Machine Vision and Applications (2014) 25:1–4
DOI 10.1007/s00138-013-0586-x

EDITORIAL

Special issue on Multimedia Event Detection

Thomas B. Moeslund · Omar Javed · Yu-Gang Jiang · R. Manmatha

Published online: 17 December 2013

Recently, the generation, storage and sharing of multimedia video data has increased at an astronomical rate. In 2012, over 100 h of videos was uploaded on YouTube every minute[1]. The multimedia data being shared covers a wide variety, ranging from homemade birthday videos to professionally produced comedy skits, and from woodworking tutorials to breaking news reports and analysis, etc. Although the storage and dissemination capacity of the network has grown exponentially, the development of automatic tools to search and retrieve this data has not kept pace, and by large, manual annotation and categorization is used for video search. Manual annotation, along with being expensive and slow, cannot express the rich content of video data. A layperson or an analyst might want to search the video not only based on the main topic (e.g., news report of a protest) but also based on the events taking place, the activities of the people and entities being viewed, the conversations taking place and the sounds recorded. To automatically detect, classify and index everything that can potentially be recorded is clearly a grand challenge and needs to be divided into manageable entities for research and development of search solutions.

One of the sub-challenges for video search is human activity and event detection. Recently, some success has been reported for large-scale object recognition, object tracking and human action detection using computer vision as well as in the automatic indexing of speech in well-defined scenarios using audio processing. But a far bigger challenge is to generalize such findings to fully unconstrained settings and significantly increase the types of searchable events and the accuracy of the retrieved results. To achieve this end, it is widely believed that exploiting multiple modalities (i.e., imagery, audio and text)—as opposed to single modalities (imagery only)—and accurately fusing information obtained from each modality for event detection is likely to help. The aim of this special issue is to promote the important topic of multi media event detection. The content of the papers published in this issue ranges from system papers focused on multi-modal frameworks, over fusion of audio–visual cues, to mono-modality event and behavior understanding, all of which are critical topics when aiming at pushing the current frontiers within Multimedia Event Detection.

In the first paper, Wei Tong et al. [1] present a method to learn a discriminative video representation that integrates audio, visual and text (from OCR and speech recognition) modalities for event recognition. This intermediate representation is a compact vector representation derived from multiple bag-of-words features, and is learned by minimizing a robust loss function. The framework can also use auxiliary information, i.e., videos not related to targeted event for intermediate representation learning. Experiments indicate that using the multi-modal intermediate representation gives better results as compared to using early or late fusion schemes to integrate the available modalities.

[1] http://www.youtube.com/yt/press/statistics.html.

T. B. Moeslund (✉)
Aalborg University, Aalborg, Denmark
e-mail: tbm@create.aau.dk

O. Javed
SRI International, Princeton, NJ, USA
e-mail: omar.javed@sri.com

Y.-G. Jiang
School of Computer Science, Fudan University, Shanghai, China
e-mail: ygj@fudan.edu.cn

R. Manmatha
School of Computer Science, University of Massachusetts, Amherst, MA 01007, USA
e-mail: manmatha@cs.umass.edu

Myers et al. [2] present a comprehensive evaluation of features and multi-modal fusion schemes for events detection in videos. A thorough analysis of usefulness of modalities and features for detection of events is presented. It is also shown that a number of complex late fusion schemes like sparse mixture models, MAP weighting, multi-stage SVM, etc. do not outperform the simple arithmetic mean for fusion in large-scale experiments.

Jhuo et al. [3] present a new joint representation for audio–visual patterns by creating bi-modal words which represent both the audio and video signals. This is done by first separately extracting visual and audio words and then building a bipartite graph to create the bi-modal word codebook. Multiple codebooks are produced. The visual and audio words are then discretized again using different pooling techniques to produce the bi-modal words. The multiple codebooks give rise to multiple representations using bi-modal words and these are combined using multiple kernel learning for event classification. Experiments show significant improvements over methods using the individual visual and audio codewords.

Oh et al. [4] describe a system for multimedia event detection involving novel (video, audio and image) features at different granularities which are fused using a novel technique. A hierarchy of features from low-level features to high-level concepts is used. First, base features such as SIFT, HoG and MFCC are extracted. Mid-level features are created from these using classifiers such as support vector machines. These mid-level features may include objects such as "tree, computer". A novel latent SVM is then used to find high-level concepts such as "skateboarding in garage" from the mid-level features. This latent SVM trained in an unsupervised manner provides the ability to model unobserved variables in training. Two novel score fusion functions are also described. Experiments are shown on the 2011 TRECVID MED dataset.

Marin-Jimenez et al. [5] introduce another interesting approach to fuse audio and visual information for human interaction recognition in unconstrained videos. The authors also adopt the bag of words representations for both audio and visual channel and evaluate several fusion methods. Several observations are discussed such as the superior performance of audio features over the visual image-based ones, which is interesting as people often only utilized visual features in the task of human interaction recognition.

Burghouts et al. [6] describe an approach that combines several interesting ideas for action detection, including negative training sample selection, a two-stage classification pipeline and the fusion of multiple features. Each of the ideas is able to lead considerable performance gains and the combination of them offers state-of-the-art results on popular benchmark datasets such as the IXMAS.

Spampinato et al. [7] investigate the interesting problem of fish behavior analysis with underwater camera, which is helpful for marine biologists who study underwater environment and climate conditions. A comprehensive system that integrates many techniques is developed to detect and track fishes. Recognition of two important behaviors of several popular fish species is then performed and promising results are reported.

John et al. [8] investigate the effects of single subspace vs. multiple subspace for human activity classification. The subspaces are learned using a Charting method, which is a very interesting and novel approach in this domain. Skeleton data is used as input to the classification system, where a layered pruning scheme reduces the complexity of the classification problem. The developed framework is evaluated on public benchmark datasets and state-of-the-art results are obtained.

Lee et al. [9] describe a system for detecting pedestrian abnormalities in outdoor video data. Reliable trajectories of pedestrians are captured by a multi-view detection and tracking approach. These trajectories are then analyzed and abnormal events, e.g., illegal entry and line forming, are detected in real-time. More sophisticated—and none real time—event detector validators are built on top of these real-time methods. The system is evaluated in different outdoor settings.

In the last paper in this special issue, Zhu et al. [10] deal with the problem of how a huge amount of video data can be summarized into a much shorter sequence. The approach uses the notion of video synopsis where spatio-temporal events are detected and stitched into one joined video sequence resulting in a highly compressed video sequence. The focus of the work is on avoiding overlapping events and hence obtain a less redundant video synopsis. The approach is evaluated on various real video sequences.

The guest editors would like to thank the authors for their efforts and interest. Moreover, we would also like to state how much we appreciate the valuable comments and suggestions provided by the reviewers. Lastly, a special thanks to the Editor-in-Chief Prof. Mubarak Shah and to the Editorial Coordinator Cherry Place for their support during the preparation and publication of this Special Issue.

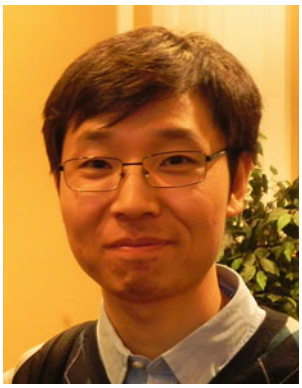
References

1. Tong, W., Yang, Y., Ma, Z., Jiang, L., Yu S.-I., Lan, Z., Sze, W., Younessian, E., Hauptmann, A.: E-LAMP: integration of innovative ideas for multimedia event Detection. Mach. Vis. Appl. doi:10.1007/s00138-013-0529-6
2. Myers, G.K., Nallapati, R., van Hout, J., Pancoast, S., Nevatia, R., Sun, C., Habibian, A., Koelma, D.C., van de Sande, K.E.A., Smeulders, A.W.M., Snoek, C.G.M.: Evaluating multimedia features and fusion for example-based event detection. Mach. Vis. Appl. doi:10.1007/s00138-013-0527-8
3. Jhuo, I.-H., Ye, G., Gao, S., Liu, D., Jiang, Y.-G., Lee, D.T., Chang, S.-F.: Discovering joint audio–visual codewords for video event detection. Mach. Vis. Appl. doi:10.1007/s00138-013-0567-0
4. Oh, S., McCloskey, S., Kim, I., Vahdat, A., Cannons, K., Hajimirsadeghi, H., Mori, G., Amitha, A.G., Megha, P., Jason, J.P.:

Multimedia event detection with multimodal feature fusion and temporal concept localization. Mach. Vis. Appl. doi:10.1007/s00138-013-0525-x
5. Marin-Jimenez, M.J., Muñoz-Salinas, R., Yeguas-Bolivar, E., de la Blanca, N.P.: Human interaction categorization by using audio–visual cues. Mach. Vis. Appl. doi:10.1007/s00138-013-0521-1
6. Burghouts, G.J., Schutte, K., Bouma, H., den Hollander, R.: Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos. Mach. Vis. Appl. doi:10.1007/s00138-013-0521-1
7. Spampinato, C., Beauxis-Aussalet, E., Palazzo, S., Beyan, C., van Ossenbruggen, J., He, J., Boom, B., Huang, X.: A rule-based event detection system for fish behaviour in underwater videos. Mach. Vis. Appl. doi:10.1007/s00138-013-0509-x
8. John, V., Trucco, E.: Charting-based subspace learning for video-based human action classification. Mach. Vis. Appl. doi:10.1007/s00138-013-0508-y
9. Lee, S.C., Nevatia, R.: Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system. Mach. Vis. Appl. doi:10.1007/s00138-013-0516-y
10. Zhu, X., Liu, J., Wang, J., Lu, H.: Key observation selection based effective video synopsis for camera-network. Mach. Vis. Appl. doi:10.1007/s00138-013-0519-8

Thomas B. Moeslund received the M.Sc.E.E. and Ph.D. degrees from Aalborg University, Denmark, in 1996 and 2003, respectively. He is currently the head of Section for Media Technology and head of the Visual Analysis of People Laboratory both at Aalborg University. He has been involved in 14 national and international research projects, as a coordinator, work package leader, and researcher. He has authored several textbooks and more than 100 peer-reviewed papers (citations: 3,864; H-index: 20). His research interests include all aspects of computer vision, with a special focus on automatic analysis of people. Professor Moeslund has co-chaired ten international conferences/workshops/tutorials and been a member of the Program Committee for a number of conferences and workshops. He serves as an Associate Editor for Machine Vision and Application and is an Editorial Board member for four international journals. He received a Most Cited Paper Award in 2009, best Paper Awards in 2010 and 2013, a Teacher of the Year Award in 2010, a Most Suitable for Commercial Application award in 2012, and the Northern Jutland University-Foundation Innovation Award in 2013.

Omar Javed is a principal scientist and technology leader at the Vision and Learning Lab at SRI International. His areas of interest include video event understanding, object tracking, multi-sensor surveillance and online machine learning. Currently, he is working on the problems of large-scale video analysis and retrieval, and persistent Intelligence, Surveillance and Reconnaissance (ISR). Javed is the author of the book, "Multi-Camera Surveillance Algorithms and Practice" and his article, "Object Tracking: A Survey" was ranked #1 in ACM's Most Popular Magazine and Computing Survey articles in 2007. His paper on "Modeling Inter-Camera Space-Time and Appearance Relationships for Tracking across Non Overlapping Views" was listed among the top 10 most cited papers in the Computer Vision and Image Understanding Journal from 2007–2011.



Yu-Gang Jiang received the Ph.D. degree in computer science from the City University of Hong Kong in 2009. During 2008–2011, he was with the Department of Electrical Engineering, Columbia University, New York. He is currently an Associate Professor of Computer Science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision. Dr. Jiang is an active participant of the annual U.S. NIST TRECVID Evaluation and has designed a few top-performing video analytic systems over the years. His work has led to a best demo award from ACM Hong Kong, the second prize of ACM Multimedia Grand Challenge 2011, a recognition by IBM T. J. Watson Research Center as one of ten "emerging leaders in multimedia" in 2009, and an award from Intel to outstanding young CS faculties in China (2013). He has served on the organizing and program committees of many international conferences and is a Guest Editor of a forthcoming special issue on Socio-Mobile Media Analysis and Retrieval, IEEE Transactions on Multimedia.

R. Manmatha is a research associate professor in the Dept. of Computer Science at the University of Massachusetts, Amherst. His research is in the areas of retrieving images/videos and handwritten documents and printed documents. He worked on the automatic annotation and retrieval of images and videos and is currently involved in a project to do event detection in videos. He proposed the idea of word spotting for handwritten documents (using word image matching to search handwritten documents). He and his students built the first automatic demonstration system for retrieving historical handwritten documents (a portion of George Washington's handwritten documents). He is also involved in a NSF project in collaboration with the Internet Archive and Tufts University to data mine scanned printed books. In addition he is involved in two projects funded by the Andrew Mellon Foundation to create tools for scholars to make such material more easily useful. He has also worked on the automatic annotation and retrieval of images and videos and is currently involved in a project to do event detection in videos. He was a co-founder of SnapTell, a mobile image search company (acquired by A9/Amazon) and is a consultant to A9/Amazon. He spent a summer as a visiting research scientist at Google working on their book project. He is an associated editor for IEEE Trans. PAMI and a program chair for ACM Intl. Conf. for Multimedia Retrieval (ICMR), 2014 and the 14th Intl. Conf. on Frontiers in Handwriting Recognition (ICFHR), 2014.