Guest Editorial: Large-Scale Multimedia Data Retrieval, Classification, and Understanding

TODAY, large collections of multimedia data are explosively created in different fields and have attracted increasing interest in the multimedia research area. Large-scale multimedia data provide great unprecedented opportunities to address many challenging research problems, e.g., enabling generic visual classification to bridge the well-known semantic gap by exploring large-scale data, offering a promising possibility for in-depth multimedia understanding, as well as discerning patterns and making better decisions by analyzing the large pool of data. Therefore, the techniques for large-scale multimedia retrieval, classification, and understanding are highly desired.

Simultaneously, the explosion of multimedia data puts urgent needs for more sophisticated and robust models and algorithms to retrieve, classify, and understand these data. For example, how is the large-scale multimedia data organized and how can it be managed to enable efficient browsing and retrieval? The researchers in this direction produce many hashing, indexing, and quantization algorithms for high-dimensional data. Another interesting challenge is, how can the traditional machine learning algorithms (proven efficient and effective in small-sized and low-dimensional data points) be scaled up to millions and even billions of items with thousands of dimensionalities? This motivated the community to design parallel and distributed machine learning platforms, exploiting GPUs as well as developing practical algorithms. Besides, it is also important to exploit the commonalities and differences between different tasks, e.g., image retrieval and classification have much in common while different indexing methods evolve in a mutually supporting way.

This special issue targets the researchers and practitioners from both academia and industry. In total, 47 submissions have been received. After two rounds of reviews, 13 papers were accepted for publication. The accepted papers are summarized below.

1) Jie Lin, Ling-Yu Duan, Shiqi Wang, Yan Bai, Yihang Lou, Vijay Chandrasekhar, Tiejun Huang, Alex Kot, and Wen Gao, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval": With emerging demand for large scale video analysis, MPEG initiated the compact descriptor for video analysis (CDVA) standardization in 2014. Unlike handcrafted descriptors adopted by the ongoing CDVA standard, in this work, the authors study the problem of deep learned global descriptors for video matching, localization, and retrieval. First, they propose a nested invariance pooling (NIP) method to derive compact deep global descriptors from convolutional neural networks, by progressively encoding translation, scale, and rotation invariances into the pooled descriptors. Second, they design hybrid pooling operations within NIP (HNIP) to further improve the discriminability of deep global descriptors. Third, the advantages and performance on the combination of deep and handcrafted descriptors are provided to better investigate the complementary effects of them. Experimental results show that HNIP outperforms state-of-the-art deep and canonical handcrafted descriptors with significant mAP gains of 5.5% and 4.7%.

2) Shanshan Yao, Baoning Niu, and Jianquan Liu, "Audio identification by sampling sub-fingerprints and counting matches": It is challenging to retrieve audio clips from large audio data sets not only due to the high dimensionality of audio but also due to the large number of audios. Fingerprinting methods primarily focus on the use of semantic-level techniques to speed up retrieval and neglect low-level support. This paper shows that the performance of audio retrieval can be exploited by properly organizing and manipulating audio fingerprint data. A sampling and counting method that markedly improves the retrieval speed while maintaining a high recall rate and high precision for short audio clips is proposed. An inverted index structure for fingerprints that quickly shrinks the scope of the candidate set while requiring considerably less memory is proposed. The experiments show that the proposed method is faster and yields more consistent performance in terms of recall rate and precision than do the state-of-the-art methods.

3) Zhixiang Chen, Jiwen Lu, Jianjiang Feng, and Jie Zhou, "Nonlinear sparse hashing": To facilitate fast similarity search, this paper proposes to encode the nonlinear similarity and image structure as compact binary codes. Rather than adopting single matrix as projection in the literature, the authors employ a nonlinear transformation in the form of multilayer neural network to generate binary codes to capture the local structure between data samples. Specifically, they train the network such that the quantization loss is minimized and the variance over all bits is maximized. In addition, they capture the salient structure of image samples at the abstract level with sparsity constraint and inherit the generalization power to unseen samples. Furthermore, they incorporate the supervisory label information into the learning procedure to take advantage of the manual label. To obtain the desired binary codes and the parameterized non-

Digital Object Identifier 10.1109/TMM.2017.2733638

linear transformation, they optimize the formulated objective problem over each variable with an iterative alternating method. Experimental results on three public data sets show a superior performance than several recent proposed hashing methods.

4) Rameswar Panda and Amit K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization": Most traditional video summarization methods are designed to generate effective summaries for single-view videos, and thus cannot fully exploit the complicated intra- and inter-view correlations in summarizing multi-view videos in a camera network. In this paper, with the aim of summarizing multi-view videos, the authors introduce a novel unsupervised framework via joint embedding and sparse representative selection. The objective function is two-fold. The first is to capture the multi-view correlations via an embedding, which helps in extracting a diverse set of representatives. The second is to use a ℓ_{21}-norm to model the sparsity while selecting representative shots for the summary. An efficient alternating algorithm based on half-quadratic minimization is introduced to solve the proposed non-smooth and non-convex objective with convergence analysis. Rigorous experiments on several multi-view data sets demonstrate that the approach clearly outperforms the state-of-the-art methods.

5) Fumin Shen, Yang Yang, Li Liu, Wei Liu, Dacheng Tao, and Heng Tao Shen, "Asymmetric binary coding for image search": Learning to hash has attracted broad research interests in recent computer vision and machine learning studies, due to its ability to accomplish efficient approximate nearest neighbor search. However, the closely related task, maximum inner product search, has rarely been studied in the literature. This work introduces a general binary coding framework based on asymmetric hash functions, named asymmetric inner-product binary coding (AIBC). In particular, AIBC learns two different hash functions which can reveal the inner products between original data vectors by the generated binary vectors. Although conceptually simple, the associated optimization is very challenging due to the highly non-smooth nature of the objective that involves sign functions. The algorithm tackles the non-smooth optimization in an alternating manner, by which each single coding function is optimized in an efficient discrete manner. Extensive experiments on several image retrieval benchmarks validate the superiority of the AIBC approaches over many recently proposed hashing algorithms.

6) Xiaofeng Zhu, XuelongLi, Shichao Zhang, Zongben Xu, Litao Yu, and Can Wang, "Graph PCA hashing for similarity search": This paper proposes a new hashing framework to conduct similarity search via the following steps: 1) employing linear clustering to obtain a set of representative data points and a set of landmarks of the big data set; 2) using the landmarks to generate a probability representation for each data point. The proposed probability representation method is further proved to preserve the neighborhood of each data point; and 3) integrating PCA with manifold learning to learn the hash functions using the probability representations of all representative data points. Therefore, the proposed hashing method achieves efficient similarity search, effective performance, and high generalization ability. Experimental results on four public data sets demonstrate the advantages of the proposed method in terms of similarity search compared to the state-of-the-art hashing methods.

7) Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen, "Video captioning with attention-based LSTM and semantic consistency": Recent progress in using long short-term memory (LSTM) for image captioning has motivated the exploration for automatically describing video content with natural language sentences. However, most existing methods compress an entire video shot or frame into a static representation, without considering attention mechanisms which allows for selecting salient features. To tackle this issue, the authors propose a novel end-to-end framework named aLSTMs, an attention-based LSTM model with semantic consistency, to transfer videos to natural sentences. This framework integrates attention mechanism with LSTM to capture salient structures of video, and explores the correlation between multi-modal representations (i.e., words and visual content) for generating sentences with rich semantic content. Experiments demonstrate that the method using single feature can achieve competitive, or even better, results than the state-of-the-art for video captioning in both BLEU and METEOR.

8) Xiantong Zhen, Feng Zheng, Ling Shao, Xianbin Cao, and Dan Xu, "Supervised local descriptor learning for human action recognition": Local features have been widely used in computer vision tasks, e.g., human action recognition, while it tends to be an extremely challenging task to deal with large-scale local features of high dimensionality with redundant information. In this paper, the authors propose a novel fully supervised local descriptor learning algorithm called discriminative embedding method based on the image-to-class distance (I2CDDE) to learn compact but highly discriminative local feature descriptors for more accurate and efficient action recognition. By leveraging the advantages of the I2C distance, the proposed I2CDDE incorporates class labels to enable fully supervised learning of local feature descriptors, which achieves highly discriminative but compact local descriptors. They apply the proposed I2CDDE algorithm to human action recognition on four widely used benchmark data sets. The results have shown that I2CDDE can significantly improve I2C-based classifiers and achieves state-of-the-art performance.

9) Lei Zhu, Zi Huang, Xiaobai Liu, Xiangnan He, Jiande Sun, and Xiaofang Zhou, "Discrete multimodal hashing with canonical views for robust mobile landmark search": Mobile landmark search recently receives increasing attention for its great practical values. However, it remains unsolved due to two important challenges. One is high bandwidth consumption of query transmission, and the other is the huge visual variations of query images sent from mobile devices. In this paper, the authors propose a novel hashing scheme, named as canonical view-based discrete multimodal hashing (CV-DMH), to handle these problems via a novel three-stage learning procedure. A submodular function is designed to measure visual representativeness and redundancy of a view set, and multimodal sparse coding is applied to transform visual features from multiple modalities into an intermediate representation. Compact binary codes are learned on intermediate representation within a tailored discrete binary embedding model which preserves visual

relations of images measured with canonical views and removes the involved noises. Experiments on real-world landmark data sets demonstrate the superior performance of CV-DMH over several state-of-the-art methods.

10) Na Zhao, Hanwang Zhang, Richang Hong, Meng Wang, and Tat-Seng Chua, "VideoWhisper: Towards discriminative unsupervised video feature learning with attention-based recurrent neural networks": In this work, the authors present VIDEOWHISPER, a novel approach for unsupervised video representation learning. Based on the observation that the frame sequence encodes the temporal dynamics of a video (e.g., object movement and event evolution), they treat the sequential frame order as a self-supervision to learn video representations. VIDEOWHISPER is driven by a novel video "sequence-to-whisper" learning strategy. Specifically, for each video sequence, they use a pre-learned visual dictionary to generate a sequence of high-level semantics, dubbed "whisper," which can be considered as the language describing the video dynamics. In this way, they model VIDEOWHISPER as an end-to-end sequence-to-sequence learning model using attention-based recurrent neural networks. This model is trained to predict the whisper sequence and hence it is able to learn the temporal structure of videos. Through extensive experiments on two real-world video data sets, they demonstrate that video representation learned by VIDEOWHISPER is effective to boost multimedia applications such as video retrieval and event classification.

11) Henning Muller and Devrim Unay, "Retrieval from and understanding of large-scale multi-modal medical data sets: A review": Content-based multimedia retrieval has been an active research domain since the mid-1990s. In the medical domain, visual retrieval started later and has mostly remained a research instrument and less a clinical tool, even though a few tools for retrieval are employed in clinical work. The limited size of data sets due to privacy constraints is often mentioned as a reason for these limitations. Nevertheless, much work has been done in medical visual information retrieval, including the availability of increasingly large data sets and scientific challenges. Annotated data sets and clinical data for the images have now become available and can be combined for multi-modal retrieval. This text is a systematic review of recent work (concentrating on the period between 2011 and 2017) on content-based multi-modal retrieval and image understanding in the medical domain. The text highlights the areas of advances in the past six years and in particular a trend to use larger-scale training data sets as well as deep learning approaches that can replace or complement hand-crafted feature extraction.

12) Sheng Tang, Yu Li, Lixi Deng, and Yongdong Zhang, "Object localization based on proposal fusion": Traditional regression framework of object localization such as Overfeat often suffers from the problem of inaccurate scoring due to the separate scoring of classification network and regression network upon inconsistent regions. To tackle this problem, in this paper, the authors propose a novel object localization framework based on multiple complementary region proposal methods from the view of classification rather than regression. On top of their framework, they first combine multiple complementary region proposals during both training and testing as a means of data augmentation to generate more dense and reliable proposals for fusion, then achieve optimal compromise between complexity and efficiency through category clustering for bounding box sharing among similar categories, and finally propose dense proposal fusion approach to merge dense region proposals near true object for fine-tuning of the final bounding box's coordinates and updating the confidence of fused proposals for final decision. Extensive experiments on the well-known large-scale ILSVRC 2015 LOC data set verify the effectiveness of their framework.

13) Guoyu Lu, Liqiang Nie, and Chandra Kambhamettu, "Large-scale tracking for images with few textures": In this paper, the authors propose a method which makes use of a limited number of discriminate features to explore other features without strong discriminant power. They develop a feature integrating surrounding salient points distribution knowledge, raw pixel value, and coordinate information to discover a significant amount of features in weak textured areas in an image. They also incorporate epipolar geometry in feature correspondence calculation by taking the distance from the matching candidate to its corresponding point's epipolar line into account. To reduce the number of unreliable features, they project the estimated 3D points back to the images. The re-projection error is normalized according to the 3D point's depth, which reduces the bias introduced by the object distance to the camera. They conduct experiments on a large data set of Arctic sea ice images, which are mainly composed by planes of ices and sea water. The experiment results demonstrate their method can perform fast and accurate tracking in weak textured images.

We would like to thank the reviewers for providing helpful questions and enlightening suggestions. We also thank the editorial support by the IEEE TRANSACTIONS ON MULTIMEDIA.

J. SONG, *Guest Editor*
University of Electronic Science and Technology of China
Chengdu 611731, China

H. JEGOU, *Guest Editor*
Facebook
France/USA

C. SNOEK, *Guest Editor*
University of Amsterdam
Amsterdam 94323, The Netherlands

Q. TIAN, *Guest Editor*
University of Texas at San Antonio
San Antonio, TX 78249-1604 USA

N. SEBE, *Guest Editor*
University of Trento
Trento 38123, Italy