Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search

Yue Gao, Meng Wang, *Member, IEEE*, Zheng-Jun Zha, *Member, IEEE*, Jialie Shen, *Member, IEEE*, Xuelong Li, *Fellow, IEEE*, and Xindong Wu, *Fellow, IEEE*

***Abstract*— Due to the popularity of social media websites, extensive research efforts have been dedicated to tag-based social image search. Both visual information and tags have been investigated in the research field. However, most existing methods use tags and visual characteristics either separately or sequentially in order to estimate the relevance of images. In this paper, we propose an approach that simultaneously utilizes both visual and textual information to estimate the relevance of user tagged images. The relevance estimation is determined with a hypergraph learning approach. In this method, a social image hypergraph is constructed, where vertices represent images and hyperedges represent visual or textual terms. Learning is achieved with use of a set of pseudo-positive images, where the weights of hyperedges are updated throughout the learning process. In this way, the impact of different tags and visual words can be automatically modulated. Comparative results of the experiments conducted on a dataset including 370+ images are presented, which demonstrate the effectiveness of the proposed approach.**

***Index Terms*— Hypergraph learning, social image search, tag, visual-textual.**

I. Introduction

THE RAPID development of multimedia and network technologies have lead to an explosive growth of social media in recent years. Therefore, efficient search technologies for social media corpus, such as Flickr[1] and Youtube,[2] are of great importance. Unlike general media search that heavily relies on the contextual text information, such as titles, surrounding text and alternative texts on web pages, and content-based multimedia retrieval [1]–[8], social media data are frequently associated with user generated tags that describe the images, provide meta information (i.e. date, location, etc.), or express any other sentiment. These tags can be used to index the multimedia data to facilitate their search. Extensive research efforts [9]–[11] have been dedicated to tag-based multimedia analysis.

However, conventional tag-based social image search methods cannot achieve satisfactory results for two reasons. First, there is too much noise in user-provided tags. Many tags are irrelevant or incorrectly spelled. As reported in [12], only about 50% of the tags provided by Flickr users are really related to the images. Second, it lacks an optimal ranking strategy. Take Flickr website as an example. There are two ranking options for tag-based social image search, namely, time-based ranking and interestingness-based ranking. The time-based ranking method ranks images based on the uploading time of each image, and the interestingness-based ranking method ranks images based on each image's interestingness in Flickr. These methods do not take the visual content and tags of images into consideration. Therefore, both of these two ranking strategies are not based on relevance measure, and thus the search results are not sufficiently good in terms of relevance. Therefore, efficient tag-based social image search methods are highly desired. The task is closely related to a key scientific challenge claimed by Yahoo research: "how do we combine both content-based retrieval with tags to do something better than either approach alone for multimedia retrieval."[3]

Several algorithms have been designed to improve the relevance of social image search [13]–[15]. However, most existing methods usually explore visual content and tags separately or sequentially. For example, Li et al. [16] propose a relevance-based ranking method for social image search, in which only the visual information is employed to calculate the relevance score for each image. Liu et al. [17] introduce a relevance-based ranking method that uses tags and visual contents sequentially. In this method, the tag information is first employed to generate initial relevance scores, and then the visual contents of images are used to refine the scores. Though many tags are noisy, there are also meaningful tags which are closely correlated with the visual content of the image and they are both informative in reflecting an image's

Manuscript received October 10, 2011; revised May 19, 2012; accepted May 21, 2012. Date of publication June 5, 2012; date of current version December 20, 2012. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2012CB316400, the National Natural Science Foundation of China under Grant 61125106, the National 863 Program of China under Award 2012AA011005, and the U.S. National Science Foundation (NSF) under Grant CCF-0905337. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bulent Sankur.

Y. Gao is with the Department of Automation, Tsinghua University, Beijing 100084, China.

M. Wang is with the Computer Science and Information Engineering Department, Hefei University of Technology, Hefei 230009, China (e-mail: eric.mengwang@gmail.com).

Z.-J. Zha is with the School of Computing, National University of Singapore, 117417 Singapore.

J. Shen is with the School of Information Systems, Singapore Management University, 178902 Singapore.

X. Li is with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Beijing 100864, China.

X. Wu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Department of Computer Science, University of Vermont, Burlington, VT 05405 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2202676

[1] Available at http://www.flickr.com.

[2] Available at http://www.youtube.com.

[3] Available at http://labs.yahoo.com/ksc/Multimedia.

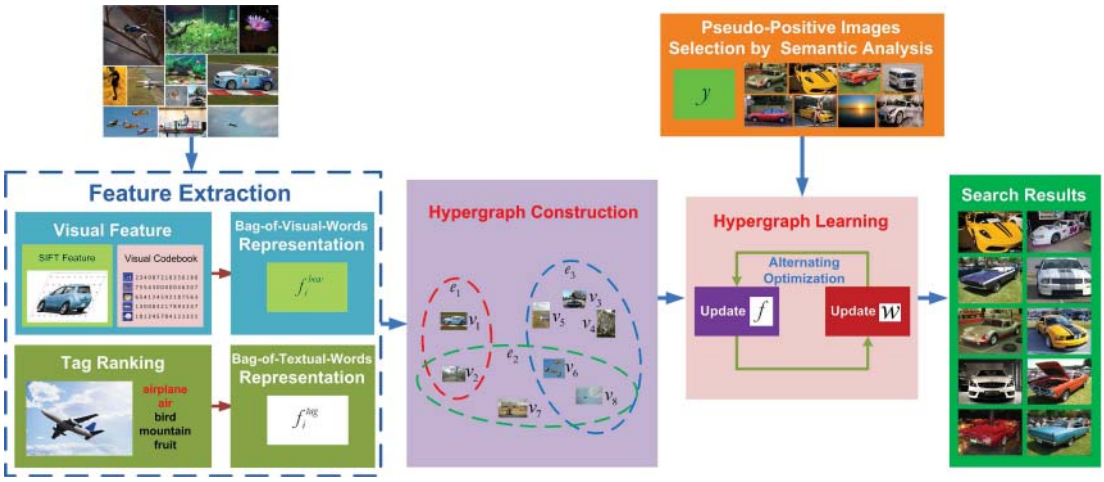

Fig. 1. Schematic illustration of the proposed visual-textual joint relevance learning approach.

relevance. Therefore, separately or sequentially using the two information sources is suboptimal for social image search.

In this paper, we propose a hypergraph-based approach to simultaneously utilize visual information and tags for image relevance learning. The scheme of our proposed approach is illustrated in Fig. 1. In the proposed method, each social image is represented by bag-of-textual-words and bag-of-visual-words features, which are generated from the tags and the visual content of the image, respectively. A hypergraph is constructed, in which the vertices denote the social images for ranking, and each visual word or tag generates a hyperedge. In such a hypergraph learning scheme, both the visual content and the tag information are taken into consideration at the same time. Different from the method [18] by using the traditional hypergraph learning approaches that adopts fixed hyperedge weights, we further learn the weights which indicate the importance of different visual words and tags. In this way, the effects of the informative visual words and tags can be enhanced. In the learning process, we first identify a set of pseudo relevant samples based on tags. Then, we calculate the relevance scores of images by iteratively updating them and the weights of hyperedges. We conduct experiments on a real-world dataset from Flickr and experimental results demonstrate the effectiveness of the approach.

The rest of the paper is organized as follows. Section II briefly reviews related work on social image search and hypergraph learning. Section III introduces the hypergraph learning. We introduce the visual-textual joint relevance learning algorithm in Section IV. Experimental results on the Flickr data set are provided in Section V to justify the effectiveness of the proposed method. Finally, we conclude the paper in Section VI.

II. Related Work

In this section, we briefly introduce the related work on social image search and hypergraph learning.

A. Social Image Search

Different from content-based image retrieval [6], [19]–[23], tag-based social image search mainly focuses on semantic queries [16]. Extensive research efforts [14], [24] have been dedicated to social image search in recent years [13]. Several methods have been proposed for tag refinement [25] or tag relevance learning [26], [27]. However, most existing methods use visual and tag information separately [16] or sequentially [17], [28]. We summarize these two schemes in Fig. 2, where separated methods can be further divided into textual content-only methods and visual content-only methods.

1) Separated Methods: In separated methods, only the textual content or the visual content is employed for tag analysis. Xu et al. [29] propose a tag refinement algorithm from the topic modeling point of view. A new graphical model named as regularized latent Dirichlet allocation (rLDA) is presented to jointly model the tag similarity and tag relevance. Sun and Bhowmick [30] propose a method to calculate the tag clarity score using the query language model and the collection language model. Liu et al. [28] propose a tag ranking approach, which is able to rank the tags that are associated with an image according to their relevance levels. Li et al. [26] introduce an approach that learns the relevance scores of tags by a neighborhood voting method. Given an image and an associated tag, the relevance score is learned by accumulating the votes from the visual neighbors of the image. They then further extend the work to multiple visual spaces [24]. They learn the relevance scores of tags and rank them by neighborhood voting in different feature spaces, and the results are aggregated with a score fusion or rank fusion method. Zhu et al. [25] propose a matrix decomposition method. The first component of their approach estimates the relevance scores of images.

2) Sequential Methods: In sequential methods, the visual content and the tags are sequentially employed for image search, in which the visual content and the tags are employed respectively. Most of existing methods first perform textual content-based analysis, and then the visual content is employed in the second image search stages. Liu et al. [17] propose a relevance-based ranking method for social image search, which first learns relevance scores based on the tags of images and then refines the relevance scores by exploring the visual content of images. Chen et al. [14] propose a tag

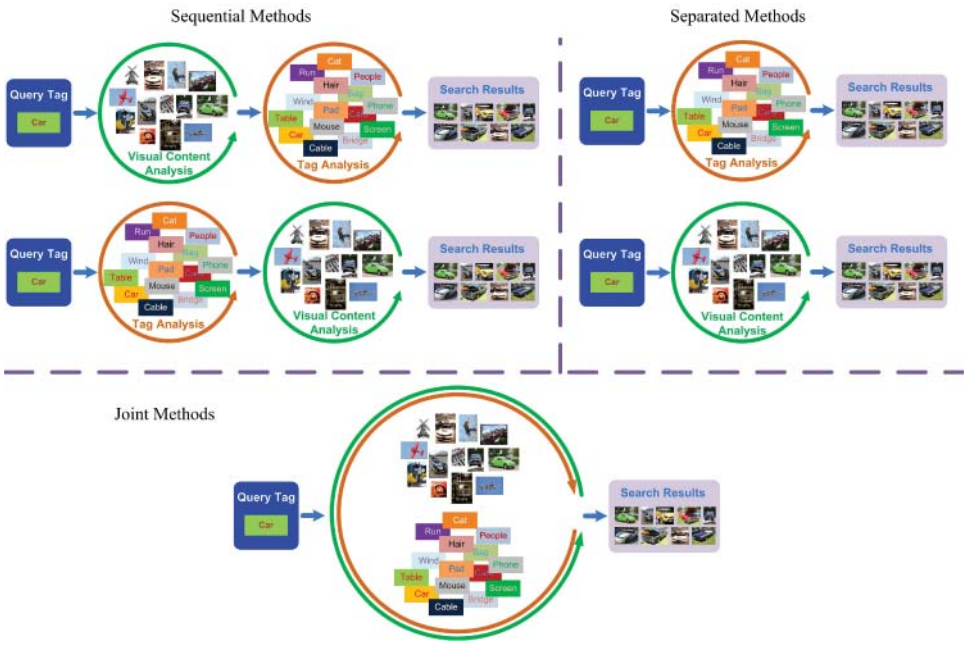

Fig. 2. Illustration of different social image search methods.

refinement method. For each tag, they first train a Support Vector Machine (SVM) classifier with the loosely labeled positive and negative samples. The classifiers are used to estimate the initial relevance scores of tags. They then further refine the scores with a graph-based method that simultaneously considers the similarity of photos and the semantic correlation of tags. Fan et al. [31] group images with a target tag into clusters. They regard a cluster as a unit, and first the initial relevance scores of the clusters are estimated and then the scores are refined via a random walk process. Liu et al. [32] adopt a three-step approach. The first step is filtering out tags that are intrinsically content-unrelated based on the ontology in WordNet. They then refine the tags based on the consistency of visual similarity and semantic similarity of images. In [33], a diverse relevance ranking scheme is proposed to re-rank social images by exploring the contents of images and their associated tags.

In separated methods, only the visual content or the tags are used for image search, in which the useful information is missing. In sequential methods, the correlation among visual content and tags are separated. Different from these two types of methods, we propose a joint method which integrates both the visual content and the tags in a unified hypergraph learning scheme such that they can be simultaneously utilized.

In addition, our approach can automatically modulate the effects of visual words and tags such that the influences of noisy textual and visual features can be reduced, and this makes our approach more robust than the existing methods.

B. Hypergraph Learning

Hypergraph has been employed in many data mining and information retrieval tasks, such as image retrieval and object recognition [34]–[36], for their effectiveness for higher-order sample relationship modeling. Zhou et al. [34] propose a general hypergraph framework and apply it to clustering, classification and embedding tasks. Zass et al. [37] propose a probabilistic hypergraph matching approach to match two sets of features. For image retrieval, Huang et al. propose [38] a transductive learning framework, where a hypergraph is constructed and each vertex in the hypergraph denotes one image. In the application of Computer-Aided Design (CAD), Wong et al. propose [39] a hypergraph-based 3D object description method, in which the vertices denote the surface patches of an object in the CAD system and the hyperedges represent the connection of the pair of boundary segments. For object recognition, Xia et al. propose [40] a class-specific hypergraph to explore both the local Scale-Invariant Feature Transform (SIFT) and the global geometric constraints for object recognition, in which the vertices of the constructed hypergraph represent the images that belong to the object category, and the selected SIFT points are employed as the feature of these vertices. This method is further extended in [41] to learn large-scale class-specific hypergraph model for 3D object recognition. Bu et al. [27] propose a unified hypergraph learning approach for music recommendation. In this method, the multi-type objects and relations in social networks (or virtual communities) interested in music are modeled by the hypergraph structure. The learning task on the constructed hypergraph is employed to measure the relationship among music tracks for music recommendation. Liu et al. [42] propose a transductive learning framework for image retrieval. In this method, each image is represented by a vertex in the constructed hypergraph, and the visual clustering results are employed to construct the hyperedges. A softer hypergraph learning procedure is introduced to rank images according the relevance levels of images.

These works have demonstrated the effectiveness of hypergraph structure in capturing higher-order relationship. In our proposed method, hypergraph is employed to model the

TABLE I
NOTATIONS AND DEFINITIONS

Notation	Definition
$\mathcal{X} = (x_1, x_2, \ldots, x_n)$	$\mathcal{X}$ indicates the image set, and x_i indicates the i-th image.
f_i^{bow}	The $n_c \times 1$ bag-of-visual-words feature vector for x_i.
f_i^{tag}	The n_t bag-of-text-words feature vector for x_i.
n_c	The size of the employed visual codebook.
n_t	The number of the employed tags.
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$	$\mathcal{G}$ indicates a hypergraph, and $\mathcal{V}$, $\mathcal{E}$ and w indicate the set of vertices, the set of edges, and the weights of hyperweights, respectively.
n	The number of images in hypergraph learning.
$\mathcal{V}$	The set of n vertices of the hypergraph.
$\mathcal{E}$	The set of edges of the hypergraph that contains n_e elements, where n_e is the number of edges.
$w = [w_1, w_2, \ldots, w_{n_e}]$	The $n_e \times 1$ weight vector of the hyperedges in the hypergraph.
$\delta(e)$	The degree of edge e.
$\mathbf{D}_v$	The $n \times n$ diagonal matrix of the vertex degrees.
$\mathbf{D}_e$	The $n_e \times n_e$ diagonal matrix of the edge degrees.
$\mathbf{H}_i$	The $n \times n_e$ incidence matrix for i-th hypergraph.
K	The number of the selected pseudo-relevant images.
y	The $n \times 1$ label vector for hypergraph learning. The elements of the pseudo-relevant images are set to 1, and the others are 0.
f	The $n \times 1$ to-be-learned relevance score vector.

relationship among social images and both the visual content and the tags of these images can be investigated. We also propose a method to automatically learn the weights of hyperedges.

III. Brief Introduction of Hypergraph Analysis

Before presenting our approach, we first briefly introduce the hypergraph learning theory.

In a simple graph, samples are represented by vertices and an edge links the two related vertices. Learning tasks can be performed on a simple graph. For instance, assuming that samples are represented by feature vectors in a feature space, an undirected graph can be constructed by using their pairwise distances, and graph-based semi-supervised learning approaches can be performed [34], [43] on this graph to categorize objects. It is noted that this simple graph cannot reflect higher-order information. Compared with the edge of a simple graph, a hyperedge in a hypergraph is able to link more than two vertices. For clarity, we first illustrate several important notations and their definitions throughout the paper in Table I.

A hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ is composed by a vertex set $\mathcal{V}$, an edge set $\mathcal{E}$, and the weights of the edges w. Each edge e is given a weight $w(e)$. The hypergraph $\mathcal{G}$ can be denoted by a $|\mathcal{V}| \times |\mathcal{E}|$ incidence matrix $\mathbf{H}$ with entries defined as:

$$h(v, e) = \begin{cases} 1 & if\ v \in e \\ 0 & if\ v \notin e. \end{cases} \tag{1}$$

For a vertex $v \in \mathcal{V}$, its vertex degree can be estimated by:

$$d(v) = \sum_{e \in \mathcal{E}} \omega(e)\, h(v, e). \tag{2}$$

For a hyperedge $e \in \mathcal{E}$, its hyperedge degree can be estimated by:

$$\delta(e) = \sum_{v \in \mathcal{V}} h(v, e). \tag{3}$$

Denote by $\mathbf{D}_v$ and $\mathbf{D}_e$ the diagonal matrices of the vertex degrees and the hyperedge degrees, respectively. Let W denote the diagonal matrix of the hyperedge weights

$$\mathbf{W}(i, j) = \begin{cases} w(i) & if\ i = j \\ 0 & otherwise. \end{cases} \tag{4}$$

In a hypergraph, many machine learning tasks can be performed, i.e. clustering, classification, and ranking. The binary classification is taken as an example here. A Normalized Laplacian method [34] is formulated as a regularization framework:

$$\arg\min_f \left\{ \lambda R_{emp}(f) + \Omega(f) \right\} \tag{5}$$

where f is the to-be-learned classification function, $\Omega(f)$ is a regularizer on the hypergraph, $R_{emp}(f)$ is empirical loss, and $\lambda > 0$ is a weighting parameter. The regularizer on the hypergraph is defined as:

$$\Omega(f) = \frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{u, v \in \mathcal{V}} \frac{w(e)\, h(u, e)\, h(v, e)}{\delta(e)} \times \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2. \tag{6}$$

Let $\Theta = \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-\frac{1}{2}}$, and $\Delta = \mathbf{I} - \Theta$. Here the normalized cost function can be rewritten as:

$$\Omega(f) = f^T \Delta f. \tag{7}$$

Algorithm 1 The Proposed Visual-Textual Joint Relevance Learning Method for Social Image Search.

Input: The image set for re-ranking $\mathcal{X} = (x_1, x_2, \ldots, x_n)$.
Output: The relevance score vector f for image re-ranking
Step 1. Hypergraph Construction
1. Regard each social image in the social image set $\mathcal{X} = (x_1, x_2, \ldots, x_n)$ as a vertex in the hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$.
2. Generate a bag-of-visual-words description f_i^{bow} for each image x_i, where $f_i^{bow}(k, 1) = 1$ indicates that x_i contains the k-th visual word.
3. Construct hyperedges by using f_i^{bow}, where the images sharing the same visual words are connected by one hyperedge. There are n_c visual hyperedges in total.
4. For each image, the tags are ranked by [28] and only top $\min(n_l, n_i)$ tags are left for further processing. Here n_i is the number of tags in x_i, and n_l is set as 10 in our experiments.
5. Let n^* be the total number of tags left in the database and the $\min(n_i, n^*)$ tags with the highest TF-IDF value are left for further hyperedge construction.
6. Generate a bag-of-textual-words description f_i^{tag} for each image x_i, where $f_i^{tag}(k, 1) = 1$ indicates that x_i contains the k-th selected tag.
7. Construct hyperedges by using f_i^{tag}, where the images sharing the same textual words are connected by one hyperedge. There are n_t textual hyperedges in total.
8. Generate the incidence matrix $\mathbf{H}_i$, the diagonal matrices of the vertex degrees and the hyperedge degrees $\mathbf{D}_v$ and $\mathbf{D}_e$, the initial weights of all hyperedges w, respectively.
Step 2. Pseudo-Relevant Sample Selection
The Flickr Distance is employed to estimate the semantic relevance of an image x_i to the query tag t_q, and the top K results are selected as the pseudo-relevant images.
Step 3. Relevance Learning on Hypergraph
Conduct semi-supervised learning on the hypergraph structure. Iteratively learn the to-be-learned relevance score vector f and the weights for hyperedge w.

Here Δ is a positive semi-definite matrix called hypergraph Laplacian.

IV. Visual-Textual Joint Relevance Learning

In this section, we introduce the proposed hypergraph-based visual-textual joint relevance learning approach by using both the visual content and the textual information. We first introduce the hypergraph construction process, and then provide the formulation of our proposed hypergraph learning approach for social image search. Finally, we detail the pseudo-relevant sample selection method. The proposed visual-textual joint relevance learning method is shown in Algorithm 1.

A. Hypergraph Construction

We regard each social image in the social image set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ as a vertex in the hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$. Let n indicate the total number of images in $\mathcal{X}$, and thus the generated hypergraph has n vertices. To create the hyperedges of the constructed hypergraph, two types of features are first selected for these social images, i.e., the visual content and the textual information.

Concerning the visual content of each social image, the bag-of-visual-words representation [44] is employed for image description as it has shown its superiority in many image retrieval tasks [45]–[49]. To generate the bag-of-visual-words representation, a dense set of uniformly distributed points are first identified for each social image, and the local SIFT [50] descriptors on these points are extracted. We then train a visual vocabulary with these data points. Let n_c indicate the size of the bag-of-visual-words codebook. With the codebook, each extracted local SIFT feature is encoded into a visual code by the nearest neighbor method. Each image x_i is represented by an $n_c \times 1$ feature vector f_i^{bow}, where $f_i^{bow}(k, 1) = 1$ indicates that x_i contains at least one data point belonging to the k-th visual code.

For the tags of social images, we adopt a bag-of-textual-words representation. In the hyperedge construction procedure, for each image x_i, the associated tags are first ranked by [28]. In [28], the initial relevance scores for these tags are first estimated based on probability density estimation, and then a random walk over a tag similarity graph is performed to refine the relevance scores. Only top $\min(n_l, n_i)$ tags are left for further processing, where n_i is the number of tags in x_i. This procedure is to keep top n_l tags for further processing. When the order of the tag is below the top n_l, it is interpreted as having small relevance to the image. When the tags for the images are less than n_l, we keep all tags for this image. Here let n^* be the total number of tags left in the database. We only employ the $\min(n_t, n^*)$ tags with the highest term frequency Ċinverse document frequency (TF-IDF) values for further hyperedge construction, where n_t is the expected number of tags left for hyperedge construction and TF-IDF is a numerical statistic to reflect the importance of a word/tag is to a document/image.

With these n_t tags, we generate the bag-of-textual-words representation for each image. Each image x_i is represented by an $n_t \times 1$ feature vector f_i^{tag}, where $f_i^{tag}(k, 1) = 1$ indicates that x_i contains the k-th selected tag.

Now we have two feature vectors f_i^{bow} and f_i^{tag} for each image. Therefore, two types of hyperedges, namely the visual content-based and the the tag-based hyperedges, are constructed in the hypergraph. For visual content-based hyperedges, each visual word generates a hyperedge, by which the social images that contain the same visual word, i.e. $f_i^{bow}(k, 1) = 1$, are connected. Therefore, there are n_c visual content-based hyperedges in total. Analogously, each tag is used to generate a hyperedge, and there are n_t tag-based hyperedges. Concerning both the visual content and the tags, there are $n_c + n_t$ hyperedges.

Fig. 3 provides an example to show how visual and textual hyperedges are constructed. In the example, there are three hyperedges constructed by visual words or tags respectively.

In the constructed hypergraph, let $\mathbf{D}_v$ and $\mathbf{D}_e$ denote the diagonal matrices of the vertex degrees and the hyperedge degrees, respectively. The incidence matrix $\mathbf{H}$ is constructed using Eq. (1). The weights of all hyperedges are initialized with $w_i = \frac{1}{n_e}$.

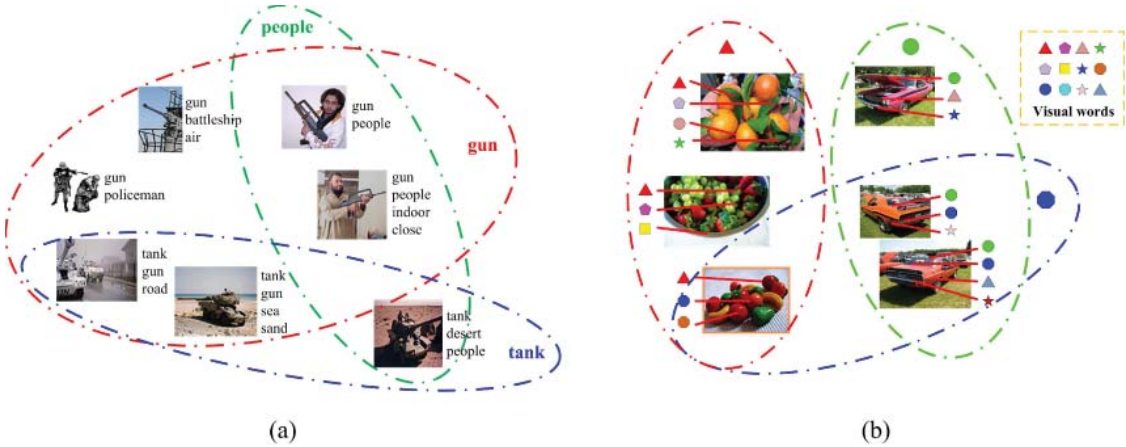

Fig. 3. Examples of hyperedge construction. (a) Example of textual hyperedge construction, where three hyperedges are generated by tags "people," "gun," and "tank." (b) Example of visual hyperedge construction, where three hyperedges are generated by three visual words.

Fig. 4 shows an example for the connection between two images by using visual information and the textual information respectively.

According to the above process, we can observe the rationality of the proposed hypergraph based approach: two social images tend to be connected with more hyperedges if they share a lot of tags or visual words.

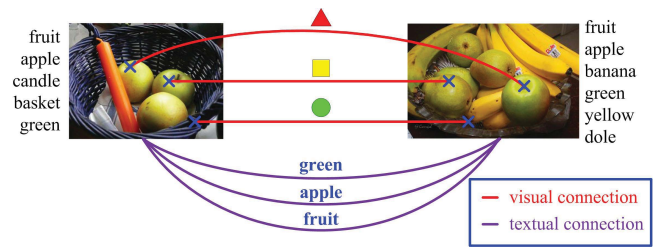

Fig. 4. Example of the connection between two images.

B. Social Image Relevance Learning Formulation on Hypergraph

In the constructed hypergraph structure, each image is denoted by a vertex, and the social image search task is regarded as a binary classification problem. We aim to measure the relevance scores among all vertices in the hypergraph, and the transductive inference is also formulated as a regularization framework $\arg\min_{f,\omega}\left\{\Omega(f)+\lambda R_{emp}(f)+\mu\Psi(\omega)\right\}$. Here the regularizer term $\Omega(f)$ on the hypergraph structure applies the formation of $\Omega(f)$ in Eq. 6 in the social image search task. $\Omega(f)$ indicates that highly related vertices should have close label results, which is defined as:

$$\frac{1}{2}\sum_{i=1}^{n_e}\sum_{u,v\in V}\frac{w_i h(u,e_i)\,h(v,e_i)}{\delta(e_i)}\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2 \quad (8)$$

where the vector f is the to-be-learned relevance score vector. Eq. (8) further turns into:

$$\begin{aligned}\Omega(f) &= \sum_{i=1}^{n_e}\sum_{u,v\in V}\frac{w_i h(u,e_i)\,h(v,e_i)}{\delta(e_i)} \times\left(\frac{f^2(u)}{d(u)}-\frac{f(u)\,f(v)}{\sqrt{d(u)\,d(v)}}\right)\\ &= \sum_{u\in V}f^2(u)\sum_{i=1}^{n_e}\frac{w_i h(u,e_i)}{d(u)}\sum_{v\in V}\frac{h(v,e_i)}{\delta(e_i)} -\sum_{i=1}^{n_e}\sum_{u,v\in V}\frac{f(u)h(u,e_i)\,w_i h(v,e_i)\,f(v)}{\sqrt{d(u)\,d(v)}\delta(e_i)}\\ &= f^T(\mathbf{I}-\Theta)\,f \end{aligned} \quad (9)$$

where $\Theta=\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_v^{-\frac{1}{2}}$. Let $\Delta=\mathbf{I}-\Theta$, where Δ is the normalized hypergraph Laplacian. Thus we rewrite the regularizer $\Omega(f)$ as:

$$\Omega(f)=f^T\Delta f. \quad (10)$$

The loss term is defined as:

$$R_{emp}(f)=\|f-y\|^2=\sum_{u\in V}(f(u)-y(u))^2 \quad (11)$$

where y is an $n\times 1$ initial label vector. This empirical loss $R_{emp}(f)$ guarantees that the new generated labeling results are not far away from the initial label information. To generate y, we usually need a set of relevant samples. A straightforward approach is to regard all the images that have the query tag as relevant, but the noise will cause performance degradation. To reduce the noise, a set of samples are selected which are not associated with the query tag but also have high relevance probabilities. We call these samples pseudo-relevant, and this strategy has been widely used in re-ranking [51]–[53]. The detailed method for pseudo-relevant sample selection will be introduced in next subsection. The corresponding elements of these images are set to 1, and other elements are 0.

In the constructed hypergraph, all the hyperedges are initialized with an identical weight. However, the hyperedges are with different effects as there exists a lot of uninformative visual words and tags for a given query. Therefore, performing a weighting or selection on the hyperedges will be helpful. Here we integrate the learning of the hyperedge weights into the formulation.

Let $\{e_1, e_2, \ldots, e_{n_e}\}$ denote the n_e hyperedges, and $w = [w_1, w_2, \ldots, w_{n_e}]$ be the $n_e \times 1$ weight vector of the hyperedges in the hypergraph, where $\sum_{i=1}^{n_e} \omega_i = 1$. We add a 2-norm regularizer on w and then simultaneously optimize w and f. The formulation thus becomes:

$$\arg\min_{f,\omega} \Phi(f) = \arg\min_{f} \left\{ f^T \Delta f + \lambda \| f - y \|^2 + \mu \sum_{i=1}^{n_e} \omega_i{}^2 \right\} \quad \text{s.t.} \sum_{\mathrm{i}=1}^{\mathrm{n_e}} \omega_{\mathrm{i}} = 1 \tag{12}$$

where λ and μ are two positive weighting parameters.

Under this formulation, we aim to seek optimal results which can minimize the cost function including the loss cost, the hypergraph regularizer and the hypergraph weight regularizer.

C. Solving the Optimization Problem

Here we adopt an alternating optimization strategy to solve the above problem. In this strategy, for the to-be-learned two variables w and f, we fix one and optimize the other one each time.

Using this iterative optimization method, the optimal f and w values are obtained. With the hypergraph edge weight w, the hypergraph structure is optimized for the query. Under this hypergraph structure, f is the optimal to-be-learned relevance score vector for social image search.

We introduce the alternating optimization strategy as follows.

First, we fix w and optimize f. The problem becomes

$$\arg\min_{f} \Phi(f) = \arg\min_{f} \left\{ f^T \Delta f + \lambda \| f - y \|^2 \right\}. \tag{13}$$

From the above equation, we can derive that:

$$\begin{aligned} f &= \left(I + \frac{1}{\lambda} \Delta \right)^{-1} y \\ &= \left(\mathbf{I} + \frac{1}{\lambda} (\mathbf{I} - \Theta) \right)^{-1} y \\ &= \frac{\lambda + 1}{\lambda} \left(\mathbf{I} - \frac{1}{\lambda + 1} \Theta \right)^{-1} y. \end{aligned} \tag{14}$$

Let $\xi = \frac{1}{\lambda+1}$, we can derive:

$$f = \frac{1}{1 - \xi} (\mathbf{I} - \xi \Theta)^{-1} y. \tag{15}$$

Now we have the optimal f when w is fixed.

Next, we fix f and optimize w, and the equation becomes:

$$\arg\min_{\omega} \Phi(f) = \arg\min_{f} \left\{ f^T \Delta f + \mu \sum_{i=1}^{n_e} \omega_i{}^2 \right\} \quad \text{s.t.} \sum_{i=1}^{n_e} w_i = 1, \mu > 0. \tag{16}$$

Here the Lagrangian is employed and the optimization problem turns to:

$$\begin{aligned} &\arg\min_{\omega,\eta} f^T \Delta f + \mu \sum_{i=1}^{n_e} \omega_i{}^2 + \eta \left(\sum_{i=1}^{n_e} \omega_i - 1 \right) \\ &= \min_{\omega,\eta} f^T \left(\mathbf{I} - \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-\frac{1}{2}} \right) f \\ &\quad + \mu \sum_{i=1}^{n_e} \omega_i{}^2 + \eta \left(\sum_{i=1}^{n_e} \omega_i - 1 \right). \end{aligned} \tag{17}$$

Let $\Gamma = \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H}$, it can be derived that

$$\eta = \frac{f^T \Gamma f - 2\mu}{n_e} \tag{18}$$

and

$$\omega_i = \frac{1}{n_e} - \frac{f^T \Gamma \mathbf{D}_e^{-1} \Gamma^T f}{2 n_e \mu} + \frac{f^T \Gamma_i D_e^{-1}(i,i) \Gamma_i{}^T f}{2\mu} \tag{19}$$

where Γ_i is the i-th column of Γ.

Since each of the above steps decreases the objective function $\Phi(f)$ which has a lower bound 0, the convergence of the alternating optimization is guaranteed.

D. Probabilistic Explanation of the Proposed Method

Here we also provide a probabilistic explanation of our approach. From probabilistic perspective, we can derive the optimal f and w with the maximum posterior probability given the samples X and the label vector y

$$\{f, w\}^* = \arg\max p(f, w | X, y). \tag{20}$$

Following Bayes rule, the above equation can turn to

$$\arg\max p(f | X, w)\, p(y | X, f, w)\, p(w). \tag{21}$$

We let

$$p(f | X, w) = \frac{1}{Z_1} \exp\left(-f^T \Delta f \right) \tag{22}$$

$$\begin{aligned} p(y | X, f, w) &= p(y | X, f) \\ &= \frac{1}{Z_2} \exp\left(-\frac{\| y - f \|^2}{1/\lambda} \right) \end{aligned} \tag{23}$$

and

$$p(w) = \frac{1}{Z_3} \exp\left(-\frac{\left\| w - \frac{1}{n_e} \mathbf{1} \right\|^2}{1/\mu} \right) \tag{24}$$

where Z_1, Z_2 and Z_3 are normalizing constants to keep the integral of the probability function to be 1, and $\mathbf{1}$ is a vector that has all the elements to be 1. By adding a constraint $\sum_{i=1}^{n_e} w_i = 1$, we can see that Eq. (20) and Eq. (12) are equivalent. The first two terms, i.e., $p(f | X, w)$ and $p(y | X, f, w)$, actually reflect the two assumptions in graph and hypergraph learning, i.e., the function should be smooth on the graph/hypergraph and it should not change too much from the initial labels. In comparison with the conventional hypergraph learning [34], we have added the term $p(w)$. Instead of fixing hyperedge weights, we assume that they have a Gaussian distribution, such that the weights w can be learned together with f.

Fig. 5. Several example images and their associated tags used in our experiments.

E. Pseudo-Relevant Sample Selection

Note that we have used a set of pseudo-relevant samples in the hypergraph learning algorithm. In this part, we introduce the pseudo-relevant sample selection method. We simply estimate the semantic relevance of an image x_i to the query tag t_q as the average semantic similarity between t_q and all tags of x_i

$$s\left(x_i, t_q\right) = \frac{1}{n_i} \sum_{t \in \mathcal{T}_i} s_{tag}\left(t_q, t\right) \tag{25}$$

where $\mathcal{T}_i$ is the tag set of x_i. With this semantic relevance, all the social images that are associated with the tag are ranked in descending order, and the top K results are selected as the pseudo-relevant images.

The semantic similarity between two tag t_1 and t_2 is calculated by:

$$s_{tag}\left(t_1, t_2\right) = \exp\left(-FD\left(t_1, t_2\right)\right), \tag{26}$$

where $FD\left(t_1, t_2\right)$ is the Flickr Distance [54] between t_1 and t_2.

Flickr Distance is selected due to its effectiveness on the distance measure based on text information as introduced in [54]. In Flickr distance, a group of images are first obtained from Flickr for each tag, and a latent topic based visual language model is built to model the visual characteristic of the tag. The Flickr distance is calculated by using Jensen-Shannon divergence between the two visual language models.

V. Experiments

A. Experimental Settings

We conduct experiments on a Flickr dataset, which has been used in [33]. The dataset is collected based on a diverse set of popular tags, including *airshow*, *apple*, *aquarium*, *basin*, *beach*, *bird*, *bmw*, *car*, *chicken*, *chopper*, *cow*, *decoration*, *dolphin*, *eagle*, *fighter*, *flame*, *flower*, *forsest*, *fruit*, *furniture*, *glacier*, *hairstyle*, *hockey*, *horse*, *jaguar*, *jellyfish*, *lion*, *matrix*, *motorcyle*, *olymplics*, *owl*, *palace*, *panda*, *rabbit*, *rainbow*, *rice*, *sailboat*, *seagull*, *shark*, *snowman*, *spider*, *sport*, *starfish*, *swimmer*, *telephone*, *triumphal*, *turtle*, *watch*, *waterfall*, *weapon*, *wildlife*, and *wolf*. These tags are employed to search images and the top 2000 searching results for each query tag are collected with their associated information, i.e. tags, uploading time, and user identifier. There are 104,000 images and 83,999 unique tags in total. Fig. 5 shows some example images and the associated tags. Each image is labeled with three relevance levels with respect to the corresponding query: very relevant, relevant and irrelevant. We use scores 2, 1, 0 to indicate the three relevance levels, respectively. To enlarge the testing dataset, we further employ the NUS-WIDE dataset [55] which is merged with the Flickr dataset. The new dataset includes 370K+ images in total. In the proposed method, simple tag-based method is employed to select top 2000 images for further processing. In our experimental setups, we have taken the tags with multiple meanings into consideration. For the tags with more than one meaning, the images corresponding to different meanings are all regarded as relevant. This means that when one tag has two meanings, if the image corresponds to either one of these two meanings, this image is relevant to the tag. In our experiments, there are two queries with multiple meanings, i.e., *apple* and *jaguar*. *Apple* could refer to fruit, mobile phone or computer, and *jaguar* could refer to animal or car.

We compare the following methods:

1) Graph-based semi-supervised learning [56]. Graph-based semi-supervised learning have been widely applied in multimedia applications, such as image/object retrieval [57]–[59] and video annotation [60], [61]. Here the distance between two social images is calculated based on the bag-of-visual-words and the bag-of-textual-words representation, and the graph-based semi-supervised learning method in [56] is employed. We also adopt the pseudo-sample selection method in Section IV-E. The method is denoted as "Graph."
2) Sequential social image relevance learning, which is proposed in [32]. In this method, initial relevance scores are estimated based on tags, and the scores are then refined with a graph-based learning based on images' visual content. The method is denoted as "Sequential."
3) Tag ranking [28]. In this method, initial relevance scores of the tags are estimated, and a random walk process on a tag graph is conducted to refine the relevance scores of tags in each image. Given a query tag, the relevance score of one image is based on the position of the query tag in the tag ranking list. The method is denoted as "Tag Ranking."
4) Tag relevance combination [24]. In this method, tag relevance estimates are combined based on the largest entropy assumption. The method is denoted as "Uniform Tagger."
5) Hypergraph-based relevance learning. In this method, the hypergraph is employed to model the relationship among social images. But different from the proposed approach, all hyperedge weights are equally treated. The method is denoted as "HG."
6) Hypergraph-based relevance learning with hyperedge weight estimation, i.e., the proposed approach. The method is denoted as "HG-WE."
7) The proposed learning method with only visual information. That means, we only employ visual information in the hypergraph learning process. The method is denoted as "HG-WE (Visual)."
8) The proposed learning method with only textual information. That means, we only employ textual information

TABLE II

THE $NDCG@20$ RESULTS OF DIFFERENT METHODS. THE BEST RESULT IN EACH ROW IS MARKED IN BOLD

Query	Graph	Seq	TagRanking	UniformTagger	HL	HL+WE	HG-WE (Visual)	HG-WE (Tag)
airshow	0.5854	0.4193	0.4514	0.4266	0.5759	**0.7183**	0.5847	0.6869
apple	0.3520	0.2433	0.2405	0.2265	0.6975	**0.8128**	0.8100	0.7875
aquarium	0.5799	0.5640	0.5704	0.5740	0.8163	**0.9346**	0.9189	0.9134
basin	0.1638	0.2981	0.3390	0.3594	0.4911	0.6115	**0.6178**	0.5946
beach	0.6508	0.5986	0.6328	0.6258	0.8270	**1.0000**	0.9949	0.9869
bird	0.7504	0.8931	0.9641	0.8870	0.9576	**0.9653**	0.9375	0.9618
bmw	0.2721	0.5910	0.5995	0.5900	0.6244	**0.7265**	0.7048	0.6826
car	0.5570	0.6996	0.7391	0.6897	0.8095	0.7991	**0.9659**	0.7484
chicken	0.6853	0.3254	0.3349	0.5982	0.7609	**1.0000**	0.9321	0.9903
chopper	0.5265	0.4660	0.5488	0.5125	0.4038	0.6390	0.6076	**0.6490**
cow	0.7214	0.5253	0.5955	0.5915	0.9065	**1.0000**	**1.0000**	0.9982
decoration	0.5413	0.7960	0.8341	0.8323	0.9468	**1.0000**	**1.0000**	0.9827
dolphin	0.3662	0.3390	0.3442	0.3487	0.6412	0.7731	**0.7952**	0.7345
eagle	**0.7978**	0.5438	0.5724	0.5719	0.5750	0.7172	0.6325	0.6862
fighter	0.8560	0.4989	0.5465	0.4920	**1.0000**	**1.0000**	0.9427	0.9947
flame	0.7461	0.3419	0.3718	0.3484	0.3747	**0.7651**	0.6999	0.6898
flower	0.7372	0.7878	0.8678	0.6296	0.7288	**0.8888**	0.7478	0.8714
forsest	0.6424	0.5062	0.6051	0.5842	0.9490	**0.9951**	0.9551	0.9827
fruit	**0.8261**	0.5300	0.5924	0.5532	0.4642	0.7596	0.7223	0.7304
furniture	0.9245	0.4451	0.4782	0.4627	0.8255	**0.9397**	0.8991	0.9098
glacier	0.5604	**0.7975**	0.7821	0.7856	0.5747	0.6099	0.7749	0.5853
hairstyle	0.8988	0.7619	0.8519	0.8208	0.5630	**0.9346**	0.8866	0.9174
hockey	0.7928	0.9915	0.9878	0.7917	**1.0000**	**1.0000**	**1.0000**	0.9767
horse	0.7399	0.1676	0.2053	0.1557	**1.0000**	0.9865	0.9398	0.9807
jaguar	0.7540	0.7957	**0.8920**	0.8540	0.8346	0.7915	0.8478	0.7713
jellyfish	0.7691	0.6005	0.6825	0.6128	0.6842	**0.9406**	0.8819	0.8846
lion	0.4164	0.5787	0.6381	0.5923	0.6971	**0.9175**	0.7180	0.9116
matrix	0.8956	0.6430	0.6986	0.6942	0.9417	**1.0000**	0.9177	0.9888
motorcyle	0.7026	0.5984	0.6204	0.5971	0.8098	0.8242	**0.9689**	0.8386
olymplics	0.4946	0.7373	0.8384	0.6055	0.9088	0.9309	**0.9537**	0.8837
owl	0.5076	0.8351	0.9251	0.8410	0.7644	**1.0000**	0.5734	0.9826
palace	0.3718	0.3718	0.4376	0.4182	0.8608	**0.9094**	0.7483	0.8827
panda	0.5148	0.5117	0.5385	0.5044	0.8233	**1.0000**	0.9856	0.9735
rabbit	0.7650	0.6780	0.7096	0.6953	0.8609	0.8216	**0.9530**	0.7766
rainbow	0.4335	0.5587	0.6052	0.5642	0.8035	0.9093	**0.9406**	0.8717
rice	0.3233	0.5068	0.5045	0.5049	0.5656	0.9878	**0.9927**	0.9410
sailboat	0.4515	0.4621	0.4876	0.4843	0.7993	0.8548	**1.0000**	0.8720
seagull	0.4066	0.6150	0.6808	0.6326	0.5216	**0.7937**	0.7208	0.7540
shark	0.4544	0.5010	0.6305	0.4824	0.5791	**0.9759**	0.9448	0.9649
snowman	**1.0000**	0.3928	0.4936	0.7072	**1.0000**	**1.0000**	**1.0000**	0.9927
spider	0.2371	0.5147	0.5563	0.5709	0.6155	**0.8306**	0.7677	0.7681
sport	0.2579	0.6334	0.7203	0.6943	0.4828	0.8701	**0.9015**	0.8468
starfish	0.5102	0.6681	0.7040	0.6793	0.5814	**0.8806**	0.7633	0.8378
swimmer	0.9696	0.8213	0.8781	0.9723	0.9657	**1.0000**	0.9625	0.9913
telephone	0.2215	0.8481	0.8800	0.8739	0.7034	0.9204	**0.9325**	0.8664
triumphal	0.1616	0.5644	0.5999	0.5697	**0.8082**	0.7651	0.5365	0.7415
turtle	0.5482	0.2982	0.4316	0.3122	0.8069	0.9537	0.8291	**0.9762**
watch	0.6740	0.7249	0.7652	0.7448	0.6832	**0.9669**	0.9442	0.9379
waterfall	0.4892	0.5315	0.5929	0.6079	0.7623	**0.9249**	0.9088	0.8724
weapon	0.2351	0.6284	**0.7100**	0.5545	0.5253	0.6338	0.5337	0.6109
wildlife	0.5265	0.6624	0.7187	0.7067	0.8124	**0.8509**	0.7714	0.8411
wolf	0.4152	0.6322	0.6640	0.6332	0.8578	**1.0000**	0.9387	0.9823
mean	0.5727	0.5778	0.6281	0.5994	0.7418	**0.8814**	0.8463	0.8578

in the hypergraph learning process. The method is denoted as "HG-WE (Tag)."

For the "Graph" method, there is a weighting parameter and we set it following the setting of [56]. For the "Sequential" method, the parameter settings are the same with [32]. For the "Hypergraph", "Hypergraph-WE" and "Hypergraph" methods, the parameters λ and μ are empirically set to 100 and 0.001, respectively. We set the size of the tag and visual word dictionary to 1000, i.e., $n_c = n_t = 1000$. For pseudo-relevant sample selection, we simply set K to 100, i.e., we use 100 pseudo-relevant images. The parameter n_l is set as 10. We will also analyze the parameters later. For the above four methods, we randomly sample 5,000 images that are not associated with the query tag as negative examples in the learning process. The Normalized Discounted Cumulative Gain (NDCG) [62] is employed for performance evaluation.

B. Experimental Results and Discussion

Table II illustrates the $NDCG@20$ comparison of different methods. Here we illustrate not only the NDCG measurements of each query but also the average NDCG measurements of the 52 queries. From the results we have the following observations:

1) The "HG" method achieves better performance than "Graph," "Sequential," "Tag Ranking," and "Uniform Tagger". This indicates that the hypergraph learning is effective in social image modeling.
2) The proposed "HG-WE" approach achieves the best results for most queries. Among the seven methods, it also achieves the best average performance. This shows that our hyperedge weighting method can greatly improve the performance of hypergraph learning. Its superiority over "HG-WE (Visual)" and "HG-WE (Tag)" also demonstrates the effectiveness of combining visual and tag information.

Fig. 6 illustrates the average NDCG measurements at different depths. From the figure we can see that our approach consistently outperforms the other methods. Fig. 7 demonstrates the top 10 results obtained by different methods for an example query *weapon*. From the figure we can intuitively see the superiority of the proposed approach.

We further investigate the performance of the proposed method on query with multiple meanings. The proposed method is not limited by multiple meanings, and any image with one meaning from all different meanings is regarded as relevant. In our method, the pseudo-relevant sample selection procedure is not limited to any special meaning. There-

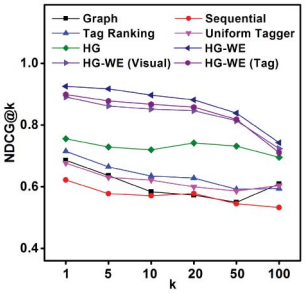

Fig. 6. Average $NDCG@k$ comparison of different methods, where k is the depth for NDCG.

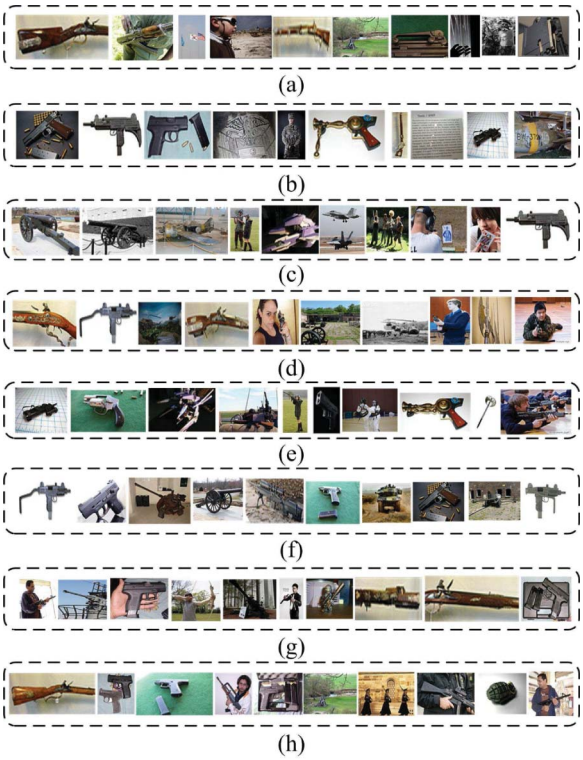

Fig. 7. Top results obtained by different methods for the query *weapon*. (a) Graph-based semi-supervised learning. (b) Sequential social image relevance learning. (c) Tag ranking. (d) Uniform tagger. (e) Hypergraph-based relevance learning. (f) Hypergraph based relevance learning with hyperedge weight estimation, i.e., the proposed method. (g) Proposed learning method with merely visual information. (h) Proposed learning method with merely tag information.

fore, the final ranking list can preserve images from all different meanings, and users can obtain searching results with different meanings. In our experiments, the query *apple* and the query *jaguar* are two queries with more than one meaning.

Fig. 8 demonstrates the top 10 results obtained by different methods for an example query *apple*. As shown in these results, the proposed method can return relevant results with different meanings, which demonstrates the superiority of the proposed approach.

To intuitively demonstrate the effects of hyperedge weight learning, here we illustrate several examples. We consider the queries *car* and *weapon*. Fig. 9(a) shows the dense sampling on the two example images. In Fig. 9, we only illustrate

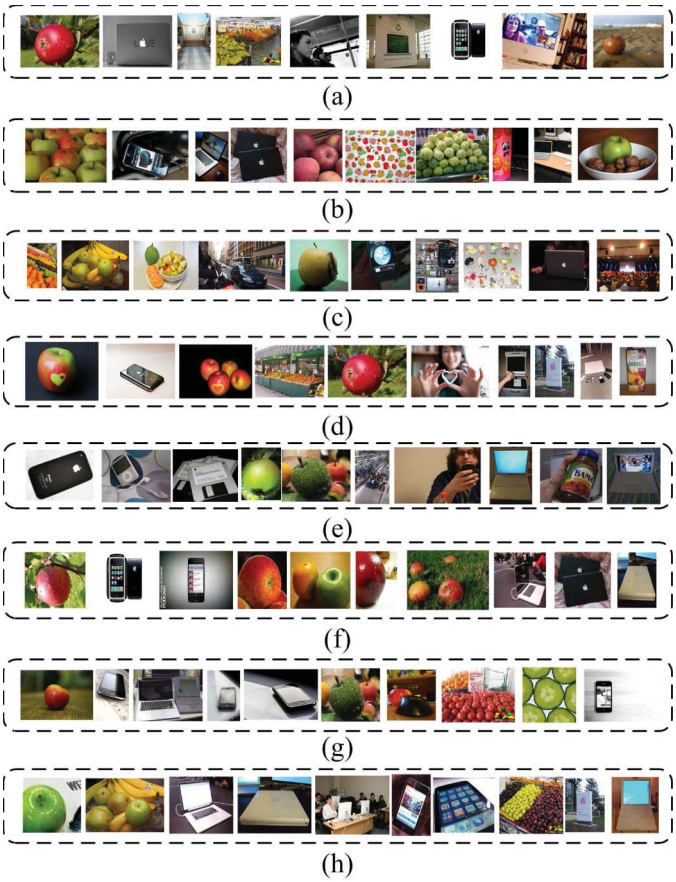

Fig. 8. Top results obtained by different methods for the query *apple*. (a) Graph-based semi-supervised learning. (b) Sequential social image relevance learning. (c) Tag ranking. (d) Uniform tagger. (e) Hypergraph-based relevance learning. (f) Hypergraph based relevance learning with hyperedge weight estimation, i.e., the proposed method. (g) Proposed learning method with merely visual information. (h) Proposed learning method with merely tag information.

the keypoints of the 100 visual words that have the highest weights. By comparing (a) and (b) in the figure, we can see that the weight learning is able to enhance the descriptive visual words for a given query. Fig. 10 shows the 10 tags that have the highest weights after the hypergraph learning process. We can also see that they are closely related to the queries.

C. On the Parameters λ and μ

From Eq. (12) we can see that there are two weighting parameters λ and μ in our formulation. They modulate the effects of the loss term $\|f - y\|^2$ and the regularizer term $\|w\|^2$ respectively. The parameter λ is widely used in graph or hypergraph learning algorithms, and its value determines the closeness of f and y. For parameter μ, if its value tends to be infinite, then the proposed algorithm will degenerate to Eq. (2), i.e., the conventional hypergraph learning algorithm in [34]. If the value of μ tends to be 0, then the optimal results is that only one weight is 1 and all others are 0. This extreme case means that there will be only one hyperedge weight used.

Fig. 11(a) and (b) demonstrate the average $NDCG@20$ performance curves with respect to the variation of λ and μ, respectively. In Fig. 11(a), we fix μ to be 0.001 and vary λ from 1×10^1 to 1×10^3. In Fig. 11(b), we fix λ to be 100

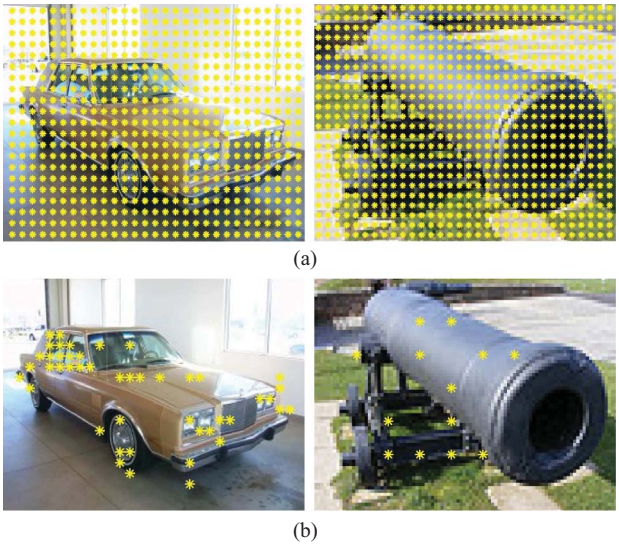

Fig. 9. (a) Densely sampled points on two example images. (b) Illustration of top 100 visual words with the highest weights after the hypergraph learning process.

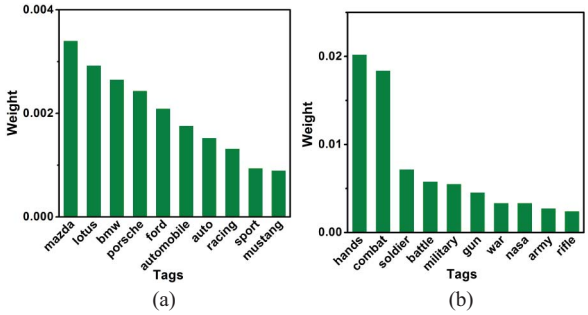

Fig. 10. Ten tags with the highest weights after the hypergraph learning process for the queries (a) *car* and (b) *weapon*.

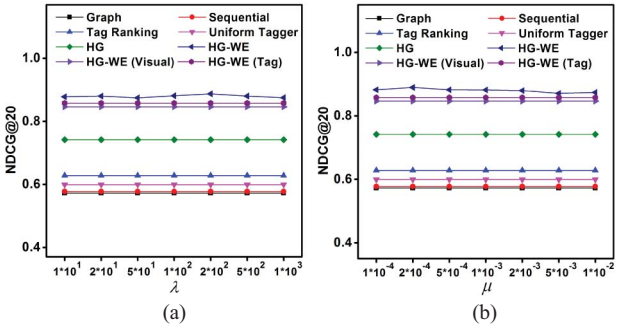

Fig. 11. Average NDCG@20 performance curves with respect to the variation of λ and μ. (a) μ is fixed as 0.001, and λ varies from 1×10^1 to 1×10^3. (b) λ is fixed as 100, and μ varies from 1×10^{-4} to 1×10^{-2}.

and vary μ from 1×10^{-4} to 1×10^{-2}. We also illustrate the NDCG results obtained by the other six methods in the figure for comparison. We can see that our approach is able to outperform the other methods when the two parameters vary in a wide range.

D. On the Size of Tag and Visual Word Dictionaries

We also provide the experimental results with different sizes of the tag and visual word dictionaries, i.e., n_c and n_t. The

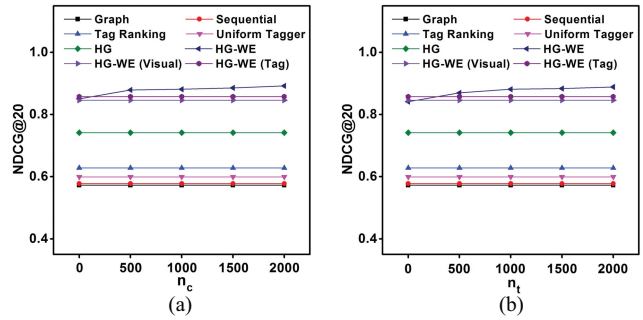

Fig. 12. Performance curves with respect to the variation of n_c and n_t. (a) $NDCG@20$ comparison of the proposed method with different n_c selection. (b) $NDCG@20$ comparison of the proposed method with different n_t selection.

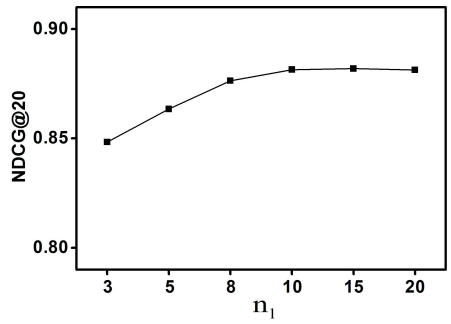

Fig. 13. $NDCG@20$ comparison of the proposed method with different n_l selection.

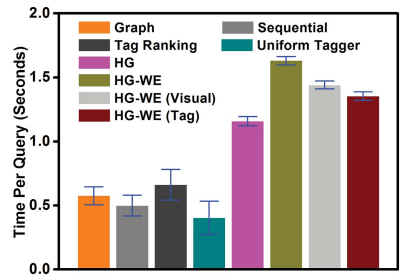

Fig. 14. Computational cost comparison.

sizes actually determine the number of hyperedges in the hypergraph learning algorithm.

Fig. 12(a) and (b) illustrate the $NDCG@20$ performance curves with respect to the variation of n_c and n_t, respectively. Here $n_c = 0$ and $n_t = 0$ indicate only using tags and visual content, respectively. We also illustrate the results of the other methods for comparison.

As shown in Fig. 12(a), when n_c is not 0, the proposed method achieves better results compared with all other methods. When n_c is 0, the proposed method is the same as HG-WE(Tag), and they achieved the same experiment results, which are also the best results in all methods. When n_c increases from 0 to 2000, the image search performance in terms of $NDCG@20$ becomes better, while the growth speed is slower when n_c is larger. As shown in Fig. 12(b), when n_t is not 0, the proposed method achieves better results compared with all other methods. When n_t is 0, the proposed method is the same as HG-WE(Visual), and they achieved the same

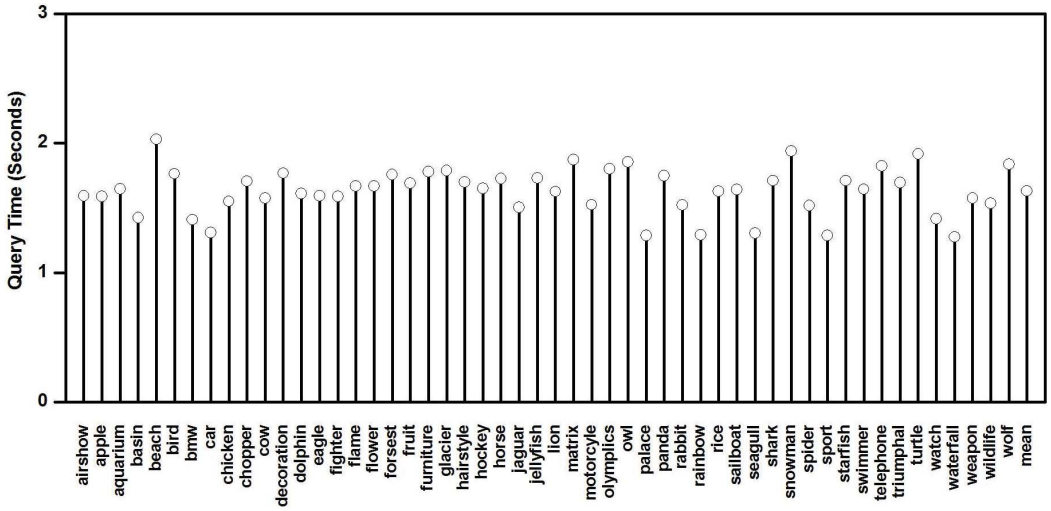

Fig. 15. Computational cost for each query tag of the proposed method.

experiment results, which are the second best results in all methods. When n_t increases from 0 to 2000, the image search performance in terms of $NDCG@20$ becomes better, while the growth speed is slower when n_t is larger.

E. On the Parameter n_l

The parameter n_l is employed to filter noise tags in the hypergraph construction procedure. Here we provide the experimental results with different n_l values. Fig. 13 illustrate the $NDCG@20$ performance curves with respect to the variation of n_l.

As shown in Fig. 13, when n_l is small, i.e., $n_l = 3$, the image search performance of the proposed method is the worst. With the increase of n_l, the image search performance in terms of $NDCG@20$ becomes better while the growth speed becomes slow. When n_l is larger than 10, the image search performance is relatively steady. The results can be explained as follows. When n_l is too small, only a few tags are kept for further processing, which may lead to information lost by removing most of the tags. With the increase of n_l, more tags are selected, which can employ more meaningful tags for further processing and improve the image search performance. When n_l is large enough, most of meaningful tags have been selected, and continuing to select more tags may not only involve useful tags but also bring in more noise tags, which could reduce the image search performance.

F. On the Running Time Comparison

We now demonstrate the time costs per query of different methods in Fig. 14. The computational cost for each query tag of the proposed method is provided in Fig. 15. But note that we have not taken the visual feature extraction step into consideration[4] The computational costs are recorded on a PC with Pentium 4 2.0GHz and 4G memory. We draw following conclusions.

[4]Visual feature extraction is a relatively time-consuming step. But many search engines host several pre-computed visual features for the indexed images in order to facilitate several services, and thus we can utilize these features to avoid the need of feature extraction process.

1) The HG methods (including HG, HG-WE, HG-WE (Visual) and HG-WE (Tag)) are with higher computational cost but better search performance compared with Graph, Sequential, Tag Ranking and Uniform Tagger.
2) In HG methods, HG is the most efficient with the worst performance due to there is no weight learning procedure. HG-WE requires the highest computational cost and also achieves the best retrieval performance.

VI. Conclusion

This paper proposes an approach that simultaneously utilizes both visual and textual information for social image search. In the proposed method, both visual content and tags are used to generate the hyperedges of a hypergraph, and a relevance learning procedure is performed on the hypergraph structure where a set of pseudo-relevant samples are employed. Different from the conventional hypergraph learning algorithms, our approach learns not only the relevance scores among images but also the weights of hyperedges. By using the learning of hyperedge weights, the effects of uninformative tags and visual words can be minimized.

To test the performance of the proposed approach, we conducted experiments on a merged dataset including 370K+ images. Experimental results demonstrate that the proposed method achieved better results compared with many baseline methods including graph-based semi-supervised learning, sequential social image ranking, sequential social image ranking, tag ranking, and uniform tagger. We also compared our method with that without hyperedge weight learning. Experimental results show that the hyperedge weight learning procedure is effective on improving the search performance. We further compared the proposed method with that of using only the visual content-based hyperedge and the textual information-based hyperedge respectively. The results show that the proposed method that simultaneously utilizes both the visual content and the textual information achieves better results than that of using them individually.

Besides the relevance performance, it has been pointed out that diversity is also important for search results [33] and

[63]. Most diversification processes, such as the methods in [33], [63], can be performed on the relevance-based social image ranking list obtained by the proposed approach, which is our future work.

REFERENCES

[1] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang, "Visual query suggestion," in *Proc. ACM Conf. Multimedia*, 2009, pp. 15–24.

[2] J. Shen, D. Tao, and X. Li, "QUC-tree: Integrating query context information for efficient music retrieval," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 313–323, Feb. 2009.

[3] J. Shen, D. Tao, and X. Li, "Modality mixture projections for semantic video event detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1587–1596, Nov. 2008.

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[5] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[6] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, and T. Chua, "Camera constraint-free view-based 3-D object retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2269–2281, Apr. 2012.

[7] R. Hong, M. Wang, M. Xu, S. Yan, and T.-S. Chua, "Dynamic captioning: Video accessibility enhancement for hearing impairment," in *Proc. ACM Conf. Multimedia*, 2010, pp. 421–430.

[8] J. Shen and Z. Cheng, "Personalized video similarity measure," *ACM Multimedia Syst. J.*, vol. 15, no. 7, pp. 421–433, 2011.

[9] D. Liu, M. Wang, X.-S. Hua, and H.-J. Zhang, "Semi-automatic tagging of photo albums via exemplar selection and tag inference," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 82–91, Feb. 2011.

[10] Y. Gao, J. Tang, R. Hong, Q. Dai, T. Chua, and R. Jain, "W2Go: A travel guidance system by automatic landmark ranking," in *Proc. ACM Conf. Multimedia*, 2010, pp. 123–132.

[11] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug. 2011.

[12] B. Sigurbjornsson and R. V. Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. ACM Int. World Wide Web Conf.*, 2008, pp. 327–336.

[13] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka, "Can social tagging improve web image search?" in *Web Information Systems Engineering* (Lecture Notes in Computer Science). vol. 5175. Berlin, Germany: Springer-Verlag, 2008, pp. 235–249.

[14] L. Chen, D. Xu, W. Tsang, and J. Luo, "Tag-based web photo retrieval improved by batch mode re-tagging," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3440–3446.

[15] K. Yang, M. Wang, X. S. Hua, and H. J. Zhang, "Social image search with diverse relevance ranking," in *Proc. ACM Conf. Multimedia Model.*, 2009, pp. 174–184.

[16] X. Li, C. G. M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proc. ACM Conf. Multimedia Inf. Retr.*, 2008, pp. 180–187.

[17] D. Liu, M. Wang, L. Yang, X. S. Hua, and H. Zhang, "Tag quality improvement for social images," in *Proc. IEEE Int. Conf. Multimedia*, Jul. 2009, pp. 350–353.

[18] Y. Gao, M. Wang, H. Luan, J. Shen, S. Yan, and D. Tao, "Tag-based social image search with visual-text joint hypergraph learning," in *Proc. ACM Conf. Multimedia*, 2011, pp. 1517–1520.

[19] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, Feb. 2012.

[20] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[21] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, and N. Zhang, "Less is more: Efficient 3-D object retrieval with query view selection," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 1007–1018, Oct. 2011.

[22] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua, "Visual query suggestion: Toward capturing user intent in internet image search," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 6, no. 3, pp. 1–13, 2010.

[23] F. Li, Q. Dai, W. Xu, and G. Er, "Multilabel neighborhood propagation for region-based image retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1692–1604, Dec. 2008.

[24] X. Li, C. G. Snoek, and M. Worring, "Unsupervised multi-feature tag relevance learning for social image retrieval," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2010, pp. 10–17.

[25] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement toward low-rank, content-tag prior and error sparsity," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 461–470.

[26] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[27] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music recommendation by unified hypergraph: Combining social media information and music content," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 391–400.

[28] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang, "Tag ranking," in *Proc. ACM Conf. World Wide Web*, 2019, pp. 351–360.

[29] H. Xu, J. Wang, X. S. Hua, and S. Li, "Tag refinement by regularized LDA," in *Proc. ACM Int. Conf. Multimedia*, 2009, pp. 573–576.

[30] A. Sun and S. S. Bhowmick, "Image tag clarity: In search of visualrepresentative tags for social images," in *Proc. SIGMM Workshop Social Media*, 2009, pp. 19–26.

[31] J. Fan, Y. Shen, N. Zhou, and Y. Gao, "Harvesting large-scale weakly-tagged image databases from the web," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 802–809.

[32] D. Liu, X. S. Hua, M. Wang, and H. Zhang, "Boost search relevance for tag-based social image retrieval," in *Proc. IEEE Int. Conf. Multimedia*, Jul. 2009, pp. 1636–1639.

[33] M. Wang, K. Yang, X. S. Hua, and H.-J. Zhang, "Toward a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, Dec. 2010.

[34] D. Zhou, J. Huang, and B. Schokopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1–8.

[35] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas, "Video object segmentation by hypergraph cut," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1738–1745.

[36] S. Xia and E. Hancock, "Clustering using class specific hyper graphs," in *Proc. Int. Joint IAPR Workshop, Struct., Syntactic, Stat. Pattern Recognit.*, 2008, pp. 318–328.

[37] R. Zass and A. Shashua, "Probabilistic graph and hypergraph matching," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[38] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3376–3383.

[39] A. K. C. Wong and S. W. Lu, "Recognition and shape synthesis of 3-D objects based on attributed hypergraphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 3, pp. 279–290, Mar. 1989.

[40] S. Xia and E. Hancock, "3D object recognition using hyper-graphs and ranked local invariant features," in *Proc. Int. Joint IAPR Struct., Syntactic, Stat. Pattern Recognit.*, 2008, pp. 117–126.

[41] S. Xia and E. Hancock, "Learning large scale class specific hyper graphs for object recognition," in *Proc. Int. Conf. Image Graph.*, 2008, pp. 366–371.

[42] Q. Liu, Y. Huang, and D. Metaxas, "Hypergraph with sampling for image retrieval," *Pattern Recognit.*, vol. 44, nos. 10–11, pp. 2255–2262, 2011.

[43] X. Zhu, "Semi-supervised learning with graphs," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, 2005.

[44] Y.-G. Jiang and C.-W. Ngo, "Bag-of-visual-words expansion using visual relatedness for video indexing," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 769–770.

[45] R. Ji, H. Yao, X. Sun, B. Zhong, P. Xu, and W. Gao, "Toward semantic embedding in visual vocabulary," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 918–925.

[46] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. ACM SIGMM Workshop Multimedia Inf. Retr.*, 2007, pp. 197–206.

[47] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2012.

[48] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Toward optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. ACM Int. Conf. Image Video Retr.*, 2007, pp. 494–501.

[49] R. Ji, H. Yao, W. Liu, X. Sun, and Q. Tian, "Task-dependent visual-codebook compression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2282–2293, Apr. 2012.
[50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[51] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo relevance feedback," in *Proc. Int. Conf. Image Video Retr.*, 2003, pp. 238–247.
[52] J. Wang, Y. G. Jiang, and S. F. Chang, "Label diagnosis through self tuning for web image search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1390–1397.
[53] W. H. Hsu, L. S. Kennedy, and S. F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Conf. Multimedia*, 2007, pp. 971–980.
[54] L. Wu, X. Hua, N. Yu, W. Ma, and S. Li, "Flickr distance," in *Proc. ACM Conf. Multimedia*, 2008, pp. 31–40.
[55] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Conf. Image Video Retr.*, Santorini, Greece, Jul. 2009, pp. 1–9.
[56] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schokopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1–8.
[57] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, "Unified video annotation via multigraph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–746, May 2009.
[58] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.
[59] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, DOI: 10.1109/TIP.2012.2199502.
[60] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.
[61] M. Wang, X.-S. Hua, T. Mei, R. Hong, G.-J. Qi, Y. Song, and L.-R. Dai, "Semi-supervised kernel density estimation for video annotation," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 348–396, 2009.
[62] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–466, 2002.
[63] R. H. V. Leuken, L. Garcia, X. Olivares, and R. Zwol, "Visual diversification of image search results," in *Proc. ACM Int. World Wide Web Conf.*, 2009, pp. 341–350.

Yue Gao received the B.S. degree from the Harbin Institute of Technology, Harbin, China, in 2005, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2008 and 2012, respectively.

He was a Visiting Scholar with Carnegie Mellon University, where he worked with Dr. A. Hauptmann from 2010 to 2011, and a Research Intern with the National University of Singapore, Singapore, and Intel China Research Center, Beijing. His current research interests include multimedia information retrieval, 3-D object retrieval and recognition, and social media analysis.

Meng Wang (M'09) received the B.E. degree from the Special Class for the Gifted Young, and the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively.

He is a Professor with the Hefei University of Technology, Hefei, China. He was an Associate Researcher with Microsoft Research Asia, and then a Core Member in a startup in Silicon Valley. He was with the National University of Singapore, Singapore, as a Senior Research Fellow. He has authored more than 100 book chapters, and journal and conference papers in his areas of expertise. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing.

Dr. Wang was a recipient of the Best Paper Awards in the 17th and 18th ACM International Conference on Multimedia and the Best Paper Award in the 16th International Multimedia Modeling Conference. He is a member of the Association of Computer Machinery.

Zheng-Jun Zha (M'08) received the B.E. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China.

He is a Senior Research Fellow with the School of Computing, National University of Singapore, Singapore. His current research interests include multimedia content analysis, computer vision, and multimedia applications, such as search, recommendation, and social networking.

Dr. Zha was a recipient of the Microsoft Research Fellowship in 2007, the President Scholarship of the Chinese Academy of Science in 2009, and the Best Paper Award in the Association of Computer Machinery (ACM) International Conference on Multimedia in 2009. He is a member of ACM.

Jialie Shen (M'07) received the Ph.D. degree in computer science in the area of large-scale media retrieval and database access methods from the University of New South Wales (UNSW), Sydney, Australia.

He is an Assistant Professor with the School of Information Systems, Singapore Management University (SMU), Singapore. Before SMU, he was a Faculty Member with UNSW, and a Researcher with the Information Retrieval Research Group, the University of Glasgow, Glasgow, U.K. His current research interests include information retrieval in the textual and multimedia domain, economic-aware media analysis, and artificial intelligence, particularly machine perception and its applications on IR and business intelligence.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

Xindong Wu (F'11) received the Bachelor's and Master's degrees in computer science from the Hefei University of Technology, Hefei, China, and the Ph.D. degree in artificial intelligence from the University of Edinburgh, Edinburgh, U.K.

He is a Yangtze River Scholar with the School of Computer Science and Information Engineering, Hefei University of Technology and a Professor of computer science with the University of Vermont, Burlington. His current research interests include data mining, knowledge-based systems, and web information exploration.

Dr. Wu is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of *Knowledge and Information Systems* (Springer), and a Series Editor of the Springer Book Series on Advanced Information and Knowledge Processing. He was the Editor-in-Chief of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (the IEEE Computer Society) from 2005 to 2008. He served as a Program Committee Chair and Co-Chair for the IEEE International Conference on Data Mining in 2003, the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in 2007, and the 19th ACM Conference on Information and Knowledge Management in 2010.