Media computing and applications for immersive communications: recent advances

Lei Xie¹ · Janne Heikkilä² · Bo Li³

Published online: 28 August 2017
© Springer-Verlag GmbH Germany 2017

1 Introduction

The vision of media computing for immersive communication is to enable natural interactions among people via advanced media technologies. Recently, various research communities have shown great interest in audio, visual and signal processing for immersive, multimodal and multimedia interactions. The goal of this special issue is to bring together researchers engaged in the development of media computing technologies and applications, and to highlight the potential and special characteristics of immersive communication.

We received 15 high-quality submissions and each paper was carefully peer-reviewed. Finally, nine manuscripts were selected to be included in this special issue. We summarize these papers in the following four major categories.

2 The importance of sound

Sound localization is essential in immersive communication and ambient intelligence. To tackle this task, the first paper

by Jia et al. is titled “Real-Time Multiple Sound Source Localization and Counting Using a Soundfield Microphone”. In this paper, a multiple sound source localization and counting method based on a relaxed sparsity of speech signal is presented. To avoid the redundancy and complexity of using microphone arrays, a soundfield microphone is adopted. The proposed method achieves a higher accuracy of direction of arrival (DOA) estimation and source counting compared with other existing techniques. Moreover, it has higher efficiency and lower complexity, which makes it suitable for real-time applications.

Voice activity detection (VAD) is a basic component in speech communication and other speech processing technologies. Recently, machine-learning-based VAD methods have drawn much attention. In the second paper, titled “Noise Robust Voice Activity Detection using Joint Phase and Magnitude based Feature Enhancement”, Phapatanaburi et al. have proposed a deep neural network (DNN) based joint phase and magnitude feature (JPMF) enhancement method for noise-robust VAD. Phase information is usually discarded in conventional methods, but this study has shown that the proposed joint phase and magnitude approach significantly outperforms the conventional magnitude-only based method.

3 Vision and multimodal

Object tracking is a classic topic in computer vision and accurately tracking objects through cameras is a fundamental goal in immersive communications. In the third paper, titled “Online Object Tracking Based on BLSTM-RNN with Contextual-Sequential Labeling”, Zhou et al. have proposed a novel appearance model by transforming the target contextual dependency into a semantic sequential representation

✉ Lei Xie
lxie@nwpu.edu.cn

Janne Heikkilä
jth@ee.oulu.fi

Bo Li
boboli@google.com

¹ School of Computer Science, Northwestern Polytechnical University, Xi'an, China

² Department of Computer Science and Engineering, University of Oulu, Oulu, Finland

³ Google Inc., Mountain View, USA

in online object tracking. Specifically, a recurrent neural network (RNN) with bidirectional long short-term memory (BLSTM) cells is used for online tracking-by-learning. The proposed tracking method has demonstrated to outperform most of state-of-the-art trackers on challenging benchmark videos.

Privacy protection is another important issue in immersive video communication. In the fourth paper, “Object Coding based Video Authentication for Privacy Protection in Immersive Communication”, Zhang et al. have proposed an object-based coding authentication strategy based on the Chinese remainder theorem for video authenticity protection. Based on the efficiency evaluation on video transmission, the proposed approach can ensure an applicable authenticity between the foreground objects and their associated background for video immersive applications.

In immersive communication, sometimes we need to interact with a robot. In the fifth paper, “Multimodal Sensory Fusion for Soccer Robot Self-localization Based on Long Short-Term Memory Recurrent Neural Network”, Wang et al. have proposed a multimodal sensory fusion method for self-localization of autonomous mobile robots. Specifically, an LSTM-RNN is used as a predictor and information from Inertia Navigation System (INS) and vision preceptors are fused at feature level. The results demonstrate that the proposed approach makes an improvement in prediction accuracy and efficiency compared with conventional methods.

4 Affective computing

Detecting the human emotion status is definitely beneficial to immersive communication and interaction. There have been increasingly interests recently in affective computing through multimodal signals. In the sixth paper, titled “Coupled HMM-based Multimodal Fusion for Mood Disorder Detection through Elicited Audio–Visual Signals”, Yang et al. have proposed a multimodal fusion approach for mood disorder detection. Facial action unit (AU) profiles and speech emotion profiles (EPs) are obtained and fused by a coupled HMM. Experimental results show the promising advantage and efficacy of the CHMM-based fusion approach for mood disorder detection.

In the seventh paper, “An Efficient Movement and Mental Classification for Children with Autism based on Motion and EEG Features”, Dang et al. are interested in activity classification based on motion and EEG features for rehabilitation training of the children with autism.

In the eighth paper, “CHEAVD: A Chinese Natural Emotional Audio–Visual Database”, Li et al. have presented an emotional audio–visual database to facilitate research on multimodal emotion analysis. Some baseline experimental results on emotion recognition have been provided as well.

5 Semantic analysis

Semantic understanding of the audio/visual contents is another popular topic, including detection of the topic changes. In the ninth paper, titled “A Hybrid Neural Network Hidden Markov Model Approach for Automatic Story Segmentation”, Yu et al. have proposed a hybrid neural network hidden Markov model (NN-HMM) approach for automatic topic/story segmentation. In this approach, the transitions between stories are modeled by an HMM and a DNN is used to directly map the word distribution into topic posterior probabilities. Experimental results show that the proposed NN-HMM approach outperforms the traditional HMM approach and achieves state-of-the-art performance in story segmentation.

6 Summary

This special issue covers wide areas of media computing in immersive communication. We sincerely hope that the readers would enjoy reading this special issue.

The guest editors would like to take this opportunity to thank all the authors for their valuable contributions. We wish to express our deepest gratitude to the referees. They have provided very useful comments and suggestions to our authors. Finally, our sincere thanks go to the Editor-in-Chief, Prof. Vincenzo Loia, for his kind support throughout the preparation of this special issue.