# Learning Semantics From Multimedia Web Resources: An Introduction to the Special Issue

## I. Introduction

RAPID advances in technology for capturing, processing, distributing, storing, and presenting visual data has resulted in a proliferation of multimedia data in the worldwide web (WWW). This is reflected in the success of many social websites, such as Flickr, Youtube, and Facebook, which drastically increased the volume of community-shared media resources, including images and videos. These websites allow users not only to create and share media data but also to rate and annotate them. Thus lots of meta-data, such as user-provided tags, comments, geo-tags, capture time and EXIF information, associated to multimedia resources, are available in the WWW. Therefore, there is a crucial demand of methods to organize and understand these repositories.

The multimedia research community has widely recognized the importance of learning effective models for understanding, organization, and access but has failed to make rapid progress due to the insufficiency of labeled data, which typically comes from users in an interactive labor-intensive manual process. In order to reduce this manual effort, many semi-supervised learning or active learning approaches have been proposed. Nevertheless, there is still a need to manually annotate a large set of images or videos to bootstrap and steer the training. The rich information clues associated with the multimedia data in the WWW offer a way out. If we can learn the models for semantic concepts effectively from user-shared data by using their associated meta-text as training labels, or if we can infer the semantic concepts of the multimedia data directly from the data in the Internet, manual efforts in multimedia annotation can be reduced. Consequently, semantic-based multimedia retrieval can benefit much from the community-contributed resource.

There is, however, a problem in using the associated meta-information as training labels: they are often very noisy. Thus how to remove the noise in the training labels or how to handle the noise in the learning process are urgent research topics. Besides modeling media items (e.g., an image or a video), the WWW is providing incredible resources to model users, through the aggregation of their traces on social media sites (e.g., the images they upload, the tags they use, the people whose content they comment on). So in addition to model multimedia data only, how to model people's behaviors or events is also important.

The goals of this special issue are three-fold: 1) introduce novel research in learning from resources in the Internet; 2) survey on the progress of this area in the past years; 3) discuss new applications based on the learned models.

## II. Review Process

This special issue solicited contributions on a wide range of related topics. We received 34 submissions ranging from those describing robust solutions for targeted applications, to those presenting models applicable to multiple problem domains.

To mirror the variety of topics and approaches, the Guest Editors represent expertise in diverse areas: Qi Tian in multimedia retrieval, Jinhui Tang in multimedia search and social media mining, Marcel Worring in multimedia semantics, and Daniel Gatica-Perez in social computing.

The corresponding authors are diverse in geographic location, including Asia, North America, and Europe. The majority of the papers were reviewed by at least three experts. The reviewer decision for each paper was discussed among the four guest editors and 13 papers were finally accepted for the special issue.

## III. Guide to Accepted Papers

The papers included in this special issue provide an excellent sampling of the recent work on learning semantics from multimedia Web resources. They fall into four categories which we address below.

### A. Knowledge Mining Based on Social Media

Social media is embedded with rich information. Knowledge mining based on social media analysis is significant for diverse applications. Chen *et al.* propose a tag-based image retrieval framework to improve the retrieval performance of a group of related personal images captured by the same user within a short period of an event by leveraging millions of training web images and their associated rich textual descriptions. A new classification method called SVM with Augmented Features is used to learn an adapted classifier by leveraging the pre-learned SVM classifiers of popular tags that are associated with a large number of relevant training web images. Based on the refined relevance scores, their proposed framework can be readily applied to tag-based image retrieval for a group of raw consumer photos without any textual descriptions or a group of Flickr photos with noisy tags.

Sang *et al.* exploit social annotations and propose a novel framework which simultaneously considers the user and query relevance to learn to personalize image search. The basic premise is to embed the user preference and query-related search intent into user-specific topic spaces. It enriches the users' annotation pool before construction of user-specific topic spaces. The TA Ranking based Multi-correlation Tensor Factorization model is proposed to perform annotation prediction, which are considered as users' potential annotations for the images. Besides, user-specific Topic Modeling maps the query relevance and user preference into the same user-specific topic space.

Zhang *et al.* present a face annotation system to automatically collect and label celebrity faces from the web. A large-scale celebrity name vocabulary is constructed to identify candidate

Digital Object Identifier 10.1109/TMM.2012.2208026

names from the surrounding text. The celebrity names are assigned to the faces by label propagation on a facial similarity graph.

Xia *et al.* address an interesting topic of relationship of people in human-centered images. They develop a transfer subspace learning based algorithm in order to reduce the significant differences in the appearance distributions between children's and old parents' facial images. By exploring the semantic relevance of the associated metadata, it predicts the most likely kin relationships embedded in an image.

Ma *et al.* propose a feature selection method and apply it to Web image annotation. It integrates two state-of-the-art innovations from shared feature subspace uncovering and joint feature selection with sparsity. This approach can be readily applied for annotation of social media on the Web.

### B. Visual Concept Learning

Sufficient training examples are essential for effective learning of semantic visual concepts. Recently the rapid popularization media data on the Web has made it possible to collect training exemplars without human assistance. Kuo *et al.* leverage both the image contents and associated textual information in the social media to approximate the semantic representations for the two modalities. It augments each image with relevant semantic features by capturing textual and visual relations among images in graphs, and automatically discovers relevant semantic features by propagation and selection of these graphs.

Zhu *et al.* propose to collect training samples from the noisily tagged Web images for visual concept learning. It maximizes two important criteria, relevancy and coverage, of the automatically generated training sets. An ontology-based hierarchical pooling method is presented to collect samples, not only based on the target concept, but also from ontologically neighboring concepts.

Wang *et al.* propose to discover image semantics in codebook derivative space. The method encodes information from the codebook derivative space to enhance image modeling, and introduces a weighted pooling strategy based on Bhattacharyya distance to generate a novel image representation.

Ewerth *et al.* present a web-supervised system for long-term learning of visual concepts that uses training data from the WWW. The system continuously updates its learned models while at the same time, it deals with scalability by retaining only a small number of training images.

Li *et al.* propose a bi-concept image search engine by establishing bi-concepts as a new method to search for the co-occurrence of two visual concepts in unlabeled images. This engine is equipped with bi-concept detectors directly, rather than artificial combinations of individual single-concept detectors.

### C. Video Analytics

Web media data lends itself as a rich source for video analysis. Wang *et al.* present an event driven web video summarization approach based on tag-localization and key shot mining. It first localizes the tags that are associated with each video into its shots. The relevance scores of the shots with respect to the event query are then estimated. After that, a set of key shots are identified by performing near-duplicate key frame detection.

Ikizler-Cinbis *et al.* show that Web images can be used to annotate videos taken in uncontrolled environments. They retrieve action images from the Web, using them to annotate generic and challenging videos.

### D. Query Difficulty Estimation

Another interesting topic is query difficulty prediction (QDP), which attempts to predict the quality of the search result for a query over a given collection. Tian *et al.* investigate the QDP problem in Web image search. A novel method is proposed to automatically predict the quality of image search results for an arbitrary query. This model is built based on a set of valuable features which are designed by exploring the visual characteristics of images in the search results. QDP can be readily used in applications, such as optimal image search engine selection and search results merging.

## Acknowledgment

We thank all the reviewers for their valuable comments that ultimately ensure the high quality of the special issue, and all the contributing authors for their interesting and innovative work. We would also like to thank past and current Editors-in-Chief Prof. S. Hemami and Prof. M. Van der Schaar for sharing our vision and providing guidance. The editorial staff of T-MM, especially R. Wollman, has been extremely supportive, helpful, and patient throughout the entire process.

Qi Tian, *Lead Guest Editor*
University of Texas at San Antonio
San Antonio, TX 78249 USA
qitian@cs.utsa.edu

Jinhui Tang, *Guest Editor*
Nanjing University of Science and Technology
Nanjing, Jiangsu, China
jinhuitang@mail.njust.edu.cn

Marcel Worring, *Guest Editor*
University of Amsterdam
Amsterdam, The Netherlands
m.worring@uva.nl

Daniel Gatica-Perez, *Guest Editor*
Swiss Federal Institute of Technology, Lausanne (EPFL)
Idiap Research Institute
Martigny, Switzerland
gatica@idiap.ch