J Multimodal User Interfaces (2013) 7:1–3
DOI 10.1007/s12193-012-0115-7

PREFACE

Preface

Patrizia Paggio · Dirk Heylen · Michael Kipp · Guest Editors for the Special Issue Multimodal Corpora

Published online: 10 November 2012

This special issue of the Journal on Multimodal User Interfaces collects a variety of papers dealing with multimodal corpora. Developing a multimodal corpus involves the recording, annotation and analysis of several communication modalities such as speech, hand gesture, facial expression, body posture, etc. As many research areas are moving from focused but single modality research to fully-fledged multimodality research, multimodal corpora are becoming a core research asset and an opportunity for an interdisciplinary exchange of ideas, concepts and data. The number of publicly available multimodal corpora is constantly growing as is the interest in studying multimodal communication by machines and through machines. This is a very positive trend, since the availability of large annotated multimodal corpora in different application domains, and for different languages, is a necessary prerequisite for the development of innovative, intelligent and flexible multimodal user interfaces.

This special issue brings out, in an extended and revised form, the best papers from two workshops on multimodal corpora held in conjunction with LREC 2010 and ICMI 2011 respectively (see http://www.multimodal-corpora.org). Both workshops presented a wide selection of work on different aspects of multimodal corpora, with contributions on collection efforts, coding, validation, analysis methods, annotation tools and applications of multimodal corpora. In addition to these general topics, the 2010 workshop focused specifically on how the field can benefit from new methods for the tracking of face, gaze, hands and body and the recording of articulated full-body motion using motion capture, while at the 2011 workshop, particular emphasis was put on multimodal corpora for machine learning.

Given the large selection and high quality of the papers presented at the two workshops, by collecting the best of them in an extended form, this special issue offers a more in-depth look into the latest research in the field and provides insight in how this research can contribute to the advancement of multimodal user interfaces. It constitutes an important update to the special issue of the *Journal of Language Resources and Evaluation* on Multimodal Corpora (2007) and the state-of-the-art book *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications* (2009), both of which were also based on extended versions of papers presented at previous workshops on multimodal corpora at LREC.

The articles in this collection describe multimodal corpora collected for very different application domains, and deal with analyses of multimodal data using different methodologies. A possible way to group them is according to the genre of communication they address.

The paper by Lücking et al. describes the Bielefeld Speech and Gesture Alignment Corpus, a collection of dialogues concerned with route and landmark description, a genre that is particularly well-suited for the elicitation of iconic and deictic hand gestures. In addition to a description of the annotated corpus, the paper discusses a number of research results achieved using the data, including the use of attribute-value matrices and unification to express the combined meaning of gesture and speech, and

P. Paggio (✉)
University of Malta, Msida, Malta
e-mail: patrizia.paggio@um.edu.mt

D. Heylen
University of Twente, Enschede, The Netherlands

M. Kipp
Hochschule Augsburg, Augsburg, Germany

the function of gesture to support dialogue structure. From the point of view of applications, the authors argue that the corpus can be used to derive better models for human–robot communication.

The contributions by Oeretl et al., Paggio and Navarretta and Sanchez-Cortes et al. all deal with face-to-face conversation, in pairs or in groups, although they discuss different aspects of the interaction. The first of these papers describes the D64 corpus. Their data are quite different from the Bielefeld Corpus. The participants have been filmed nonstop for 2 days, about 4 h a day. The interaction is neither task-oriented nor controlled. The features provided are annotations of conversational involvement, speech activity, pauses and degree of change of movement. Possible applications for which such data would be useful are systems that predict the degree of involvement of people in the conversation, e.g. systems in which virtual characters need to monitor or moderate a conversation. Face-to-face conversation, albeit in a more limited and controlled context, is one of the two genres targeted in the paper by Paggio and Navarretta, a study of how head movement and facial expression features can be used for the automatic classification of multimodal feedback in a corpus of first acquaintance dialogue. The study also deals with the automatic classification of linguistic feedback expressions in multimodal map-task dialogues. Finally, Sanchez-Cortes et al. present a multimodal analysis of emergent leadership in small groups using audio-visual features. Contrary to the results in the study by Paggio and Navarretta (and in the paper by Bigi discussed below), they find that multimodal features only provide a moderate effect for the task of identifying the leader, whereas simple auditive features are much more effective.

Political debates, although a kind of face-to-face interaction, are a genre in their own right. The two papers by Bigi and Poggi et al. both deal with analyses of data from political debates, but they use very different methodologies. Bigi's contribution is a study of disruption in parliamentary debates. The author shows that machine learning methods mixing shallow and deeper multimodal features achieve good results in a classification task in which the speaker's responses to disruptions are classified according to a small number of different answer types. The article by Poggi et al., on the other hand, provides a qualitative analysis of comments in political debates. A better understanding of phenomena such as disruptions and comments, as well as annotated data modeling them, are important for the development of systems that must be able to cope with such phenomena, for example tutoring systems.

Communication between humans and robots is the topic of the paper by Xavier Alameda et al., which introduces Ravel (robots with audiovisual abilities), a dataset covering examples of human–robot interaction scenarios. The paper describes the data acquisition setup, the process of sensor calibration, the data annotation and content, and describes several ways in which the data can be used to carry out experiments in this application domain.

Another form of human–machine interaction is the one that takes place in so-called smart homes, i.e. homes equipped with ambient intelligence, particularly smart homes for health- and care-related assistance aimed at the fast ageing population of industrialised countries. This genre is dealt with in the paper by Fleury et al., which describes a multimodal corpus acquired from one such fully equipped smart home. The corpus includes examples of daily life routines, interactions with objects, speech interaction, information on the user's position etc.

An aspect of communication that is receiving growing attention is that of affective behaviour. Affective behaviour and emotional content can occur in different genres, but a significant amount of current research in this field is based on corpora of acted emotions. Three of our articles fall in this group. The paper by Dubois et al, explores the use of different interface designs to achieve optimal emotion recognition in communicative devices, while Caridakis et al. provide a discussion of design and implementation issues concerning the development of a multimodal, cross-cultural corpus of affective behaviour for three different languages. The issue of cultural specificity in connection with affective behaviour is also discussed in the paper by Cu et al., which presents the Filipino multimodal emotional database, a collection of over 10,000 short annotated video sequences of acted emotions. This paper also shows that on these data, good results can be achieved for automatic emotion recognition using a mixture of voice and face features. Finally, the contribution by Tkalcic et al. presents the results of four studies based on a corpus of video clips of subjects' affective responses to visual stimuli. The studies deal with an affective content-based recommender system, a personality-based collaborative filtering recommender system, an emotion detection algorithm and a qualitative study of the latent factors.

Finally, the last paper of this collection, by Essid et al., presents a multimodal corpus collected in a very different domain from those mentioned so far, namely the domain of body movement, specifically dance. Visual and auditive modalities in this corpus refer to the frames of dance movements and various sounds including the sound of the dancer's steps. The annotation provided consists of musical and choreography annotations, as well as performance ratings. The application domain for this kind of data is one in which students can learn dance movements in a virtual environment through the use of avatars.

Overall, we hope that this issue provides a broad overview of current trends in the development of multimodal corpora

and many concrete examples of how they can be analysed with the ultimate goal of the development of user interfaces in different domains. We would like to thank all the authors for their contributions, and especially all the reviewers who have worked with us to ensure the excellent quality of this publication.

July 2012