Multimed Tools Appl (2015) 74:463–465
DOI 10.1007/s11042-014-2419-y

GUEST EDITORIAL

Guest Editorial: Ad Hoc Web Multimedia Analysis with Limited Supervision

Yahong Han · Yi Yang · Jingdong Wang

Published online: 10 January 2015

With the popularity of social media applications and Web 2.0 techniques, there is an explosive growth of web multimedia data generated from user-sharing web sites such as Flickr, YouTube, and Facebook. The social characteristics and the increased scalability turn out to be a great challenge in the semantic understanding and retrieval of web multimedia. Though the users' comments and tagging can be well exploited to provide more semantic cues for the web multimedia analysis, the annotations of these data contain a lot of noisy tags and are always weakly tagging. Thus, the supervision information available is limited due to the huge output space. Furthermore, for web multimedia analysis, the negative examples come from an infinite semantic space and we have no clue about the semantics these negative examples include. Thus, how to construct the generic model for each ad hoc multimedia analysis task (a.k.a. Ad Hoc Web Multimedia Analysis) is a challenging problem for the social media applications on the web.

The research of ad hoc web multimedia analysis with limited supervision is an interesting and fundamental research area, which involves several fields, ranging from machine learning, multimedia retrieval, and computer vision to data mining. This issue consists of 11 papers, which are briefly discussed as follows.

Visual content analysis is a fundamental problem in ad hoc multimedia analysis, which will facilitate the semantic understanding of multimedia data. In this issue, two papers investigate visual content analysis and its applications in super-resolution and image-based localization on mobile phone. The "Depth map Super-Resolution based on joint dictionary learning" (10.1007/s11042-014-2002-6) paper utilizes unsupervised dictionary learning for depth map super-resolution. This method transforms a low-resolution depth map to a high-resolution depth map. Different from previous depth map super-resolution methods, the proposed algorithm uses a joint dictionary learning method with both low- and high-resolution depth

Y. Han (✉)
School of Computer Science and Technology, Tianjin University, Tianjin, China
e-mail: yahong@tju.edu.cn

Y. Yang
Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia
e-mail: yee.i.yang@gmail.com

J. Wang
Microsoft Research Asia, Beijing, China
e-mail: jingdw@microsoft.com

Y. Han
Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China

maps, which can be used for 3D reconstruction. The "Memory efficient large-scale image-based localization" (10.1007/s11042-014-1977-3) paper introduces a new method to reduce the dimensionality of local descriptors and thus can be used in memory-efficient image-based localization. As image-based localization is a key application in mobile applications, the proposed method will facilitate large-scale applications owing to its advantages of low memory cost. With this new method, all descriptors are projected into a lower-dimensional space. The low-dimensional descriptors are then mapped into a Hamming space for further reducing the memory requirement.

Because it is difficult to have precisely labeled training data in the real-world applications of web multimedia analysis, how to utilize unlabeled data or weakly supervised information for the analysis of ad hoc test multimedia data is a challenging problem. In "Max-Margin Adaptive Model for Complex Video Pattern Recognition," (10.1007/s11042-014-2010-6) a max-margin adaptive (MMA) model for complex video pattern recognition was proposed, which can utilize a large number of unlabeled videos to assist the model training. The MMA model considers the data distribution consistence between labeled training videos and unlabeled auxiliary ones by learning an optimal mapping function. Experimental results on public benchmark dataset show the better performance of the method. The "Evaluation of semi-supervised learning method on action recognition" (10.1007/s11042-014-1936-z) paper conducts an evaluation research on applying different combination of pooling and semi-supervised learning method on both the synthetic and realistic action recognition datasets to see which combination or method performs better. Comprehensive evaluation results are obtained that semi-supervised learning and multi-feature fusion are effective in video action recognition. In "Boosted MIML method for weakly-supervised image semantic segmentation," (10.1007/s11042-014-1967-5) authors propose a Boosted Multi-Instance Multi-Label (BMIML) learning method for image semantic segmentation, in which the image-level labels as weakly-supervised constraints are utilized for fine-grained image analysis.

Cross-media analysis and adaptation or multi-modal fusion is an effective way to utilize limited supervision information. This issue consists of four papers about this topic. In "Accumulated Reconstruction Error Vector (AREV) A Semantic Representation for Cross-Media Retrieval," (10.1007/s11042-014-1968-4) authors explore the fundamental problem of cross-media retrieval, i.e., cross-media semantic representation. The proposed method projects individually their original feature descriptions into a shared semantic space, in which each component is semantic consistent for various media types due to the consistency in category information. Experimental results demonstrate the effectiveness of the proposed method. Toward the same problem of cross-media representation, the "Nonnegative Cross-media Recoding of Visual-Auditory Content for Social Media Analysis" (10.1007/s11042-014-1970-x) paper proposes a novel nonnegative cross-media recoding approach, which learns co-occurrences of cross-media feature space by explicitly learning a common subset of basis vectors. The nonnegative constraint makes the proposed modal more interpretable. Experiments on image-audio dataset show the better performance of the proposed method. In "Polysemious Visual Representation Based on Feature Aggregation for Large Scale Image Applications," (10.1007/s11042-014-1975-5) authors investigate the correlations between multiple image features and multiple semantic concepts. The proposed polysemious image representation consists of three levels of aggregation, i.e., codebook level, semantic level, and multiple feature level aggregation. The final polysemious representation is obtained by a weighted pooling approach. As unlabeled multimedia data always accompany multimodal examples on the social website, the "Markov Random Field Based Fusion for Supervised and Semi-supervised Multi-modal Image Classification" (10.1007/s11042-014-2018-y) paper proposes a semi-supervised multi-modal image classification method, which uses both the labeled and unlabeled examples for training. Experimental results on benchmark dataset show that the proposed method can well exploit the multi-modal data and unlabeled examples.

The issue also includes the study of learning to rank in multimedia retrieval and social multimedia mining. In "Sparse Structure Regularized Ranking," (10.1007/s11042-014-1939-9) authors propose a novel learning to rank algorithm by exploring the sparse structure and using it to regularize raking scores. In the algorithm, multimedia objects are assumed to be represented as a sparse linear combination of all other objects, and the sparse combination coefficients and the ranking scores are simultaneously learned. Experiments results show that the proposed algorithm has potential practical application in multimedia retrieval. The application in "Mining near duplicate image groups" (10.1007/s11042-014-2008-0) is interesting, which aims at the challenge of finding the near duplicate image group from web-scale social images. Instead of the searching scheme, authors utilize adaptive global feature clustering and local feature refinement methods, in which a two-layer hierarchical model is developed for both the feature clustering and near duplicate image group verification.

These 11 papers cover a wide range of methods and applications about ad hoc web multimedia analysis with limited supervision. As learning with limited supervision and its applications in multimedia analysis have attracted increasing research interest, we hope this issue appeal to both the experts in the field as well as to those who wish a snapshot of the current breadth of practical web multimedia analysis.

Acknowledgements This paper was partially supported by the NSFC (under Grant 61202166, 61472276).