Guest Editorial: Special Section on “Multimedia Understanding via Multimodal Analytics”

37

Information in the real world often comes through multiple input channels, in the form of multimodalities. For instance, Web images are frequently associated with tags and textual descriptions; beside textual captions, videos contain visual and acoustic signals; sensory perception includes simultaneous inputs from visual, auditory, motor, and haptic pathways. Rather than independent, the multimodalities are correlated in a complex way, which can be temporally synchronized (e.g., video clips and corresponding audio transcripts), spatially related (e.g., object relationships in 3D space), or otherwise semantically connected (e.g., images and their tags). We thus firmly believe that joint learning by discovering and modeling the relatedness among multiple modalities will remarkably strengthen the learning performance.

This special section, “Multimedia Understanding via Multimodal Analytics,” in the *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* timely targets this opportunity and brings together active researchers from all over the world to share their recent advances in multimodal learning. As guest editors, we were happy to receive 12 submissions in this special section. Among them, four articles have been accepted, which cover interesting multimodal applications of social media event summarization, RGB-D scene understanding, image captioning, and question answering.

The first article in this section, “Multimodal Multiplatform Social Media Event Summarization” by Tiwari et al., proposes a novel framework for multiplatform and multimodal summarization. This includes an Markov Random Field (MRF)-based similarity measure that links topics across platforms as well as across time, which has shown promising results on a crowdsourced user study. To select content from different topics, the authors model the conventional summarization objectives of coverage, novelty, and significance as submodular functions in a social media setting. Traditional summarization solutions rely on gold-standard summaries, which are not practical to create manually for social media content given the size of the involved data. For the evaluation, the authors undertake crowdsourced as well as expert studies to evaluate the quality of generated summaries.

The second article in this section, “Structure-aware Multi-modal Feature Fusion for RGB-D Scene Classification and Beyond” by Wang et al., proposes a structure-aware multi-modal feature fusion framework. The authors apply it to RGB-D scene classification to effectively make use of high-dimensional Fisher Vector (FV) features from RGB, Horizontal disparity, Height above ground, and Angle (HHA), and surface normal modalities, where the authors enforce inter-component sparsity to select informative Gaussian Mixture Model (GMM) components of FV and inter-modal non-sparsity to encourage contributions from all modalities. By combining the regression results of the proposal-based multi-modal FV features and the full-image-based multi-modal Convolutional Neural Network (CNN) features, the authors achieve state-of-the-art scene classification performance on the SUNRGBD Dataset and the NYU Depth Dataset V2.

The third article in this section, “Image Captioning with Deep Bidirectional Long Short-Term Memory (LSTM) and Multi-Task Learning” by Wang et al., proposes a bidirectional LSTM model that generates descriptive sentence for image by taking both history and future context into

2018 Copyright is held by the owner/author(s).

1551-6857/2018/04-ART37

<https://doi.org/10.1145/3192334>

account. The authors further design deep bidirectional LSTM architectures to embed image and sentence at high semantic space for learning visual-language model. We proved multi-task learning of Bi-LSTM is beneficial to increase model generality and further confirmed by transfer learning experiment. The authors also qualitatively visualized internal states of the proposed model to understand how multimodal bidirectional LSTM generates words at consecutive time steps.

The fourth article in this section, “Towards Personalized Activity Level Prediction in Community Question Answering Websites” by Liu et al., proposes a method termed PALP to address the Personalized Activity Level Prediction for expert users in Community Question Answering websites. The authors have taken into consideration the behavior change of an expert over time to enable a more accurate prediction. Further, the authors have constructed and evaluated an extensive list of potential indicative features and classified them into three categories. Experimental results demonstrate the superior performance of our approach over the existing methods.

All these articles show the diversity of methods and applications for multimodal analytics. We hope that these novel articles will further research in this area and in relation to other exciting developments in the multimedia and computer vision areas.

Yan Yan

Department of Computer Science, Texas State University, USA

Liqiang Nie

Department of Computer Science, Shandong University, China

Rita Cucchiara

Department of Computer Engineering and Science,
University of Modena and Reggio Emilia, Italy

Guest Editors