Special Issue on Multimodal Affective Interaction

COMPUTERS (and computing devices) are becoming pervasive in our daily life; it is widely believed that computers will be embedded everywhere in human environments in the future. Designing intelligent human-computer interaction (HCI) systems is one of the main challenges facing the realization of pervasive interactive computing. Given that affective state recognition is an indispensable component for human-human interaction, to allow natural humanlike interactions with computing devices, HCI designs need to possess the ability of recognizing and reacting to affective state of the user.

Affect sensing and recognition has been studied in various disciplines in the past three decades. Human affective information is conveyed via a broad range of modalities and channels, including visual (facial expression, head pose, gesture, body movement and postures, etc.), auditory (pitch, loudness, speaking rate, etc.), tactile (heart rate, skin conductivity, etc.), brain signals, and so on. In natural human-human interaction, the sensed emotional signals from multiple modalities are combined in our brains to allow affect perception in variable conditions. However, most of the existing works on automatic affect recognition focus on a single modality or cue (e.g., facial expression or speech), not taking into account the multiple sources of affective information. Each modality or cue in isolation has its inherent weakness and limitation. Multimodal affect sensing and recognition is the most promising approach for building an automated affect recognition system. This has been one of the main trends in the affective computing community.

There are many challenges in handling multimodal affective interaction. Examples include how to acquire and annotate affect data using multiple sensors or modalities, especially the spontaneous data in natural settings; how to effectively extract and select representative features from different modalities for affect recognition; how to synchronize data or features from different modalities; how to select the fusion strategy of multimodal affect data for a given application, etc. It is also necessary to investigate which modalities and cues are most suitable for the application context. To address these challenges, we have to adapt existing single-mode approaches to work together at the fusion level or to devise new techniques for multimodal affective interaction. This special issue aims at presenting and highlighting the latest research and development in these areas. The papers included in this issue cover broad topics related to multimodal affective interaction.

The special issue solicited 26 submissions on a wide range of topics related to multimodal affective interaction. The corresponding authors represented broad diversity in geographic location, including North America, Europe, and Asia. We enlisted more than 70 reviewers to provide expert opinions. Most papers were reviewed by at least two and up to four experts. The decision reached for each paper was discussed among the five guest editors, and 11 papers were eventually accepted for the special issue. The 11 papers included can be categorized as follows.

Digital Object Identifier 10.1109/TMM.2010.2052315

I. Emotional Speech Synthesis and Recognition

Emotional speech synthesis and recognition is a vital component of multimodal affective interaction. To obtain accurate labels when annotating large affective speech corpora, in "Reliable Pitch Marking of Affective Speech at Peaks or Valleys Using Restricted Dynamic Programming", Alías and Munné introduce a methodology for reliable pitch marking of affective speech, which adjusts the pitch marks at the signal peaks or valleys after applying a three-stage restricted dynamic programming algorithm. For emotion identification in speech, it is believed that the prosody carries most of the emotional information. In "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech", Luengo *et al.* investigate the characteristics and capabilities of features derived from prosody, spectral envelope, and voice quality. Their experimental results suggest that spectral envelope features outperform the prosodic ones for emotion identification in speech. Affect recognition by human is largely influenced by the context information. How to exploit contextual information in automatic affect recognition remains an open problem. In "Context Analysis in Speech Emotion Recognition", Tawari and Trivedi address this problem for speech emotion recognition. They present the collection of a new audiovisual database in an automobile setting, and introduce a set of features based on cepstrum analysis of pitch and intensity contours. They systematically analyze the effects of different contexts on two different databases.

II. Affective Video Content Analysis

With vast amounts of video data being generated, affective video content analysis has been a topic of recent interest. Most works aim to identify the emotional information in videos by extracting affective features and fusing those features in some established affective models. In "Affective Visualization and Retrieval for Music Video", Zhang *et al.* present an integrated system for personalized music video (MV) affective analysis, visualization, and retrieval. A method for visualizing the abstract affective states is introduced. Both comprehensive experiments and subjective user studies on a large MV dataset demonstrate that their personalized affective analysis is more effective than the state-of-the-art algorithms. Irie *et al.* present a method for affective movie scene classification in "Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification". They introduce the so-called affective audio-visual words to extract emotion-specific audio-visual features, while a classification model named latent topic driving model is proposed for emotion category recognition.

III. Facial Expressions and Head Movements

Facial expressions and head movements are major visual channels for expressing and interpreting affective states. In "Robust Symbolic Dual-View Facial Expression Recognition with Skin Wrinkles: Local versus Global Approach", Huang *et al.* argue that the sketch of facial feature contour could be adequate for facial expression classification. Using skin wrinkles, they compare local and global approaches for expression recognition. Calix *et al.* study automatic emotion detection in descriptive sentences and how this can be used to tune facial expression parameters for 3-D character generation in "Emotion Recognition in Text for 3-D Facial Expression Rendering", where mutual information is adopted for word feature selection. In order to generate head movements for virtual agents, in "Predicting Speaker Head Nods and the Effects of Affective Information", Lee and Marsella adopt a machine learning approach for learning head movement models from gesture corpora. The linguistic and affective features are considered in their work, and experiments show that the learned models can predict speaker head nods with high precision and recall rates, and that using affective information can help improve the performance.

IV. Affect Analysis in Small Groups

Two other papers in this issue focus on affect analysis in small groups, an important but rarely addressed problem. In "Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior", Hung and Gatica-Perez estimate the level of cohesion in small groups using audio-visual nonverbal behavior. A series of audio and video features are proposed, which are designed and inspired by findings in the social sciences literature. In "A System for Real-time Multimodal Analysis of Nonverbal Affective Social Interaction", Varni *et al.* present a multimodal system for real-time analysis of nonverbal affective social interaction in small groups of users, with particular focus on the synchronization of the affective behavior within a group and the emergence of leadership. They develop techniques to compute quantitative measures for behavior synchronization and leadership.

V. Audio-Visual Affective Corpus

Collection of affective corpora is prerequisite of the study on emotion recognition and synthesis. In the correspondence paper "Audio-Visual Corpus of Affective Communication", Fanelli *et al.* present a new audio-visual corpus (speech and facial expression) in the form of dense dynamic 3-D face geometries. They acquire high-quality data by working in a controlled environment and resort to video clips to induce affective states.

Acknowledgment

We would like to express our sincere thanks to all the reviewers who provided the authors with timely, critical, yet constructive feedback. We also thank Mrs. D. Tomaro, Ms. R. Wollman, and the staff at the IEEE for their efficiency in organizing the review process. Last but not least, our special thanks go to the Editor-in-Chief of the IEEE Transaction on Multimedia, Prof. S. S. Hemami, for her guidance and assistance with this special issue. We hope this special issue can spark and stimulate further research and development effort in the area of multimodal affective interaction.

Nicu Sebe
University of Trento
Trento, 38100 Italy
(sebe@disi.unitn.it)

Hamid Aghajan
Stanford University
Stanford, CA 94305 USA
(aghajan@stanford.edu)

Thomas Huang
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA
(huang@ifp.uiuc.edu)

Nadia Magnenat-Thalmann
University of Geneva
Geneva, CH-1227 Switzerland
(thalmann@miralab.ch)

Caifeng Shan
Philips Research
Eindoven, 5656AE, The Netherlands
(caifeng.shan@philips.com)

Nicu Sebe is an Associate Professor in the Department of Information Engineering and Computer Science, University of Trento, Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He is an author of two monographies.

Prof. Sebe was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008 and ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, and a Program Co-Chair of the ACM Multimedia 2007 and 2011 conferences. He has served as the guest editor for several special issues in IEEE Computer, *Computer Vision and Image Understanding*, *Image and Vision Computing*, *Multimedia Systems*, ACM TOMCCAP, and the IEEE Transactions on Multimedia. He is the co-chair of the IEEE Computer Society Task Force on Human-Centered Computing and is an associate editor of the IEEE Transactions on Multimedia, *Machine Vision and Applications*, *Image and Vision Computing*, *Electronic Imaging*, and *Journal of Multimedia*.

Hamid Aghajan received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1995.

He has been a Professor of electrical engineering (consulting) at Stanford University since 2003, where he is the Director of the Ambient Intelligence Research (AIR) Lab and the Wireless Sensor Networks Lab. The areas of research in his group consist of multi-camera networks and human interfaces for smart, vision-based reasoning environments, with application to smart homes, occupancy-based services, assisted living and well being, smart meetings, and avatar-based communication and social interactions. He has coauthored three edited volumes on: *Human-Centric Interfaces for Ambient Intelligence* (New York: Academic, 2009), *Multi-Camera Networks—Principles and Applications* (New York: Academic, 2009), and *Handbook of Ambient Intelligence and Smart Environments* (New York: Springer, 2009).

Dr. Aghajan is Co-Editor-in-Chief of the *Journal of Ambient Intelligence and Smart Environments*. He has been co-founder and program co-chair of the first International Conference on Distributed Smart Cameras (ICDSC 2007), and general co-chair of ICDSC 2008. He has organized short courses on distributed vision processing, multi-camera networks, and human-centered vision systems at CVPR 2007, 2008, 2010, ICCV 2009, ICDSC 2009, and ICASSP 2009, and has served as chair at: workshops on Use of Context in Vision Processing at CVPR 2010 and ICMI-MLMI 2009, Human-Computer Interaction at ICCV 2009, Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications at ECCV 2008, Behaviour Monitoring and Interpretation at German AI Conference 2008–2009, and Vision Networks for Behaviour Analysis (VNBA) at ACM Multimedia 2008.

Thomas S. Huang received the Sc.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was on the faculty of MIT and Purdue University. He joined the University of Illinois at Urbana–Champaign in 1980 and is currently William L. Everitt Distinguished Professor of Electrical and Computer Engineering, Research Professor of Coordinated Science Laboratory, Professor of the Center for Advanced Study, and Co-Chair of the Human Computer Intelligent Interaction major research theme of the Beckman Institute for Advanced Science and Technology. He has published 21 books and more than 600 technical papers in network theory, digital holography, image and video compression, multimodal human computer interfaces, and multimedia databases.

Prof. Huang is a member of the National Academy of Engineering and has received numerous honors and awards, including the IEEE Jack S. Kilby Signal Processing Medal and the King-Sun Fu Prize of the International Association of Pattern Recognition.

Nadia Magnenat-Thalmann received several bachelor's and master's degrees in various disciplines (psychology, biology, and chemistry) and the Ph.D. degree in quantum physics from the University of Geneva, Geneva, Switzerland, in 1977.

She has pioneered research into virtual humans over the last 30 years. From 1977 to 1989, she was a Professor at the University of Montreal, Montreal, QC, Canada. Since 1989, she has been a Professor at the University of Geneva, where she founded the interdisciplinary multimedia research group MIRALab. She is the coordinator of several European Research Projects, among them the European Center of Excellence INTERMEDIA (http://intermedia.miralab.ch/) and the European Center of Excellence 3D ANATOMICAL HUMANS (http://3dah.miralab.ch/).

Prof. Magnenat-Thalmann is Editor-in-Chief of the *Visual Computer Journal* published by Springer Verlag and Co-Editor-in-Chief of the journal *Computer Animation and Virtual Worlds* published by Wiley. She was given a Doctor Honoris Causa from the Leibniz University of Hanover and an honorary doctorate from the University of Ottawa. Recently this year, she was awarded the Eurographics Distinguished Career Award.

Caifeng Shan received the B.Eng. degree in computer science from the University of Science and Technology of China (USTC), the M.Eng degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences (CAS), and the Ph.D. degree in computer science from Queen Mary, University of London, London, U.K.

He is currently a Senior Scientist with Philips Research, Eindhoven, The Netherlands. His research interests include computer vision, pattern recognition, image/video processing and analysis, multimedia, machine learning, human-computer interaction, and related applications. He has authored more than 30 refereed book chapters, journal, and conference papers and filed three patent applications. He has edited the book *Video Search and Mining* (New York: Springer-Verlag, 2010) and the book *Multimedia Interaction and Intelligent User Interfaces* (New York: Springer, 2010),

Dr. Shan was co-chair of the IEEE International Workshop on Video Mining (VM'08) at ICDM 2008, co-chair of the IEEE International Workshop on Human-Computer Interaction (HCI'09) at ICCV 2009, and co-chair of the International Workshop on Interactive Multimedia for Consumer Electronics (IMCE'09) at ACM Multimedia 2009. He has served for many major international journals and conferences as referees.