Multimodal Human-Computer Interfaces

Editors
Thierry Dutoit - Faculté Polytechnique de Mons, Belgium
Laurence Nigay, Université Joseph Fourier – Grenoble 1, France
Michael Schnaider, Zentrum für Graphische Datenverarbeitung, Germany

Editorial

The goal of multimodal interfaces is to extend the sensory-motor capabilities of computer systems to better match the natural communication means of human beings. Multimodal interfaces represent a very active interdisciplinary research area which has expanded rapidly. Since the seminal "Put that there" demonstrator by R. Bolt (1980) that combines speech and gesture, significant achievements have been made in terms of both modalities and real multimodal systems. Indeed, in addition to more and more robust modalities, conceptual and empirical work on the usage of multiple modalities is now available for guiding the design of efficient and usable multimodal interfaces.

Closely linked with the actions of the Network of Excellence SIMILAR on Multimodal Interfaces (www.similar.cc), the goal of this special issue is to highlight the interdisciplinary work on multimodal interfaces that includes several research domains. The selected papers cover at least three research domains: Human-Computer Interaction, Speech Processing and Image Processing.

There are of course many ways of organizing a collection of papers on multimodal interfaces; we have chosen to organize them into two sets: the papers mainly focusing on input multimodal interfaces (from the user to the system) and the ones dedicated to output multimodal interfaces (from the system to the user).

In this issue, several papers focus on ***input multimodal*** *interfaces. The first three papers of this part involve speaker/speech recognition. The three following papers then focus on the interpretation of the combined usage of speech and gesture while we conclude this part on input multimodality with error handling in the course of multimodal communication.*

In their paper on audio-visual perception of a lecturer in a smart seminar room, Stiefelhagen *et al*. aim at providing valuable cues for annotating and indexing multimedia recordings of seminars. They describe the various combined components (tracking, speech recognition, head orientation, identification) of a smart teaching system. One important conclusion they draw, for instance, based on results on several multimodal recordings of seminars, is that there is a direct relation between the accuracy of speaker localization and the word error rate obtained from speech recognition.

Given a particular sound event, we are all perfectly capable of locating its origin if it has been generated by some visible mechanical action. In other words, as Monaci *et al.* clearly note in their paper, "we are particularly efficient at assessing audiovisual

synchrony". They investigate video modelling (in terms of geometric primitives evolving over time) and deliver a speaker mouth tracking system which can be compared with the evolution of an audio feature to assess the correspondence between acoustic and visual signals.

The paper by Cetingul *et al.* also investigates the bimodal nature of speech input, by proposing a multimodal speaker/speech recognition system that integrates audio, lip texture and lip motion modalities. They show that inclusion of the lip motion modality provides further performance gains over those which are obtained by fusion of audio and lip texture alone, in both speaker identification and isolated word recognition scenarios.

The paper by Carbini *et al.* reports on a Wizard of Oz experiment for analyzing multimodal user behaviour of their MOWGLI (Multimodal Oral With Gesture Large display Interface) real time system. The information context is that of a cooperative story telling experiment, where user speech-gesture actions are interpreted in order to cooperatively build a story with another person, partner of the interpreter.

Building on his recent book on multimodal man-machine dialogue, Landragin focuses on the interpretation of verbal referring expressions together with pointing gestures. The paper proposes to integrate perceptual, linguistic and cognitive aspects of communication within a "multimodal reference domain", for the correct interpretation of references to objects (known to be very ambiguous).

Martin *et al.* propose an in-depth discussion of how children gesture and combine their gesture with speech when conversing with a 3D character. They study the errors occurring in the context of a conversational system: the NICE project, which gives rise to two prototypes, one for conversation with fairytale author Hans Christian Andersen and another one for playful computer game-style interaction with some of his fairytale characters in a fairytale world.

The paper by Bourguet focuses on error handling. Indeed any communication is error-prone by nature. Understanding where and why errors occur is a necessary step towards increasing the efficiency of multimodal interfaces. Being able to handle errors correctly in the course of communication is also a critical component of multimodal interfaces. In her paper on the taxonomy of error-handling strategies in recognition-based multimodal HCI, Bourguet browses spoken human-machine dialogues, handwriting systems as well as multimodal natural interfaces, and derive a classification scheme which can serve as a design tool for the development of more efficient and natural multimodal human-machine interfaces.

Multimodal output *interfaces are specifically studied in four papers. The two first papers are dedicated to audiovisual speech (combining sound and lip movements), after the visionary work of Christian Benoît and his colleagues, who first coined the term "visemes" at the First ESCA Workshop on Speech Synthesis in 1991 . The third paper focuses on the sense of touch and the last one depicts a more general approach for multimodal output interfaces.*

In their paper on lip synchronization, Zorić and Pandžić focus on the animation of the face of a speaking avatar in such a way that it realistically pronounces a given speech, not knowing the underlying text in advance. For a realistic result, lip movements must be perfectly synchronized with the audio (albeit some delay cannot be avoided before speech analysis results can be used for face synthesis).

Zelezny *et al.* detail the design, implementation and evaluation of an audio-visual speech synthesis system in Czech, with its acoustic synthesis emulating human speech and its 3D facial animation emulating human lip articulation.

Mode (or sensory) substitution is of high interest for the development of aids for people with sensory impairments. In this issue, Wall and Brewster provide a review of the developments in tactile displays for such sensory substitution, along with relevant literature on perceptual and psychophysical issues related to the sense of touch.

Last but not least, Rousseau *et al.* focus on the design of output multimodal systems and study the intelligent multimodal presentation of information. They propose a model (WWHT- What, Which, How and Then) and a platform (ELOQUENCE) for the specification, the simulation and the execution of output multimodal systems.

Most authors (and the editors of this issue!) agree on one point: the current lack of multimodal corpora suitable for the evaluation of recognition/synthesis approaches and interaction strategies. In spite of the efforts of data distributors such as ELDA and LDC, one must admit that most corpora available today target the study of a limited number of modalities, if not one. Let us hope that such open challenges will be further developed in the multimodal community, as they appeared in speech or image processing a few years ago and will contribute to the worldwide sharing of resources.

* * *

We thank the many reviewers who spent time reviewing manuscripts for this special issue on multimodal interfaces. We are indeed grateful for all of the reviewers' hard work.

Aderito Marcos	Centro de Computação Gráfica, Guimarães, Portugal
De Amicis Raffaele	Center for Advanced Computer Graphics Technologies, Trento, Italy
Arisoy Ebru	Bogazici University, Turkey
Bellik Yacine	Université de Paris 11, France
Bernsen Niels	University of Southern Denmark, Denmark
Boite Jean-Marc	Multitel ASBL, Belgium
Bourguet Marie-Luce	Queen Mary University of London, UK
Bruns Willi	Universität Bremen, Germany
Calvary Gaëlle	Université Joseph Fourier – Grenoble 1, France
Caplier Alice	Institut National Polytechnique de Grenoble, France
Carbini Sébastien	France Telecom R&D, Lannion, France
Carbonell Noëlle	Université Henri Poincaré Nancy 1, France
Cernak Milos	Insitut Eurecom, Sophia Antipolis, France
Chen Jingdong	Bell Labs, Lucent Technology, USA
Coutaz Joëlle	Université Joseph Fourier – Grenoble 1, France
Deketelaere Stéphane	Multitel ASBL, Belgium
Dupont Stéphane	Multitel ASBL, Belgium

Encarnacao Miguel	University of Tübingen, Germany
Karpov Alexey	SPIIRAS, St. Petersburg, Russia
Klein Konrad	Fraunhofer IGD, Darmstadt, Germany
Lecolinet Eric	ENST Paris, France
Malerczyk Cornelius	ZGDV eV, Darmstadt, Germany
Martin Jean-Claude	Université de Paris 8, France
Nedel Luciana	University of Rio Grande do Sul, Brazil
Nicolaidis Nikos	Aristotle University of Thessaloniki, Greece
Paterno Fabio	CNR-ISTI, Italy
Pietquin Olivier	Institut Supérieur d'Electricité, Metz, France
Potamianos Gerasimos	Thomas J. Watson Research Center, IBM, USA
Ris Christophe	Multitel ASBL, Belgium
Ronzhin Andrey	SPIIRAS, St. Petersburg, Russia
Siivola Vesa	Helsinki University of Technology, Finland
Vanderdonckt Jean	Université Catholique de Louvain, Belgium

Table of contents