Computer Vision and Image Understanding 118 (2014) 1

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Editorial

Special section on learning from multiple evidences for large scale multimedia analysis

With the popularity of digital cameras and smartphones, an explosive growing number of multimedia data are generated every day. The size of personal and internet image/video collections keeps growing rapidly. In the Web 2.0 era, the success of social multimedia websites, such as Facebook, Flickr, and Youtube, provides us a plenty of internet multimedia data. Even in a personal digital archive, there may be over ten thousand pictures and the length of video data could be over hundreds hours. Therefore, effective and efficient multimedia data analysis, which substantially benefits multimedia data utilization and management at large scales, turns into one of the greatest research challenge in the community.

The information obtained from multimedia data consists of multiple evidences, e.g., internet images are usually accompanied with a textual description and social network metadata. Learning from such multiple evidences for large scale multimedia content analysis is an interesting research topic, with a range of important applications, such as multimedia retrieval, multimedia event detection, concept detection, indexing, etc. For example, it has been reported in several recent papers that combining metadata with low level features would benefit web image analysis. As another example, the 15-year Informedia project at Carnegie Mellon University has demonstrated that combining Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR) with visual features usually yields higher multimedia event detection accuracy than only using visual features. It is therefore a promising research direction to appropriately exploit multiple evidences derived from visual, auditory, textual features and social metadata.

This special issue is presenting the latest research on combing multiple evidences for multimedia analysis. Among the 18 submissions, 5 were accepted by this special issue. Given an action specified by a user, Nga and Yanai propose a novel method to automatically retrieve the video shots of that action from Internet by jointly exploiting the metadata and visual features of web videos. An experiment on large scale dataset demonstrates that combing the two cues would help reduce human labor for building action dataset, compared to the traditional exhausted manual way. Zhang et al. build an Object Patch Net (OPN) from loosely labeled Internet images, and then perform large scale image annotation and retrieval by combining semantic information and visual features. Liu et al. use social information to better understand user intention, and combine social and visual information for web image retrieval. The fourth paper proposes a uniformed saliency model, in which semantic information and visual information are considered. The last paper proposes a new machine learning algorithm for multiple feature analysis. The experiment on image annotation using more than 10 features has demonstrated the effectiveness of the proposed algorithm.

We would like to thank the authors and the reviewers for their efforts in producing the contents of this special issue. We would also like to thank Professor Avi Kak, the editor-in-chief of Computer Vision and Image Understanding, and the Elsevier staff for making this special issue possible.

Guest Editors
Yi Yang
The University of Queensland, Australia
E-mail address: yi.yang@uq.edu.au

Nicu Sebe
University of Trento, Italy
E-mail address: sebe@disi.unitn.it

Cees Snoek
University of Amsterdam, The Netherlands
E-mail address: cgmsnoek@uva.nl

Xian-Sheng Hua
Microsoft Research, USA
E-mail address: xshua@microsoft.com

Yueting Zhuang
Zhejiang University, China
E-mail address: yzhuang@cs.zju.edu.cn

Available online 5 September 2013

http://dx.doi.org/10.1016/j.cviu.2013.09.001