Computational analysis of human behavior has been receiving a lot of attention in the past few years. The main drive behind this interest is the widespread penetration of computer-based systems and Internet applications that increasingly enter the domain of social relations. This requires more responsive systems, capable of adapting to the rich behavior patterns exhibited by interacting humans. The present special issue grew out of the First International Workshop on Human Behavior Understanding (HBU ’10, held as a satellite to ICPR 2010) [1], which demonstrated that two major areas of present research focus in this domain were activity recognition and affect sensing. This special issue deals with the latter.

We received 17 submissions to this special issue, and only a few were extended papers from the original workshop. The applications tackled in this set of papers covered a broad area, dealing with human-human interactions (including interviews, meetings, social gatherings, and social games), human-virtual agent interactions (for application interfaces, as well as for tutoring and coaching scenarios), and improved multimedia applications. The affective content was analyzed by visual inspection of facial cues, evaluation of affective gestures, nonverbal speech and voice cues, timing of interactions, proximity and body language of interacting parties, and physiological signals. The four selected papers in this issue represent the spectrum of affect-based human behavior analysis very well, in the variety of their settings, as well as in the modalities they consider.

The paper by Pfister and Robinson is an extended version of the authors’ work presented at the HBU ’10 Workshop. It describes a classification scheme for real-time speech assessment, evaluated in the context of public speaking skills. In this application, nonverbal speech cues are extracted and used for assigning affective labels (absorbed, excited, interested, joyful, opposed, stressed, sure, thinking, unsure) to short speech segments, as well as for assessing the speech in terms of its perceived qualities (clear, competent, credible, dynamic, persuasive, pleasant). The authors collected a corpus of natural data from 31 people attending speech coaching sessions. The presented work is a promising demonstration that expert systems can be built to make use of real-time affective cues, which opens up new venues and application areas for this mature field. It is also very timely, considering the success of the movie The King’s Speech at the Oscars.

Automatic detection and accurate quantification of facial actions is a difficult problem that is on the agenda of face analysis researchers for quite some time. The last few years have seen progress in this problem, and as witnessed by the Facial Expression Recognition and Analysis Challenge organized at the FG ’11 conference [2], there are also collaborative benchmarking efforts. The state-of-the-art in facial expression analysis places emphasis on identifying Facial Actions (FACS), evaluation of expressions in natural settings, as opposed to posed expressions, and a more detailed analysis of the temporal evolution of expressions as opposed to analysis from static images. Zhu, De la Torre, Cohn, and Zhang describe a system that meets all of these challenges, and additionally considers training sample selection as a point of improvement. As the complexity of the classification problem grows (and the identification of muscle activities and their magnitude from videos is certainly a much more complex task in comparison to recognizing basic expressions from images), the training regime becomes more important, and the need for incorporating domain-specific knowledge into the learning system is increased. The authors propose a dynamic cascade bidirectional bootstrapping scheme to select positive and negative examples for each action class, and adapt a cascaded boosting classifier for final classification. Different feature descriptors (like SIFT, DAISY, and Gabor wavelet) are compared, and the paper reports some of the best results so far in AU detection on the RU-FACS database.

The paper by Nicolau, Gunes, and Pantic also evaluates facial expressions, but combines these with movement cues obtained from the shoulder area, as well as with audio cues, to predict emotions in the valence-arousal space. Their application setting is an artificial listener, which monitors the interacting human for affective signals to give appropriate responses in real-time. The authors use particle filters to track facial and shoulder motion, Mel-frequency Cepstrum Coefficient and prosody features to process audio information and bring everything together in an innovative multimodal fusion framework taking neither a feature-level nor a model-level fusion approach, but proposing to first learn valence and arousal predictions from individual cues...
with Bidirectional Long Short-Term Memory Neural Networks (BLSTM-NN), and taking the predicted values as input to a second level of BLSTM-NN to perform the final valence-arousal classifications. The improvements obtained in this associative scheme shows that it is useful to learn the correlations between valence and arousal, and that rough estimates of these high-level cues can serve as good intermediate representations.

While the three papers mentioned so far mainly focused on face and voice cues, the fourth paper by Glowinski, Dael, Camurri, Volpe, Mortillaro, and Scherer deals with the affective content of upper-body movements. The authors stress in this work the richness of head and hand movements for communicating emotions. These sources of expression are certainly becoming more relevant for recent human-centric computing paradigms. The presented work seeks a parsimonious representation for describing a set of 12 emotion classes (elation, amusement, pride, pleasure, relief, interest, hot anger, fear, despair, cold anger, anxiety, sadness) grouped into high/low arousal and positive/negative valence clusters. The Geneva multimodal emotion portrayals (GEMEP) corpus was used in the evaluations, and the proposed system is implemented as an extension to the EyesWeb XMI Expressive Gesture Processing Library. Among the features used in the representation, the system detects activity levels, spatial extent, symmetry and jerkiness of gestures, thereby providing the potential application interface with a useful set of tools to build on.

**Acknowledgments**

We would like to thank our authors and anonymous reviewers for their contributions, and Jonathan Gratch for making this special issue possible.

Albert Ali Salah  
Theo Gevers  
Alessandro Vinciarelli  
*Guest Editors*

**References**
