High level data fusion on a multimodal interactive applications platform

Olga Vybornova, Hildeberto Mendonça, Jean-Yves Lionel Lawson, Benoit Macq
Communications and Remote Sensing Lab (TELE),
Université catholique de Louvain (UCL), Belgium
{olga.vybornova, hildeberto.mendonca, jean-yves.lawson, benoit.macq}@uclouvain.be

Abstract

We demonstrate a multimodal high-level data fusion tooling integrated on a platform aimed at the rapid development and prototyping of multimodal applications through user-centered design. The platform embeds a set of pure and combined modalities as reusable components and generic mechanisms for combining modalities with a rich support for multimodal fusion in order to improve the human-computer interaction.

1. Motivation overview

Multimodal fusion is a central question to solve and provide effective and advanced human-computer interaction. To understand and formalize the coordination and cooperation between modalities involved in the same multimodal interface it is required to extract relevant features from signal representations and thereafter to proceed to high-level (semantic-level) fusion. Signal- and semantic-level processing is deemed tightly interrelated since signals are seen as bearers of the meaning. To provide natural and robust interaction with the user(s) the system must be able to interpret as a whole the various modalities used by a human to interact, i.e., to extract information arriving simultaneously from different sources and to combine them into one or several unified and coherent representations of the user's behavior. A challenge is to fuse these different input modalities effectively to augment their natural strengths and use redundancy to increase robustness. Moreover, it is critical to decide what information should be fused and what should not.

We see multimodality as a set of variables describing states of the world (user's input, an object, an event, behavior, etc.) represented in different media and through different information channels. In this respect the goal of data fusion (merging semantic content from multiple streams) should give an efficient **joint interpretation** of the multimodal behavior of the user(s) – to provide effective and advanced interaction.

We have been working with multimodal fusion considering two main modalities, which are speech recognition and human behavior analysis through image processing tools. We have conducted experiments in a hypothetical context, with two or more people interacting with each other and with objects in the scene. We observed that when intending to perform an action, the user(s) might:

- speak about it describing their actions or intentions (the ideal case, then the linguistic and action recognition data are complementary);
- be doing things, but speaking about something absolutely irrelevant to those actions (in this case the data from the two modalities should be analysed separately, without merging);
- the users' words might contradict the actions performed (then the actions have priority, since "actions speak louder than words");
- be just silent when performing actions (in this case the unimodal operation mode is required).

2. Implementation

Our experiments are supported by a set of tools integrated and managed by the OpenInterface (OI) platform (www.openinterface.org). OI is a platform to manage multimodal components and integrate them with applications which demands multimodal interactions. The main differential of OI is that it minimizes the need for complex changes in the application, injecting a part of multimodal functionalities and sometimes requiring some additional customization according to the user needs. Therefore, the code to process signals is kept outside of the application, simplifying the maintenance, allowing modality extensibility and on-line configuration which are necessary when external factors from the context influence the interaction. In order to perform data fusion we extended the OI functionalities to support parallelism between modalities, which is obligatory to allow fusion, since we have to consider many sources of data at the same time. We also implemented a

978-0-7695-3454-1/08 $25.00 © 2008 IEEE
DOI 10.1109/ISM.2008.21

timing mechanism to register all detected signals as modality events. It helps us observe all events in a timeline, since the sequence of events is important to analyze certain kinds of behaviors and validate the consistency of speech.

The existent OI functionalities plus our contributions allows adding new modalities on demand, since it offers a component composition support, where components developed for different purposes in different platforms can cooperate synchronously or asynchronously, with a declarative communication mechanism, to solve larger fusion challenges.

We use a multi-agent system through a middle-ware that complies with the FIPA specifications and through a set of graphical tools that supports the debugging and deployment phases. The agent platform is integrated with OI through a plugin to interchange data. The configuration can be even changed at run-time by moving agents from one machine to another one when required.

3. Application description

We observe two people in a room. The system tracks them, analyzes their behavior, movements, speech, and takes decisions about how to prompt them necessary information when required or provide any other assistance. The main challenges that we face in this application are unrestricted human language and free natural behavior within home/office environment. In order to analyze behavior efficiently, the system has to correctly process, interpret and create joint meaning of the data coming from speech analysis and video scene analysis. We consider that human behavior is goal-oriented, so our main aim is to recognize the users' plan, to see what they want to achieve.

Everything that humans say or do makes sense only in a particular context. In some situations we can see clearly what data should be integrated, when human actions correspond to their words and vice versa, e.g. a person is moving to telephone and says that he is going to make a phone call. This is the ideal case, and then the linguistic and action recognition data are complementary. However our system is able to handle various cases and more complex ones too.

Generally speaking, we perform the high level fusion of the input streams in three stages: (i) early fusion that merges information already at the signal or recognition stage to provide reliable reference resolution, (ii) reinforced fusion that is a cycling process performing crossmodal verification, intra- and inter-modality processing enhancement aimed at improving the final interpretation, and (iii) late fusion that integrates all the information at the final stage to give interpretation of the user's behavior.

We will demonstrate how the system manages the data streams arriving from two sources – video scene and speech. In particular, we show a technique distinguishing between the data from different modalities that should be fused and the data that should not be fused but analyzed separately.

The application employs various components integrated on the OpenInterface platform: speech recognition, person identification, syntactic parsing, natural language semantic analysis, video analysis, human behavior analysis component, a cognitive architecture that serves as the context-aware controller of all the processes, manages the domain ontology and provides the decision-making mechanism.

4. Acknowledgement

The work on the OpenInterface platform described herein was funded by the European FP6 SIMILAR Network of Excellence (www.similar.cc) and continues by the European FP6-35182 OI STREP project (www.oi-project.org).

5. References

[1] eNTERFACE Workshops, http://www.enterface.net/

[2] Lawson L. and Macq B. OpenInterface platform for multimodal applications prototyping. – ICASSP Show&Tell 2008, March 30-April 4, Las Vegas, Nevada, USA.

[3] Vybornova O., Gemo M., Moncarey R. and Macq B.: Ontology-Based Multimodal High Level Fusion Involving Natural Language Analysis for Aged People Home Care Application // In: Proceedings of INTERSPEECH 2007, August 27-31, Antwerpen, Belgium.