Multimedia Systems (2017) 23:1–3
DOI 10.1007/s00530-016-0527-4

EDITORIAL

Special issue on "visual semantic analysis with weak supervision"

Luming Zhang[1] · Yang Yang[1] · Rongrong Ji[2] · Roger Zimmermann[1]

Published online: 24 January 2017
© Springer-Verlag Berlin Heidelberg 2016

Every day, the volume of visual data that we handle is increasing exponentially due to the various factors, i.e., the availability of ubiquitous and cheap sensors, sharing platforms, and social trends. In this circumstance, effectively managing the large amount visual data is becoming an important yet challenging research topic. Artificial intelligence has been proven to be effective for interpreting and managing this preponderance of data. It is generally accepted that a tough challenge in managing the large-scale visual data is the "semantic-gap" problem. To effectively fill the semantic gap of these visual media, weakly supervised learning paradigms are developed, focusing on an intelligent mechanism that transfers the image/video-level semantics to different regions. Compared with the labor-intensive labeling in fully supervised setting, the transferring mechanism can greatly reduce the human effort, especially given the tremendous visual data on the Websites. Extensive research efforts have been dedicated to weakly supervised learning the visual semantics, while effective tools to manipulate these data are still at their infancy. This special issue will target the most recent technical progresses on learning techniques for visual semantic understanding with weak supervision, such as weakly labeled image segmentation, photo cropping, video summarization, and so on. This special issue also targets on applying new types of weak supervision in visual semantic modeling, e.g., image retrieval based on user feedback and interactive image rendering.

Submissions came from an open call for paper, and finally, more than ten papers are selected after rigorous reviews. These papers cover wide applications based on weakly supervised graph inference, weakly supervised feature selection, image retrieval using weak labels, and so on. A brief introduction of some representative papers is as follows:

In "A Discriminative Graph Inferring Framework towards Weakly Supervised Image Parsing", Yu et al. introduced a unified Discriminative Graph Inferring (DGI) framework by simultaneously inferring patch labels and learning the patch appearance models. The graph is constructed by connecting the nearest neighbors which share the same image label, and multiple correlations among patches and image labels are imposed as constraints to the inferring. For each label, the patches which do not contain the target label are adopted as negative samples to learn the appearance model. In this way, the predicted labels will be more accurate in the propagation. Graph inferring and the learned patch appearance models are finally embedded to complement each other in one unified formulation.

In "3D Object Retrieval based on Viewpoint Segmentation", Leng et al. proposed a normal method for view segmentation, based on Markov random field (MRF) model, which consider not only the difference between the content of views but also the relative locations. Each view is obtained by projecting at certain view-points and angels; therefore, these locations can be applied to depict each view, with content of views. The authors use the MRF to implement view segmentation and choose the

✉ Luming Zhang
zglumg@nus.edu.sg

Yang Yang
dcsyangy@nus.edu.sg

Rongrong Ji
rrji@ee.columbia.edu

Roger Zimmermann
rogerz@comp.nus.edu.sg

[1] National University of Singapore, Singapore, Singapore

[2] Xiamen University, Xiamen, China

representative views. Finally, they present a framework based on the proposed views segmentation method for 3D object retrieval and the experimental results demonstrate that the proposed method can achieve better retrieval effectiveness than the state-of-the-art methods under several standard evaluation measures.

In "Tag Relevance Fusion for Social Image Retrieval", Li et al. presented a systematic study, covering tag relevance fusion in the early and late stages, and in supervised and unsupervised settings. Experiments on a large present-day benchmark set show that tag relevance fusion leads to better image retrieval. Moreover, unsupervised tag relevance fusion is found to be practically as effective as supervised tag relevance fusion, but without the need of any training efforts. This finding suggests the potential of tag relevance fusion for real-world deployment.

In "Graph-based Clustering and Ranking for Diversified Image Search", Yan et al. described a novel framework for Web image search results clustering and re-ranking, the goal is to improve diversity at high ranks. To cluster the Web image search results, the authors explore the surrounding text to mine the correlations between words and images and use the correlations to improve clustering results. The linkage between word and image is reinforced by combining tf—idf method with a novel feature of nouns, i.e., /emphvisibility. For the correlations between two words, we define topic relevance of words by maximum probability of a latent topic distributing over these two words simultaneously.

In "Organizing photographs with geospatial and image semantics", Zhu et al. proposed a two-level clustering scheme that exploits both the geographic location and the semantics in photos' annotations for clustering. To improve precision and recall, the authors also proposed a semantic enhancement method combined with a new term weight function to measure the semantic similarity between photos. In this way, the proposed method can effectively measure the semantic correlation for photos that have synonym representation but not the same words. In the second-level spectral clustering step, the authors also proposed a method to determine the parameters for the semantic clustering algorithm. The method provides a novel approach towards organization and manipulation of massive geotagged photo collections.

In "Semi-supervised Tensor Learning for Image Classification", Zhang et al. proposed a new tensor-based representation algorithm for image classification. The algorithm is realized by learning the parameter tensor for image tensors. One novelty is that the parameter tensor is learned according to the Tucker Tensor Decomposition as the multiplication of a core tensor with a group of matrices for each order, which endows the algorithm, preserved the spatial information of image. We further extend the proposed tensor algorithm to a semi-supervised framework, to utilize both labeled and unlabeled images. The objective function can be solved using the alternative optimization method, where at each iteration, the authors solve the typical Ridge Regression problem to obtain the closed-form solution of the parameter along the corresponding order.

In "Robust Visual Tracking via Discriminative Appearance Model Based on Sparse Coding", Zhao et al. formulated visual tracking as a binary classification problem using a discriminative appearance model. To enhance the discriminative strength of the classifier in separating object from the background, an over-complete dictionary-containing structure information of both object and background is constructed which is used to encode the local patches inside the object region with sparsity constraint. These local sparse codes are then aggregated for object representation, and a classifier is learned to discriminate the target from the background. The candidate sample with largest classification score is considered as the tracking result. Different from recent sparsity-based tracking approaches that update the dictionary using a holistic template, the authors introduce a selective update strategy based on local image patches which alleviate the visual drift problem, especially when severely occlusion occurs.

In "A Human Motion Feature based on Semi-supervised Learning of GMM", Qi et al. presented a novel statistic feature to represent each motion according to the pre-labeled categories of key-poses. A probabilistic model is trained with semi-supervised learning of the Gaussian Mixture Model (GMM). Each pose in a given motion could then be described by a feature vector of a series of probabilities by GMM. A motion feature descriptor is proposed based on the statistics of all pose features. The experimental results and comparison with existing work show that our method performs more accurately and efficiently in motion retrieval and annotation.

In "Semantic classification for hyperspectral image by integrating distance measurement and relevance vector machine", Liu et al. focused on creating a semantic annotation-based image classification method with relevance and similarity measurement. First, the computational model of relevance vector machine is utilized to perform cluster computation for hyperspectral image data. Then, multi-distance learning algorithm is optimized as holding capability for multiple dimensions data. The proposed multi-distance learning algorithm with multiple dimensions is used to measure the similarity, according to the result of cluster computation through relevance vector machine. Finally, semantic annotation is introduced to complete classification of hyperspectral image with semantic concept. Validation with the ground truth data illustrates that the proposed method can provide more accurate and integrated classification result compared with the other methodologies.

Therefore, the integration of similarity and relevance measurement is able to improve the performance of hyperspectral image classification.

In "Image Feature Detection Algorithm Based on the Spread of Hessian Source", Zhu et al. proposed an image feature point detection method based on second-order characteristics of point and the image feature detection algorithm based on the Hessian matrix to detect more feature points. By combining the gray-scale-based image-matching technology with the feature-based image feature detection technology, they propose a Hessian algorithm to obtain more matching points, which can search for matching more quickly. The proposed algorithm overcomes the traditional matching methods that have Ergodic properties of the search strategy. Experiments demonstrate the speed and accuracy of the proposed algorithm, and we use the correct detected feature points to realize image registration, image fusion, and image stitching.

In "Vehicle Collision Risk Estimation Based on RGB-D Camera for Urban Road", Shan et al. proposed a framework for collision risk estimation using RGB-D camera for vehicles running on the urban road, where the depth information is fused with the video information for accurate calculating the position and speed of the vehicles, two essential parameters for motion trajectory estimation. Considering that the motion trajectory or its differences can be considered as a steady signal, a method based on autoregressive integrated moving average (ARIMA) models is presented to predict vehicle trajectory. Then, the collision risk is estimated based on the predicted trajectory. The experiments are carried out on the data from the real vehicles. The result shows that the accuracy of position and speed estimation can be guaranteed within urban road and the error of trajectory prediction is very minor which is unlikely to have a significant impact on calculating the probability of collision in most situations, so the proposed framework is effective in collision risk estimation.

In "Learning for classification of traffic-related object on RGB-D data", Shi et al. argued that since RGB-D data can provide the depth information and thus make it capable of tackling the bagging issues, such as overlapping, clustered background, the depth data obtained by Microsoft Kinect sensor is introduced in the proposed method for efficiently detecting and extracting the objects in the traffic scene. Moreover, we construct a feature vector, which combine the Histograms of Oriented Gradients, 2D features, and 3D Spin Image features, to represent the traffic-related objects. The feature vector is used as the input of the random forest for training a classifier and classifying the traffic-related objects. In experiments, by conducting efficiency and accuracy tests on RGB-D data captured in different traffic scenarios, the proposed method performs better than the typical support vector machine (SVM) method. The results show that traffic-related objects can be efficiently detected and the accuracy of classification can achieve higher than 98 %.

In "Effective Optimizations of Cluster-Based Nearest Neighbor Search in High-Dimensional Space", Le et al. extend the HB method to address exact NN search in correlated, high-dimensional vector data sets extracted from large-scale image database by introducing two new pruning/selection techniques, we call it HB+ . The first approach aims at selecting more quickly the subset of hyperplanes/clusters that must be considered. The second technique prunes irrelevant points in the selected subset of clusters. Performed experiments show the improvement of HB+ with respect to HB in terms of efficiency (I/O cost and CPU response time) and also demonstrate the superiority over other exact NN indexes.

In "Multiple Level Visual Semantic Fusion Method for Image Re-ranking", Qi et al. tried to bridge the semantic gap by looking for the complementary of different mid-level features. In this paper, a framework is proposed to improve image re-ranking by fusing multiple mid-level features together. The framework contains three mid-level features (DCNN-ImageNet attributes, Fisher vector, sparse coding SPM) and a semi-supervised multi-graph-based model that combining these features together. In addition, the framework can be easily extended to utilize arbitrary number of features for image re-ranking. The experiments are conducted on the a-Pascal data set, and our approach that fuses different features together is able to boost performance of image re-ranking efficiently.

To conclude, the papers in this special issue cover different techniques and applications of weakly supervised model. This special issue has attracted concentrated attentions in the related research and will benefit researchers and practitioners in this emerging topic.

Acknowledgments We also thank the reviewers for their efforts to guarantee the high quality of this special issue. Finally, we would like to thank all the authors who have contributed to this special issue. Thanks to all the people who help us to make this special issue a successful one.