpose on the other hand is to draw conclusions about the metric value.

IV. INFRARED: A METHOD FOR UI EVALUATION

With respect to the previous metric-based evaluation model, we propose INFRARED, a method to analyze and evaluate UIs with a pretty simple and potentially iterative process in 4-steps based on the following sequence: select InterFace, draw Regions, analyze Aesthetics and REDesign user interface.

- 1) Select interface - choose a UI in a defined context (screen resolution, rendering application, etc.)
- 2) Draw regions - detect and manually draw regions of interest on top of this interface in order to retrieve basic forms attributes (e.g. position, dimensions) from the geometric plane.
- 3) Analyze aesthetics - perform aesthetics metrics computations with regions attributes, discuss and interpret results obtained in a metrics report.
- 4) Redesign UI - proceed to design modifications considering the previously established report.

It is obvious that this process can be repeated until the designed UI reaches a permissible level of satisfaction. Fig. 9 illustrates this iterative evaluation and improvement process through a concrete example.

First of all, a UI is selected (1). Basically, it could be a website but also all kind of displays integrating visual components (e.g. mobile UI, desktop UI, slides, posters).

Then, to allow for an evaluation of layout arrangement, regions of interest representing the interface objects (IO) are defined on top of the UI (2). Regions of interest represent, inter alia, navigation bars, text paragraphs, ads and pictures; thus, any kind of visible IO with a specific static or dynamic purpose. It should even be conceivable to adopt a recursive process by defining regions into regions.

From the regions attributes (i.e. coordinates, dimensions, etc.), metric analytics are automatically computed and presented to the designer in the form of recommendations (3). Typically, the recommendations are part of a metrics report indicating for each metric - using gradient colour codes - if its value is high (green) or low (red). The level of acceptable threshold depends on the metric in question and can be shade scaled (e.g. balance) or shade centered (e.g. density).

This finally provides the designer with a feedback helping him to achieve the design of a new interface respecting aesthetics principles while keeping the same original design idea (4).

V. TOOL IMPLEMENTATION

A tool is currently being developed with the aim of automating the INFRARED process. QUESTIM (Quality Estimator using Metrics) is a Web-based evaluator tool that allows for evaluating UI quality using aesthetics metrics [6]. It is a direct implementation of the aforementioned model and process in a webservice generated with Google Web Toolkit (GWT). The

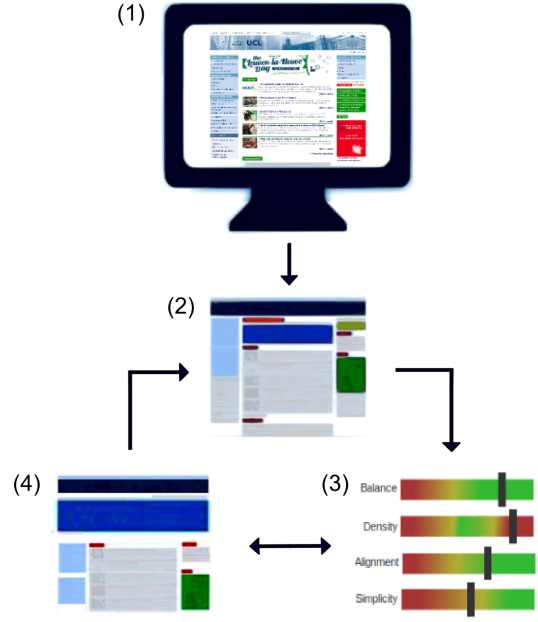


Fig. 9: INFRARED method illustration

source code is written in Java and compiled to JavaScript. The main advantages of using this framework are the following:

- it allies the performance of a Java application and the flexibility of an AJAX application that runs across all browsers, including mobiles;
- the programme can immediately be deployed and published on an application cloud provided by Google App Engine.

QUESTIM is accessible via the following URL: <http://questimapp.appspot.com> and its interface consists of two main screens. The first one, accessible when you first connect to the webpage, contains a form that the user has to complete with two possible inputs. Either he/she indicates a website URL or he/she upload an interface screenshot. When this is done, the user presses the button “continue” to access the second page.

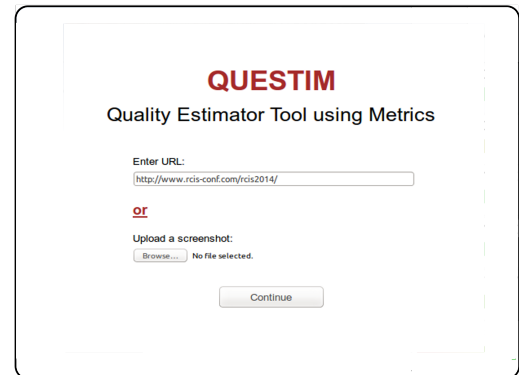


Fig. 10: Homepage of the tool QUESTIM

On this second page, several information are presented to the user:

- a *drawable canvas* in the center of the screen with in background, the webpage or screenshot previously loaded as shown on Fig. 11. This is the main part of the tool where the user is able to draw regions on top of the webpage or modify existing drawn regions;
- a *vertical tool panel* on the far left part containing buttons offering certain functionalities such as zoom, browse - allowing to browse the webpage - and capture - a feature which captures a screenshot of the website as it is exposed at that moment and give the possibility afterwards to directly modify superposed screenshot regions in order to redesign IO while receiving an immediate visual feedback;
- a *vertical analytics panel* on the far right part of the interface showing some analytics divided in two tabs: a regions tab providing information about drawn regions (x, y, width, height) on top of canvas - and a metrics tab providing aesthetics metrics in the form of a report (Fig. 12).



Fig. 11: Drawable canvas with regions drawn on top of RCIS webpage



Fig. 12: Analytics panel showing metrics of balance and density

To put it in a nutshell, the tool's ultimate goal is to provide web designers with an objective feedback about their designed webpages. At the moment, QUESTIM allows the user to load a webpage (URL) or screenshot and proceed to semi-automatic computation with interest regions drawn by this same user. Then, the tool provides an immediate response in the form

of a metrics report. For each metric result, an interpretation can be defined and a link can be made with the incidence on users preferences and/or performances. Fig. 13 is a concrete illustration of the use that may be made of QUESTIM to evaluate and redesign a website (i.e. RCIS 2014 webpage).

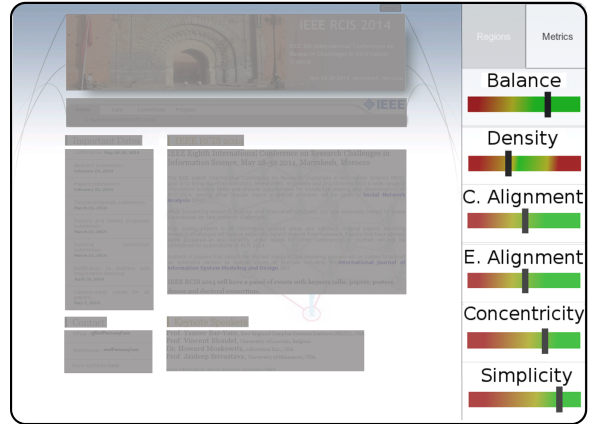
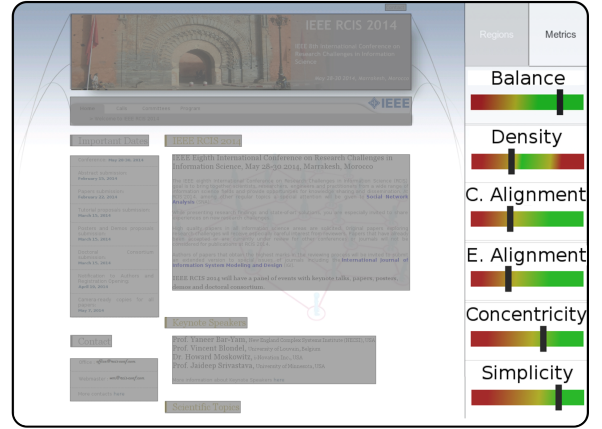


Fig. 13: Evaluation of the RCIS webpage using QUESTIM

VI. PILOT STUDY

This section presents a pilot study performed to provide empirical validation to Ngo's metrics [26]. We addressed 12 metrics - balance, sequence, equilibrium, unity, symmetry, proportion, economy, density, homogeneity, regularity, rhythm and order - for which a mathematical formula was provided and gave a quantitative value to each existing arrangement property. These formulas were based on UI objects attributes such as measures of area, size, angle, color and form.

A. Goal

The goal of this pilot study was to show relevance of the metrics while considering final users perceptions. To reach this objective, we basically calculated each of Ngo's metrics for a set of 4 interfaces and confronted the results to users reviews on the perception of the arrangement property related to that

metric. In other words, the aim was therefore to verify whether the results that can be obtained using these formulas could be legitimate as corresponding to users reviews, what the users really perceive. In order to complete this task, we applied the following methodology:

- 1) select metrics to validate
- 2) select UIs onto which could be applied the metrics
- 3) quantify for each UI the corresponding metrics
- 4) conduct a survey among users to get their own opinions about the related visual techniques for each UI
- 5) confront users reviews with metrics values and discuss the results

B. Description

The experiment took place in a controlled environment where respondents, the number of which was 25, replied each to the survey at a different time in the presence of an interviewer. The persons who took part in the experiment were under and postgraduate students of Université catholique de Louvain (UCL) in various study domain from Literature to Sciences. Each participant daily uses a computer or an electronic device with an interface. No remuneration was given for their participation. The survey form presented 4 UIs - exposed in Fig. 14 - and requested the respondents to analyze those 4 UIs in the light of 12 visual techniques - related to the selected metrics - by giving a score on a 5-points Likert scale.

C. Interfaces

As mentioned, four different interfaces were chosen based on their potential aesthetic properties (two from websites, one from a video game and the last one from an ATM). Each UI has been named as follows:

- BELGIUM - the front page of the Belgian government official website.
- ATM - the ATM interface of the bank BNP Paribas-Fortis.
- GTR2 - the main menu of the PC game GTR 2, a racing simulation video game.
- UCL - the front page of the UCL official website.

D. Procedure

Each participant was given a questionnaire to fill, a document explaining the metrics and a representation of the four interfaces printed on paper. The questionnaire contained thirteen sections, one for each metric, and four subsections, one for each interface. For each metric and for each interface, the participants had to give a score on a 5-points Likert scale. A value of 1 meaning that the score given for the considered aesthetic property was low up to a value of 5 considered as a high score for the aesthetic property.

Interviewers provided an explanation for each visual technique as well as two illustrations, each of an interface with high and low values for the metric. The UIs were printed in color on a single recto-verso sheet of A4 paper. Two UIs per page, each with the same size.



Fig. 14: Selected UIs for an exploratory experiment

Each UI was transformed through a grid definition process (Fig. 15), the only purpose of which was to distinguish the interface objects composing each UI in order to apply the metric analysis.

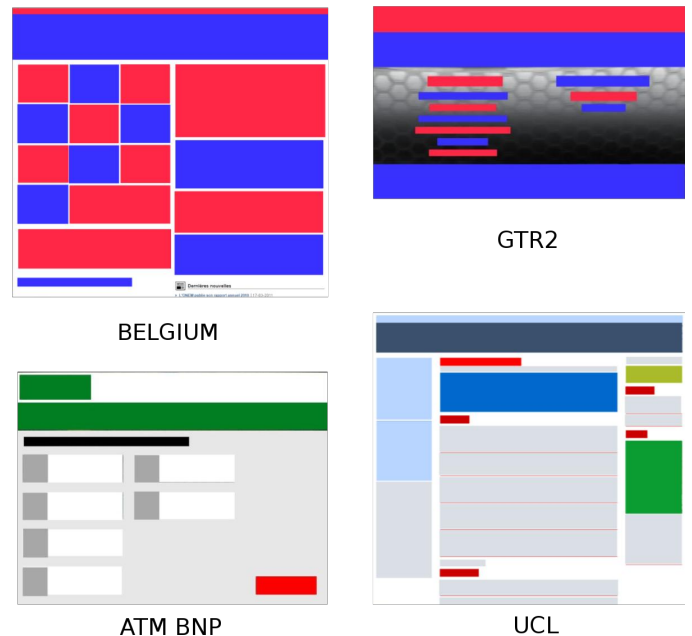


Fig. 15: Grid transformation of each selected UI

E. Hypotheses

As the goal of the experiment was to show representativeness of the predefined metrics, we chose to consider the following hypotheses:

- H1: There is a *perfect* similarity between scores of metrics and users reviews.
- H2: There is a *relative* similarity between scores of metrics and users reviews.

F. Statistical tests

In order to define the necessary tests, we conducted initially some tests to assess the variables distribution normality, namely the Kolmogorov-Smirnov test and the Shapiro-Wilk test. Both were negatively concluded (Sig. < 0.05) as shown in Table I meaning our datas were not following a normal probability law.

TABLE I: Normality tests for assessing type of variables distribution

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BalanceATM	0.243	25	0.001	0.779	25	0.000
BalanceBELGIUM	0.205	25	0.008	0.894	25	0.014
BalanceGTR2	0.208	25	0.007	0.909	25	0.029
BalanceUCL	0.244	25	0.000	0.895	25	0.014

For this reason, we chose to conduct exclusively non-parametric tests in order to investigate the aforementioned hypotheses. One-sample Wilcoxon signed-rank test, Match-paired Wilcoxon signed-rank test and Friedman's 2-way ANOVA by ranks were performed. Results of those 3 tests are respectively presented in Tables II, III and Fig. 16 only for the balance metric for the sake of conciseness and reading simplicity. The full experiment results are available at the following URL: <http://sites.uclouvain.be/questionnaire/pilotstudy/>

TABLE II: One-sample test to assess exact matching between users scores and metric considering balance

Variable	Median	Metric	H0	Sig.
BalanceATM	2	3.24	=	0.000
BalanceBELGIUM	3	3.44	=	0.636
BalanceGTR2	3	3.96	=	0.004
BalanceUCL	4	4.08	=	0.008

In the unpaired One-sample Wilcoxon test, each metric for each UI is compared to the median of users scores. If the Significance is inferior to 5%, the meaning is that the null hypothesis is rejected and we cannot consider that values are similar. For instance, in Table II, only the users perceptions of the balance for interface BELGIUM are validating the metric. We proceeded similarly for analyzing the remaining metrics.

In the Matched-paired Wilcoxon test, we compare each variable by pairs and test the null hypothesis according to which the difference of the medians is equal to 0. By analyzing the results provided in Table III, we note that balance property of interface ATM gets the lowest scores (negative statistic and Sig. < 0.05) in comparison with the 3 other interfaces.

The Friedman ANOVA test could have been discarded because it does not cover directly our hypotheses. However,

TABLE III: Matched-paired test to establish interface ranking in the light of the balance metric

Null Hypothesis	Statistic	Sig.
BalanceATM - BalanceBELGIUM = 0	-3.124	0.002
BalanceATM - BalanceGTR2 = 0	-3.324	0.001
BalanceATM - BalanceUCL = 0	-3.725	0.000
BalanceBELGIUM - BalanceGTR2 = 0	0.560	0.576
BalanceBELGIUM - BalanceUCL = 0	-2.034	0.042
BalanceGTR2 - BalanceUCL = 0	-2.083	0.037

Related-Samples Friedman's Two-Way Analysis of Variance by Ranks

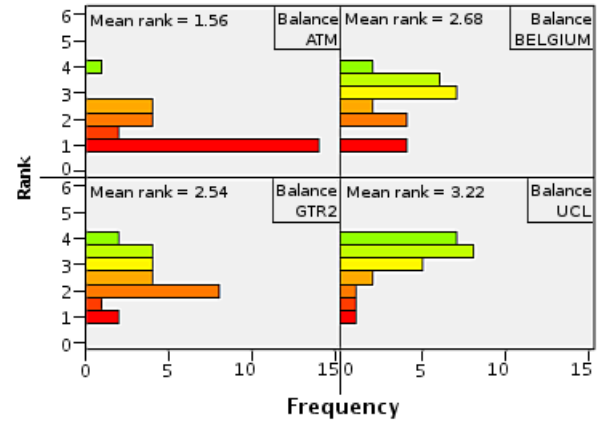


Fig. 16: Friedman ANOVA by ranks graph comparing balance perceptions of each interface

it is another evidence showing that the users scores for each metric are different when they are compared together. We chose not to add a table just for showing significance of this test which is equal to 0.000 which induces a rejection of the null hypothesis according to which all medians are equal (< 0.05). However, we found the resulting graph comparing each rank relevant for showing dispersion of responses (Fig. 16) and for supporting the ranking established with the Wilcoxon test.

G. Results

At first sight, the results provided by the unpaired Wilcoxon tests were not advocating aesthetics metrics with 17 matching out of 52 items - an item represents the value and the score given by user for one metric for one specific UI (i.e. 13 metrics * 4 UIs = 52 items) - or in other words with only 1 metric out of 12 (proportion) significantly represented users reviews for each UI - which is not a surprise per se since it is not really possible to obtain exactitude when comparing formula's values and Likert scale values. However, the results of the paired Wilcoxon tests were much more interesting because they were providing an idea of how the UI were ranked for a specific metric. Indeed, in Table III we compare each median of the reviews score for one UI with one other UI. If the p-value is inferior to 10%, the null hypothesis according to which the score are equal is rejected. It is even possible to

determine the superior UI taking into account the positive or negative sign of the statistic outcome. That leads to a possible comparison between UI rankings (as exposed in Fig. IV only for the metric of balance) for each metric by formulas on one hand and by users on the other hand. The result is then a number of 4 metrics out of 12 for which UIs were ranked mainly similarly by both criteria showing that metrics formulas were representative of human eye. Those metrics were balance, equilibrium, density, and economy.

TABLE IV: Ranking of UIs exclusively considering balance according to the users and the metric

Visual technique	UI ranking according to	
	METRIC	USERS
Balance	UCL	UCL
	GTR2	GTR2/BELGIUM
	BELGIUM	
	ATM	ATM

H. Discussion

First of all, the fact that results indicate 4 representative metrics out of 12 does not imply that the other metrics have to be rejected when considering potential elements for aesthetics evaluation. Indeed, it just points out that the results were not significant for the remaining 8 metrics. Another possible interpretation is relative to the metric itself that may be reconsidered in order to obtain relevant results.

Regarding the validation of the whole experiment, a critic has to be made. Indeed, we used a set of only 4 UIs and that implies a lack of significance in the results. Therefore, we cannot assess a sufficient margin error in the constitution of previous rankings to be assured that they are representative. The results obtained give thus an indication of a possible GUI evaluation by metrics but need to be treated with caution and request further empirical support.

Finally, another critic that can be made is about the subjective nature of the UI grid transformation process, that does not follow specific rules and is let to the judgement of the researchers. Indeed, another grid proposition for a UI could change considerably results obtained by the metrics. It should be appropriate for a metrics-based evaluation to rely on a consistent process for defining UI regions of interests.

VII. FINAL REMARKS

Qualitative evaluation of a design provided by several users judgements remains an appropriate solution to improve UI design. However, this kind of evaluation process is cumbersome and can often only be undertaken when the UI is already released. A better approach for first design evaluation could be to objectively quantify aesthetics through an evaluation by metrics. The interest of such an approach lies in the quick and objective feedback the designer may receive. Indeed, rather than having to gather people and optionally conduct a survey to receive comments on some way to improve the UI aesthetics, he/she may simply define UI regions and request a

metrics report. Afterwards, he/she has the necessary elements to modify the GUI in order to improve its aesthetics.

Through this research work, we hope to provide interface designers with methodologies and tools using validation by metrics that can be called upon at any time during design process. That is why we have chosen to focus our efforts on the definition of a method (INFRARED) and a tool (QUESTIM) in order to give concrete examples of what can be achieved in the design process when a particular attention is accorded to metrics for GUI aesthetics evaluation.

Therefore, improvements can be made in the way of addressing the model. At the moment, it only takes into account UI objects arrangement on a layout. It is necessary, as previously mentioned by other works [36, 37], to propose a model considering other potential UI aesthetics aspects such as colors combinations and shapes complexity. Another suggestion is to take into consideration the fact that UI aesthetics vary with the domain in which it belongs. Indeed, it may be nonsense for some designers to achieve a UI in which objects are completely balanced. Therefore, an idea is to split each metric in such a way that it represents no more one positive and one negative value (Balance/Unbalance) but rather two distinct components. With this point of view, it is possible to analyze joint effects on UI aesthetics (e.g. effect of an unbalanced and misaligned UI).

Finally, the last but not the least point to consider is the necessity to provide more robust and significant empirical validation of how metrics express aesthetics aspects through several experiments focusing on the relevance - the extent to which a metric is an important aesthetics criterion - and the representativeness - the extent to which the value provided by a metric is representative of what a human being may assess. We made a small step in supporting metrics representativeness through a pilot study, the results of which are encouraging since we were able to highlight some correlation between a number of metrics and the users perception of the related visual technique. However, the experiment keeps an exploratory property and triggers the need for advanced empirical support.

VIII. FUTURE WORK

This section describes some lines of research to undertake or to further investigate in the context of GUIs aesthetics evaluation:

- Regarding the evaluation by metrics, we made a small step in providing a tool for drawing objects of an interface and get a quantitative result. It would be interesting to automate most of this evaluation process and propose a system of recognition of shapes and colors. The interest would be, on the one hand, to avoid manual drawing of objects and, on the other hand, the evaluation could be applied to any interface regardless of the language which is used for defining the interface.
- Some metrics may be optimized to better reflect perceptions of human eye. For example, it is necessary to provide an accurate and validated threshold for categorizing correctly values of density that are considered by users as efficient or unefficient from an

aesthetics perspective. Moreover, the thresholds used in the analytics panel of QUESTIM were arbitrary defined. It is therefore necessary in a research process to provide empirical validation for these thresholds in order to propose a reliable evaluation tool.

- In the same vein, it is necessary to provide further empirical validation for the previously defined metrics with a final goal of solving the representativeness and relevance issues raised by such a quantitative approach of aesthetics evaluation. An attempt to solve the representativeness issue could be to reproduce the pilot study at a larger scale (i.e. exploring a larger corpus of interfaces with a larger sample of users) in order to support formulas represented for each metric. In another experiment, it could be possible to tackle the relevance issue by investigating professional designers methods and principles when it comes to GUI aesthetics. It could be in the form of a survey on the visual techniques they actually use in their work. Confronting the results of such a survey to a set of metrics would allow to make a distinction between those that are relevant for aesthetics validation and those that are irrelevant.

ACKNOWLEDGMENT

This work is partially supported by the QualiHM project (Tool-Supported Methodology for the Quality of Human-Computer Interfaces) of the CQuality research programme by Direction Générale DGO6 of Région Wallonne. A special thank goes to our colleague Ugo Sangiorgi for comments on earlier version of this paper.

REFERENCES

- [1] Cisco, "Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update , 2012 – 2017," pp. 2012–2017, 2013.
- [2] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!" *Behav. Inf. Technol.*, vol. 25, no. 2, pp. 115–126, Mar. 2006.
- [3] E. Papachristos and N. Avouris, "Are first impressions about websites only related to visual appeal?" in *Human-Computer Interact.* Lisbon: Springer, 2011, pp. 489–496.
- [4] P. Zhang and N. Li, "Love at first sight or sustained effect? The role of perceived affective quality on users' cognitive reactions to IT," in *Proc. Int. Conf. Inf. Syst.*, Washington, DC., 2004, pp. 283–296.
- [5] B. A. Myers and M. B. Rosson, "Survey on user interface programming," in *Proc. SIGCHI Conf. Hum. factors Comput. Syst. - CHI '92*, vol. 7, no. 1. New York, New York, USA: ACM Press, Mar. 1992, pp. 195–202.
- [6] M. Zen, "Metric-based evaluation of graphical user interfaces: model, method, and software support," *Proc. 5th ACM SIGCHI Symp. Eng. Interact. Comput. Syst.*, 2013.
- [7] M. J. Gregor, "Baumgarten's Aesthetica," *Rev. Metaphys.*, vol. 37, p. 357, 1983.
- [8] F. Clay, "The origin of the aesthetic emotion," *Sammelbande der Int. Musik.*, pp. 282–290, 1908.
- [9] W. Van Damme, *Beauty in Context: Towards an Anthropological Approach to Aesthetics*, E.J. Brill, Ed., 1996.
- [10] T. Lavie and N. Tractinsky, "Assessing dimensions of perceived visual aesthetics of web sites," *Int. J. Hum. Comput. Stud.*, pp. 1–41, 2004.
- [11] A. N. Tuch, J. a. Bargas-Avila, and K. Opwis, "Symmetry and aesthetics in website design: It's a man's business," *Comput. Human Behav.*, vol. 26, no. 6, pp. 1831–1837, Nov. 2010.
- [12] A. Sonderegger and J. Sauer, "The influence of design aesthetics in usability testing: effects on user performance and perceived usability," *Appl. Ergon.*, vol. 41, no. 3, pp. 403–10, May 2010.
- [13] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm, "Visual complexity of websites: Effects on users' experience, physiology, performance, and memory," *Int. J. Hum. Comput. Stud.*, vol. 67, no. 9, pp. 703–715, Sep. 2009.
- [14] A. Altoboli and Y. Lin, "Effects of Unity of Form and Symmetry on Visual Aesthetics of Website Interface Design," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 56, no. 1, pp. 728–732, Oct. 2012.
- [15] F. Alsudani and M. Casey, "The effect of aesthetics on web credibility," in *Proc. 23rd Br. HCI Gr. Annu. Conf. People Comput. Celebr. People Technol.* British Computer Society, 2009, pp. 512–519.
- [16] C. Salimun, H. C. Purchase, D. R. Simmons, and S. Brewster, "The effect of aesthetically pleasing composition on visual search performance," *Proc. 6th Nord. Conf. Human-Computer Interact. Extending Boundaries - Nord. '10*, p. 422, 2010.
- [17] G. Moss, R. Gunn, and J. Heller, "Some men like it black, some women like it pink: consumer implications of differences in male and female website design," *J. Consum. Behav.*, vol. 341, no. 2006, pp. 328–341, 2006.
- [18] N. Bevan, "Classifying and selecting UX and usability measures," in *Proc. Meaningful Meas. Valid Useful User Exp. Meas.*, 5th COST294-MAUSE Open Workshop, Ed., vol. 11, no. June, Reykjavik, 2008.
- [19] W. O. Galitz, *The Essential Guide to User Interface Design : An Introduction to GUI Design Principles and Techniques*. John Wiley & Sons, 2002.
- [20] W. Horton, *Illustrating Computer Documentation : The art of presenting information*. John Wiley & Sons, 1991.
- [21] K. Mullet and D. Sano, *Designing visual interfaces: Communication oriented techniques*. Englewood Cliffs NJ: SunSoft Press, Apr. 1995, vol. 28, no. 2.
- [22] F. Montero, V. Manuel, and L. Jaquero, "GUILayout ++ : Supporting Prototype Creation and Quality Evaluation for Abstract User Interface Generation," in *Proc. 1 st Work. User Interface eXtensible Markup Lang. UsiXML'2010*, vol. 2010, no. June. Berlin: Thalès, 2010, pp. 39–44.
- [23] R. Mahajan and B. Shneiderman, "Visual and textual consistency checking tools for graphical user interfaces," *IEEE Trans. Softw. Eng.*, vol. 23, no. 11, pp. 722–735, 1997.
- [24] T. Tullis, "A computer-based tool for evaluating alphanumeric displays," in *INTERACT '87, 2nd IFIP Conf. Human-Computer Interact.*, 1987, pp. 123–127.
- [25] J. Vanderdonckt and X. Gillo, "Visual techniques for traditional and multimedia layouts," in *Proc. Work. Adv. Vis. interfaces.* ACM, 1994, pp. 95–104.
- [26] D. Ngo, A. Samsudin, and R. Abdullah, "Aesthetic measures for assessing graphic screens," *J. Inf. Sci. Eng.*, vol. 16, pp. 97–116, 2000.
- [27] S. González, F. Montero, and P. González, "BaLOReS," in *Proc. 13th Int. Conf. Interacción Pers. - INTERACCION '12*. New York, New York, USA: ACM Press, 2012, pp. 1–2.
- [28] R. Arnheim, *Art and Visual Perception: a Psychology of the Creative Eye*. University of California Press, 1954.
- [29] D. A. Dondis, *A Primer for Visual Literacy*. The MIT Press, Jun. 1973, vol. 3, no. 2.
- [30] S. Lok, S. Feiner, and G. Ngai, "Evaluation of visual balance for automated layout," in *Proc. 9th Int. Conf. Intell. user interface - IUI '04*. New York, New York, USA: ACM, 2004, p. 101.
- [31] G. Markowsky, "Misconceptions about the golden ratio," *Coll. Math. J.*, vol. 23, no. 1, pp. 2–19, 1992.
- [32] F. Boselie, "The aesthetic attractiveness of the golden section," *Psychol. Res.*, no. 45, pp. 367–375, 1984.
- [33] G. T. Fechner, "Various Attempts to Establish a Basic Form of Beauty: Experimental Aesthetics, Golden Section, and Square," pp. 115–130, 1997.
- [34] F. Phillips, J. F. Norman, and A. M. Beers, "Fechner's aesthetics revisited," *Seeing Perceiving*, vol. 23, no. 3, pp. 263–71, Jan. 2010.

- [35] L. Brady and C. Phillips, "Aesthetics and usability: A look at color and balance," *Usability News*, vol. 5, no. 1, pp. 2–5, 2003.
- [36] D. Ngo and J. Byrne, "Another look at a model for evaluating interface aesthetics," *Int. J. Appl. Math. Comput. Sci.*, vol. 11, no. 2, pp. 515–535, 2001.
- [37] J. M. Zain, M. Tey, and Y. Goh, "Probing a Self-Developed Aesthetics Measurement Application (SDA) in Measuring Aesthetics of Mandarin Learning Web Page Interfaces," vol. 8, no. 1, 2011.
- [38] H. Purchase and J. Hamer, "Investigating objective measures of web page aesthetics and usability," in *Proc. Twelfth Australas. User Interface Conf.*, vol. 117. Australian Computer Society, 2011, pp. 19–28.
- [39] M. Möttus, D. Lamas, M. Pajusalu, and R. Torres, "The evaluation of interface aesthetics," *Proc. Int. Conf. Multimedia, Interact. Des. Innov. - MIDI '13*, p. 1, 2013.
- [40] A. Sears, "AIDE: A tool to assist in the design and evaluation of user interfaces," *Interact. Syst. Res. Center, Balt.*, pp. 1–19, 1993.
- [41] W. C. Kim and J. D. Foley, "DON : User Interface Presentation Design Assistant," vol. 20052, no. Figure 1, 1990.
- [42] D. Streveler and A. Wasserman, "Quantitative measures of the spatial properties of screen designs," in *Proc. IFIP TC13 First Int. Conf. Human-Computer Interact.*, 1984, pp. 81–89.
- [43] A. Altaboli and Y. Lin, "Investigating Effects of Screen Layout Elements on Interface and Screen Design Aesthetics," *Adv. Human-Computer Interact.*, vol. 2011, pp. 1–10, 2011.
- [44] I. Alsmadi and M. Al-Kabi, "The Introduction of Several User Interface Structural Metrics to Make Test Automation More Effective," *Open Softw. Eng. J.*, vol. 3, no. 1, pp. 72–77, Jul. 2009.
- [45] M. Ivory, R. Sinha, and M. Hearst, "Empirically validated web page design metrics," in *Proc. SIGCHI Conf. Hum. factors Comput. Syst. - CHI '92*. ACM, 2001, pp. 53–60.
- [46] A. Seffah, M. Donyae, R. B. Kline, and H. K. Padda, "Usability measurement and metrics: A consolidated model," *Softw. Qual. J.*, vol. 14, no. 2, pp. 159–178, Jun. 2006.
- [47] M. Nebeling, F. Matulic, and M. C. Norrie, "Metrics for the evaluation of news site content layout in large-screen contexts," *Proc. 2011 Annu. Conf. Hum. factors Comput. Syst. - CHI '11*, p. 1511, 2011.
- [48] T. Comber and J. Maltby, "Layout complexity: does it measure usability?" in *Human-Computer Interact. INTERACT'97*. Springer US, 1997, pp. 623–626.
- [49] J. Noble and L. Constantine, "Interactive design metric visualization: visual metric support for user interface design," *Proc. Sixth Aust. Conf. Comput. Interact.*, pp. 213–220, 1996.
- [50] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos, "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '13*, p. 2049, 2013.
- [51] A. Parush, R. Nadir, and A. Shtub, "Evaluating the layout of graphical user interface screens: Validation of a numerical computerized model," *Int. J. Hum. Comput. Interact.*, vol. 10, no. 4, pp. 343–360, 1998.
- [52] R. Kothari, J. Basak, and I. Block, "Perceptually motivated measures for capturing proximity of web page elements: Towards automated evaluation of web page layouts," *11th Int. World Wide Web Conf.*, pp. 1–14, 2002.
- [53] C. Carswell, "Graphical information processing: The effects of proximity compatibility," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 34, no. 19, pp. 1494–1498, 1990.
- [54] Q. Limbourg and J. Vanderdonckt, "USIXML: A language supporting multi-path development of user interfaces," *Eng. Hum. Comput. Interact. Interact. Syst.*, pp. 200–220, 2005.