12-2015

Content-Based Visual Landmark Search via Multimodal Hypergraph Learning

Lei ZHU
Huazhong University of Science and Technology

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Hai JIN
Huazhong University of Science and Technology

Ran ZHENG
Huazhong University of Science and Technology

Liang XIE
Huazhong University of Science and Technology

DOI: https://doi.org/10.1109/TCYB.2014.2383389

Published in IEEE Transactions on Cybernetics, 2015 Dec, 45 (12), pp. 2756-2769.
http://doi.org/10.1109/TCYB.2014.2383389

Content-Based Visual Landmark Search via Multimodal Hypergraph Learning

Lei Zhu, Jialie Shen, Hai Jin, *Senior Member, IEEE,* Ran Zheng, and Liang Xie

***Abstract*—While content-based landmark image search has recently received a lot of attention and became a very active domain, it still remains a challenging problem. Among the various reasons, high diverse visual content is the most significant one. It is common that for the same landmark, images with a wide range of visual appearances can be found from different sources and different landmarks may share very similar sets of images. As a consequence, it is very hard to accurately estimate the similarities between the landmarks purely based on single type of visual feature. Moreover, the relationships between landmark images can be very complex and how to develop an effective modeling scheme to characterize the associations still remains an open question. Motivated by these concerns, we propose multimodal hypergraph (MMHG) to characterize the complex associations between landmark images. In MMHG, images are modeled as independent vertices and hyperedges contain several vertices corresponding to particular views. Multiple hypergraphs are firstly constructed independently based on different visual modalities to describe the hidden high-order relations from different aspects. Then, they are integrated together to involve discriminative information from heterogeneous sources. We also propose a novel content-based visual landmark search system based on MMHG to facilitate effective search. Distinguished from the existing approaches, we design a unified computational module to support query-specific combination weight learning. An extensive experiment study on a large-scale test collection demonstrates the effectiveness of our scheme over state-of-the-art approaches.**

***Index Terms*—Content-based visual landmark search, high-order relations, multimodal hypergraph (MMHG), visual diversity.**

I. Introduction

OVER the past decades, with the prevalence of intelligent mobile devices and advanced mobile multimedia services, various kinds of geo-referenced multimedia data has significantly reshaped the way how we search and represent the knowledge about a geolocation. One of the typical examples is landmark image. Due to a wide range of real applications such as tour guide recommendation [1] and geo-localization [2], developing intelligent algorithms to facilitate accurate visual landmark search enjoys great importance.

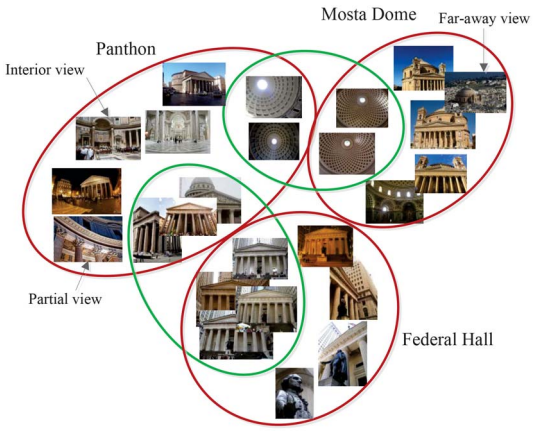

Fig. 1. Landmark images have diverse visual contents. There are complex relations among them. Several different image views describe the same landmark (images belong to red ellipse). Different landmarks share several close image views (images belong to green ellipse).

Comparing to the image data in traditional content-based image retrieval (CBIR) systems, landmark images have a few distinguishing characteristics. In particular, a landmark can be photographed from different viewpoints, under different lighting conditions and for different attractive spots. All of these significantly increase the visual diversity of recorded query and images stored in the database. As pointed out in [3], images in each landmark category can be further classified into several sub-categories according to the views they describe, such as partial view, interior view, far-away view, and etc. Fig. 1 gives a few examples of several real-world landmark images. As shown, several different image views describe the same landmark, and some different landmarks may have several views which are very similar to each other. This characteristic makes the distance between landmarks and the high-order relations among images hard to be estimated directly and accurately. Further, it poses great challenges on the design of content-based visual landmark search (CBVLS) system.

Existing CBVLS systems can be generally categorized into two major families: 1) similarity-based search [3], [4] and 2) graph-based search [5]–[7]. Similarity-based search leverages low-level feature matching to return images sequentially. This solution suffers from two main

Manuscript received June 6, 2014; revised September 28, 2014 and November 30, 2014; accepted December 3, 2014. Date of publication January 6, 2015; date of current version November 13, 2015. This work was supported in part by the National High Technology Research and Development Program of China under Grant 2012AA01A306, and in part by the National Natural Science Foundation of China under Grant 61133008. This paper was recommended by Associate Editor L. Shao. *(Corresponding author: Ran Zheng.)*

L. Zhu, H. Jin, and R. Zheng are with the Services Computing Technology and System Laboratory, Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zhraner@hust.edu.cn).

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

L. Xie is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China.

disadvantages. First, it is hard to effectively learn query-specific combination weights for different features. In this case, multiple features are usually combined with the same weights for all query images. Since query images can have high visual diversity, if CBVLS system fails to capture discriminative capabilities of the involved features on describing visual contents of query images, search performance may be degraded greatly. Second, similarity-based search relies on simple similarities, which cannot model the high-order relations among landmark images effectively. To represent more complex relations, graph-based search applies graph to model the search process. In graph, image is usually considered as vertex and the edge connecting two vertices represents a relation. Although graph enjoys a good capability to model more complex relations, one edge in a graph only connects two vertices and two edges only share one vertex.[1] Indeed, high-order relations still cannot be fully characterized by using simple graph. Therefore, graph also fails to capture many important characteristics of landmark images.

In order to mitigate the drawbacks of current methods, we explore hypergraph-based approach to solve the problem of CBVLS. One of the most desirable advantages of hypergraph is that it can capture high-order relations among images. This can lead to high-quality model for characterizing landmark images. Thus, in the hypergraph-based model, images in databases are considered as vertices, and several similar images in database form an edge.[2] In this way, since a hyperedge connects several vertices and two hyperedges also share several vertices, high-order relations among images can be modeled effectively and comprehensively.

The similar visual features or patterns such as color, texture, or shape can be easily found in the contents of landmark images from same or similar geolocation. Also, landmark images from same or similar geolocation could have diverse contents. As such, the uni-modal hypergraph (UMHG) might be not able to comprehensively model or characterize them. Inspired by recent success in visual data modeling by using multimodal-based fusion [8]–[11], we develop multimodal hypergraph (MMHG)-based approach in this paper. Basic idea is to construct multiple hypergraphs based on different visual modalities. The combination weights are exploited to measure the importance of each hypergraph. Multiple hypergraphs are integrated into an unified framework to take heterogeneous discriminative information in account. To capture the distinctiveness of query images, online search, and weight learning are formulated into an unified computational module, which iteratively calculates similarity scores of database images, and assigns proper weights to the associated hypergraphs. The main contributions of this paper can be summarized as follows.

1) Hypergraph is leveraged to model the high-order relations between landmark images. To the best of our knowledge, no existing study explores multimodal hypergraph-based approach for the task of CBVLS.
2) *A Novel Model*: MMHG is developed to integrate multiple hypergraphs based on heterogeneous visual modalities for the purpose of representing highly complex relations and effectively capturing the diverse visual contents. Combination weights are learnt to measure the importance of the various associated hypergraphs.
3) Search and weight learning is conducted using an unified computational framework, which iteratively calculates similarity scores of the images stored in database, and assigns proper query specific weights to the involved hypergraphs.

The remainder of this paper is structured as follows. Section II reviews the related work. Overview of the proposed CBVLS system is provided in Section III. Section IV introduces the components of MMHG-based CBVLS system in detail. Experimental configuration and empirical experimental results are presented in Section V. Section VI finally summarizes this paper with the conclusion.

II. Related Work

A. Landmark Search

Most existing approaches on CBVLS mainly focus on improving accuracy or diversity of the search results [3], [4], [12]–[14]. They can be generally classified into two major categories: single modal-based and multimodal-based. State-of-the-art single feature-based CBVLS systems are based on bag-of-visual-words (BOVW) [15] or its improved variants. Visual synonyms and 3-D visual phrases are proposed in [4] and [16], respectively to describe the visual contents of landmark images, by identifying pairs of distant words with similar appearance or characterizing the spatial structure of 3-D landmark. It has been reported in [4] that, with less visual-words, visual synonyms can also achieve comparable performance. However, extraction of visual synonyms relies on spatial verification-based image reranking and thus cannot be extracted in a one-shot process. 3-D visual phrase [16] also cannot be applied to our case directly, because it relies on time-consuming 3-D reconstruction which is hard to be implemented on diverse landmark images. In [17], structure from motion-based [18] 3-D reconstruction is exploited for CBVLS. This approach suffers from the same drawback with [16]. Cheng *et al.* [3] proposed a multiple feature fusion strategy for CBVLS. Their experimental results demonstrate that search accuracy is low and far from practical applications. They point out that the potential reason for low accuracy is that high-order relations among landmark images cannot be simply captured by low-level features and raw similarity measures. In addition to the work presented above, there are also some efforts that are made to diversify search results of CBVLS [13], [19].

Many researchers also propose methods to conduct CBVLS on mobile platform [14], [20]. These approaches are mainly based on variants of BOVW [15]. Efforts are mainly made on converting BOVW as a compact descriptor to reduce memory consumption and speedup network transmission.

[1] In this paper, edge and vertex represent relation and image, respectively.
[2] In hypergraph, edge is also termed as hyperedge.

B. Graph and Hypergraph Learning

Graph learning has been proven to be an effective method to improve the performance of unsupervised and semi-supervised data modeling by uncovering the underlying structures of image collections (e.g., correlation and dependency). It has been widely applied to many fields, such as clustering [21], [22], dimensionality reduction [23], [24], image retrieval [25], [26], and etc. For example, Jones and Shao [27] proposed a feature grouped spectral multigraph by aggregating results from multiple graphs, which are built in mutually independent subsets of the original feature space. Shao *et al.* [28] used graph to embed the penultimate hierarchical discriminative manifolds into a compact representation. Although their reported experimental results are promising, their employed graph cannot be applied to our case directly because of its inability to capture high-order relations among landmark images.

Hypergraph does not suffer from the drawbacks of graph, and due to high capability to model complex data, hypergraph has attracted more and more attention in [29]–[32]. One of typical example is a hypergraph-based framework for video object segmentation proposed by Huang *et al.* [29]. In this paper, over-segmented image patches are taken as vertices, and vertices with the same attribute construct a hyperedge. Via hypergraph modeling, inherent spatio-temporal neighborhood relations among patches are naturally captured. Zhang *et al.* [30] constructed a feature correlation hypergraph to capture the high-order correlation among multimodal features, which directly extends the previous binary correlations. Huang *et al.* [31] formulated CBIR in a probabilistic hypergraph (PHG) framework. Images are taken as vertices and a hyperedge is comprised of a centroid vertex and its several nearest neighbors. A vertex is assigned to hyperedge in a probabilistic way. Although multiple features can be fused effectively using single framework [31], it only adopts simple equal weights (EW) to measure the importance of the employed features. Consequently, it fails to describe their discriminative capabilities. This paper constructs the hypergraph without the above drawback, and formulates search and weight learning into an unified computational module, which iteratively performs searching process and learns the optimal feature combination weights for each query image. In [32], hypergraph is leveraged to solve the problem of view-based 3-D object retrieval. Each vertex is an object and a cluster of views constructs a hyperedge. K-means is adopted to generate multiple overlapping hyperedges. Although promising experimental results are reported, their hyperedge construction approach cannot be applied for CBVLS directly. This is because images in different landmarks have different visual diversity degrees, which makes the key parameter—number of clusters hard to be adjusted. More importantly, their learning framework only takes single modality into account, which fails to leverage discriminative information from multiple modalities.

C. Multimodal Visual Feature Fusion

Visual contents of landmark images can be complex and highly diverse. Different kinds of visual features could have various contribution on visual recognition task. Thus, fusion of multimodal visual features can effectively improve the discriminating ability of single feature-based systems or algorithms. Early-fusion [33] and late-fusion [34] are two classic strategies in literature. Early fusion concatenates multiple features into a holistic one and conducts learning on high-dimensional vectors. This technique inevitably results in information loss and suffers from "curse of dimensionality." In contrast, late-fusion performs learning processes independently in each single modality and then fuses the obtained results. Proper combination weights are essential important in this case for good performance.

In [35] and [36], multiple kernel learning [37] is proposed to combine multiple features for object recognition and detection, respectively. These methods suffer from high computation burden brought by kernel matrix computation. Yang *et al.* [6] fused multiple features in framework of hierarchical regression to annotate concepts and recognize actions in videos. Ma *et al.* [38] proposed to detect events by multiple video attributes learned from different types of features. Wang *et al.* [39] used multiple feature learning for action recognition by applying structural analysis to discover the common subspace shared by multiple features. Yang *et al.* [26] proposed to rank the image search results with multiple features in a multigraph framework with Laplacian matrix learned by local regression and global alignment (LRGA). However, the approaches introduced above [6], [26], [38], [39] all adopt the same weights to fuse multiple features, which may lose discriminative information, especially when the discriminative capabilities of the involved features are not equal. In addition, Xu *et al.* [40] adopts optimal thresholding-based feature weighting (FWOT) to fuse multiple features for video analysis, which can simultaneously learn the weights, thresholding, and smoothing parameters in a joint framework. However, FWOT is specially designed for visual classification, where learning optimal combination weights needs large quantities of labeled images.

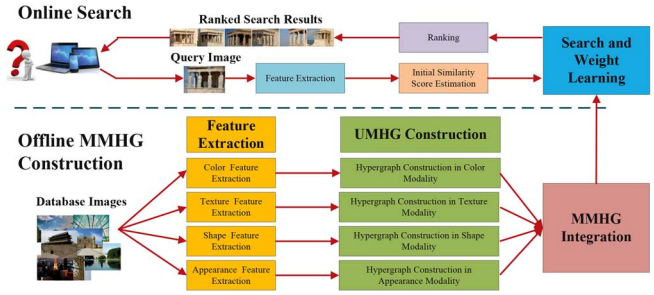

Fig. 2. Overall view of the MMHG-based CBVLS system. Best viewed in color and with pdf magnification.

III. System Overview

This section briefly provides overview of the proposed MMHG-based CBVLS system. As shown in Fig. 2, the system architecture consists of two key components: 1) offline MMHG construction and 2) online search.

1) *Offline MMHG Construction:* It is designed to build MMHG, based on which the whole CBVLS process

is performed. More specifically, this process can be further divided into three independent sub-processes: 1) low-level feature extraction; 2) UMHG construction; and 3) MMHG integration. In the system, five visual features from different modalities are extracted to describe the diverse visual contents of landmark images from different aspects. Then, UMHG which represents the relations of images stored in database is constructed in each visual modality. Finally, these UMHGs are integrated into an unified MMHG with combination weights to represent more complex relations.

2) *Online Search*: Query image is first submitted by user. Low-level features of it are obtained by the same feature extraction processes performed on database images. Then, initial similarity scores of database images are set to represent the similarity between query image and database images. Next, based on the constructed MMHG, similarity scores are updated iteratively in module of search and weight learning. Finally, the estimated similarity scores are ranked in descending order, and their corresponding database images are returned back to user.

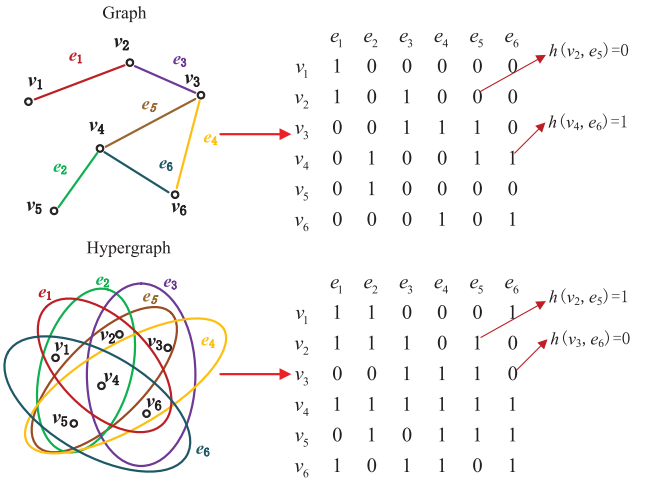

Fig. 3. Left: toy example of graph and hypergraph which include six vertices. Right: corresponding incidence matrices. The element $h(v_i, e_j) = 1$ means that the vertex v_i belongs to edge e_j, and vice versa. In graph, edge is comprised of two vertices and two edges share with only one vertex. In hypergraph, centroid vertex and its three nearest neighbors comprise a hyperedge. Two hyperedges share with two or more vertices. Note that hypergraph degenerates to graph when $k = 1$. Best viewed in color and with pdf magnification.

IV. MMHG-Based Content-Based Visual Landmark Search

In this section, we provide the details of proposed MMHG-based CBVLS system. First, we introduce five visual features used in this paper. Second, we formulate the hypergraph construction in a single modality. Third, we extend the UMHG formulation to MMHG to integrate hypergraphs constructed in multiple modalities. Finally, we present the module of search and weight learning.

A. Low-Level Feature Extraction

Our system considers five widely used low-level visual features. Their details are as follows.

1) *Color Moments (CM) [41]:* First, image is partitioned into regions without overlapping with 3×3 grid. Then, in each segmented region, color mean, color variance, and color skewness are extracted in each color channel. Finally, features calculated from regions are concatenated to form 81-D vector.
2) *Local Binary Pattern (LBP) [42]:* LBP is simple yet powerful texture descriptor to describe local structure of image by comparing centering pixel with surrounding pixels. It has good property of tolerating illumination changes. In this paper, 58-D LBP is adopted for texture description.
3) *Histogram of Oriented Gradients (HOG) [43]:* HOG is an effective descriptor which describes the shape information of image. It counts occurrences of gradient orientation in localized portions of an image and normalizes the result using a block-wise pattern. In this paper, 31-D HOG is adopted for shape description.
4) *BOVW [15]:* BOVW quantizes order-less local features to visual-words and represents image as frequency histograms of visual-words, and it has been widely used in [44]. In this paper, densely sampling strategy is employed to detect interest points and scale invariant feature transform [45] is used to describe image patches. In our case, each interest point is represented by a vector of 128 dimensions. The best performance of BOVW is obtained when the visual-word vocabulary size is set to 400.
5) *GIST [46]:* GIST is a widely used global feature, which exploits a set of perceptual dimensions (naturalness, openness, roughness, expansion, and ruggedness) to describe the spatial structure of the image. First, Gabor filter with fixed parameters is built. Then, image is filtered and segmented into grid cells where orientation histograms are extracted. In this paper, image is segmented by a 4×4 grid and features extracted from sub-regions are concatenated into a 512-D feature vector.

B. UMHG Construction

This section introduces the details of UMHG construction. Fig. 3 shows a toy example of graph, hypergraph, and their corresponding incidence matrices. Hypergraph is different from graph on its inner structure. In graph, edge is comprised of two vertices, while hyperedge in hypergraph is comprised of three or more vertices.[3] In addition, edge in graph can only share one vertex, while hyperedge can share two or more vertices. From this perspective, graph can only describe simple relations among images, while hypergraph can describe high-order relations. More specifically, in CBVLS, it can describe the high-order relations, such as whether three or more images describe the same landmark, whether three or more landmarks share close views, and etc.

[3]In our case, a hyperedge contains $k + 1$ vertices.

A hypergraph constructed in modality m can be denoted as $G_m = (V_m, E_m, W_m)$ $(m = 1, 2, \ldots, M)$, where M is the number of visual modalities, $V_m = \{v_{im}\}_{i=1}^{N}$ denotes the vertex set, $E_m = \{e_{jm}\}_{j=1}^{N}$ denotes the hyperedge set, $W_m = \{w_m(e_{jm})\}_{j=1}^{N}$ denotes the weight set for hyperedges, $w_m(e_{jm})$ is weight of hyperedge e_{jm}. Each image in database is considered as a vertex. A vertex and its k nearest vertices comprise a hyperedge. For example, in Fig. 3, centroid vertex v_1 and its nearest vertices v_2, v_4, v_6 comprise hyperedge e_1. The number of hyperedges is equal to the number of vertices (images). Similarities of images in modality m are measured by the distances of their corresponding features. In this case, the size of hyperedge is $k + 1$. As visual contents of images in landmarks are distributed with different visual diversity degree, it is hard to estimate the number of visual clusters. Therefore, the hypergraph construction method here is different from the approach proposed in [32], which adopts simple K-means clustering to generate hyperedges.

Formally, a hypergraph built in modality m can be represented with a $|V_m| \times |E_m|$ incidence matrix H_m ($|\cdot|$ denotes the cardinality of set, $|V_m| = |E_m| = N$, N is the number of database images). For example, the incidence value between vertex v_{im} and hyperedge e_{jm} in H_m is given as $(i, j = 1, 2, \ldots, N)$

$$h_m(v_{im}, e_{jm}) = \begin{cases} 1 & dist\left(x_{im}, x_{jm}\right) \le k\_dist\left(v_{jm}\right) \\ 0 & \text{otherwise} \end{cases}$$
$$dist\left(x_{im}, x_{jm}\right) = \left|\left|x_{im} - x_{jm}\right|\right|_F \tag{1}$$

where $dist(x_{im}, x_{jm})$ denotes the distance between image i and j in modality m, $||\cdot||_F$ denotes the Frobenius norm, $h_m(v_{im}, e_{jm})$ is the ith row, jth column element in H_m, image j is centroid vertex of hyperedge e_{jm}, x_{im}, x_{jm} are features of image i and j extracted in modality m, respectively, $k\_dist(v_{jm})$ denotes the distance between image j and its kth nearest database images. Different from graph, the degree of hyperedge is defined as the number of images that belong to it. For example, for hyperedge e_{jm}, its degree $\delta_m(e_{jm})$ is defined as

$$\delta_m\left(e_{jm}\right) = \sum_{i=1}^{N} h_m\left(v_{im}, e_{jm}\right). \tag{2}$$

Since each hyperedge contains $k + 1$ vertices in our approach, the number of nonzero elements in incidence matrix is $N \times (k + 1)$, and the degrees of hyperedges are all equal to $k + 1$. In this paper, we explore visual consistence to measure the importance of hyperedges. The hyperedge with more visual consistence is assigned with higher weight, and vice versa. In our case, visual consistence is measured by the similarities of images in hyperedge. Formally, for hyperedge e_{jm}, its weight $w_m(e_{jm})$ is defined as

$$w_m\left(e_{jm}\right) = \sum_{a,b=1}^{N} h_m\left(v_{am}, e_{jm}\right) h_m\left(v_{bm}, e_{jm}\right) exp\left(-\frac{||x_{am} - x_{bm}||_F}{\sigma_{e_{jm}}}\right) \tag{3}$$

where a and b is arbitrary database images, $\sigma_{e_{jm}}$ is the normalization factor, which is calculated as the mean distance of all images that belong to the hyperedge

$$\sigma_{e_{jm}} = \frac{1}{(k+1)^2} \sum_{a,b=1}^{N} h_m\left(v_{am}, e_{jm}\right) h_m\left(v_{bm}, e_{jm}\right) ||x_{am} - x_{bm}||_F. \tag{4}$$

With the calculated weights of hyperedges, the degree of each vertex is defined as the sum of the weights of hyperedges that the vertex belongs to

$$d(v_{im}) = \sum_{j=1}^{N} w_m\left(e_{jm}\right) h_m\left(v_{im}, e_{jm}\right). \tag{5}$$

Based on the constructed hypergraph, the search process can be formulated in a transductive learning framework. The formulation objective is to make the images, that belong to the highly weighted hyperedge, be assigned with similar similarity scores, so that similar database images with query image can be ranked at top positions. Denote f as the similarity score vector of database images. Optimization function $\Phi^m(f)$ of hypergraph construction in modality m can be formulated as

$$\Phi^m(f) = \arg\min_f \quad \Omega^m(f) + \lambda R^m(f) \tag{6}$$

where loss term $\Omega^m(f)$ is used to reduce empirical loss, regularization term $R^m(f)$ is used to avoid overfitting, $\lambda > 0$ is the balance factor that plays a trade-off between two terms. Specifically, $\Omega^m(f)$ is calculated as

$$\Omega^m(f) = \frac{1}{2} \sum_{i,j,l=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{lm}, e_{jm}\right) h_m\left(v_{im}, e_{jm}\right)}{\delta_m\left(e_{jm}\right)} \left(\frac{f(v_{lm})}{\sqrt{d(v_{lm})}} - \frac{f(v_{im})}{\sqrt{d(v_{im})}}\right)^2 \tag{7}$$

where $f(v_{lm})$ and $f(v_{im})$ denote the similarity score for vertex v_{lm} and v_{im}, $d(v_{lm})$ and $d(v_{im})$ denote the degree of vertex v_{lm} and v_{im}. The learning objective is to assign more similar scores for images that belong to many incidental hyperedges. Denote y as the initial score vector, $R^m(f)$ can be defined as

$$R^m(f) = ||f - y||_F = \sum_{i=1}^{N} (f(v_{im}) - y(v_{im}))^2 \tag{8}$$

where $y(v_{im})$ is 1 if the image that corresponds to vertex v_{im} is considered as query one. Then

$$\begin{aligned} \Omega^m(f) &= \sum_{i,j,l=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{lm}, e_{jm}\right) h_m\left(v_{im}, e_{jm}\right)}{\delta_m\left(e_{jm}\right)} \left(\frac{f(v_{im})^2}{d(v_{im})} - \frac{f(v_{lm}) f(v_{im})}{\sqrt{d(v_{lm}) d(v_{im})}}\right) \\ &= \sum_{i=1}^{N} f(v_{im})^2 \sum_{j=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{im}, e_{jm}\right)}{d(v_{im})} \sum_{l=1}^{N} \frac{h_m\left(v_{lm}, e_{jm}\right)}{\delta_m\left(e_{jm}\right)} \\ &\quad - \sum_{i,j,l=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{lm}, e_{jm}\right) h_m\left(v_{im}, e_{jm}\right) f(v_{lm}) f(v_{im})}{\delta_m\left(e_{jm}\right) \sqrt{d(v_{lm}) d(v_{im})}}. \end{aligned} \tag{9}$$

From (2) and (5), we can easily derive that

$$\sum_{l=1}^{N} \frac{h_m\left(v_{lm}, e_{jm}\right)}{\delta_m\left(e_{jm}\right)} = 1 \quad \sum_{j=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{im}, e_{jm}\right)}{d(v_{im})} = 1. \tag{10}$$

Therefore

$$\begin{aligned}\Omega^m(f) &= \sum_{i=1}^{N} f(v_{im})^2 \\ &\quad - \sum_{i,j,l=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{lm}, e_{jm}\right) h_m\left(v_{im}, e_{jm}\right) f(v_{lm}) f(v_{im})}{\delta_m\left(e_{jm}\right) \sqrt{d(v_{lm}) d(v_{im})}} \\ &= f^T \Delta^m f \end{aligned} \tag{11}$$

where Δ^m is Laplacian matrix of hypergraph. It can be calculated as $\Delta^m = I - \Theta^m$. Θ^m can be calculated as $\Theta^m = D_{v_m}^{-1/2} H_m D_{w_m} D_{e_m}^{-1} H_m^T D_{v_m}^{-1/2}$, where D_{v_m}, D_{e_m}, and D_{w_m} are the diagonal matrices of the vertex degrees, edge degrees, and hyperedge weights in modality m, respectively

$$\begin{aligned} D_{v_m}(i,j) &= \begin{cases} d(v_{im}) & i = j \\ 0 & i \neq j \end{cases} \\ D_{e_m}(i,j) &= \begin{cases} \delta_m\left(e_{jm}\right) & i = j \\ 0 & i \neq j \end{cases} \\ D_{w_m}(i,j) &= \begin{cases} w_m\left(e_{jm}\right) & i = j \\ 0 & i \neq j. \end{cases} \end{aligned} \tag{12}$$

Equation (6) can be reformed as

$$\Phi^m(f) = \arg\min_f \quad f^T \Delta^m f + \lambda ||f - y||_F. \tag{13}$$

C. MMHG Construction

To gain more comprehensive modeling capability, we develop MMHG-based on UMHG. The key motivation for the multimodal-based extension is mainly based on the two observations as below.

1) Different landmark images may have similar visual appearance in terms of color, texture, and shape. Integrating multiple hypergraphs from heterogeneous sources can represent more complex relations.
2) Images in a landmark category have diverse visual contents. Hypergraph built from a single modality may be not effective to capture diverse visual contents of query and database images, while combing discriminative features from multiple modalities may improve the performance of single one.

Moreover, hypergraphs constructed from heterogeneous visual modalities generally possess different discriminating ability. Weighting them equally cannot strengthen the discriminative hypergraph and attenuate the weak ones, while assigning proper weight for each hypergraph and query image specifically can improve the performance further. Therefore, in this paper, we explore combination weights $\{\alpha_m\}_{m=1}^M$ to measure the importance of UMHGs. The weight is higher if hypergraph is more discriminative, and vice versa. In this paper, we denote hypergraphs constructed in M modalities using $G_1 = (V_1, E_1, W_1)$, $G_2 = (V_2, E_2, W_2), \ldots, G_M = (V_M, E_M, W_M)$. $\{V_m\}_{m=1}^M$, $\{E_m\}_{i=m}^M$, $\{W_m\}_{m=1}^M$ are the vertex set, hyperedge set, and weight set for M hypergraphs, respectively. Denote $\{H_m\}_{m=1}^M$, $\{D_{v_m}\}_{m=1}^M$, $\{D_{e_m}\}_{m=1}^M$, $\{D_{w_m}\}_{m=1}^M$ are incidence matrices, vertex degree matrices, hyperedge degree matrices, and hyperedge weight matrices for M hypergraphs, respectively. The combined hypergraph is constructed by fusing M hypergraphs with linear combination weights $\{\alpha_m\}_{m=1}^M$. In MMHG, we redefine the empirical loss $\Omega(f)$ as the sum of weighted empirical loss of M UMHGs

$$\begin{aligned}\Omega(f) &= \sum_{m=1}^{M} \alpha_m \Omega^m(f) \\ &= \frac{1}{2} \sum_{m=1}^{M} \alpha_m \sum_{i,j,l=1}^{N} \frac{w_m\left(e_{jm}\right) h_m\left(v_{lm}, e_{jm}\right) h_m\left(v_{im}, e_{jm}\right)}{\delta_m\left(e_{jm}\right)} \\ &\quad \left(\frac{f(v_{lm})}{d(v_{lm})} - \frac{f(v_{im})}{d(v_{im})}\right)^2 \\ &= \sum_{m=1}^{M} \alpha_m f^T \left(I - \Theta^m\right) f = f^T \sum_{m=1}^{M} \alpha_m \left(I - \Theta^m\right) f \\ &= f^T \sum_{m=1}^{M} \alpha_m \Delta^m f = f^T \Delta f \end{aligned} \tag{14}$$

where $\Delta = \sum_{m=1}^M \alpha_m \Delta^m = \sum_{m=1}^M \alpha_m (I - \Theta^m)$ is Laplacian matrix of the MMHG. Since linear weights are adopted for combination, (14) may result in trivial results when $\{\Delta^m\}_{m=1}^M$ are equal with each other. To avoid this case, we introduce a smooth factor γ to relax α to α_m^γ. Therefore, (14) is transformed to

$$\Omega(f) = f^T \sum_{m=1}^{M} \alpha_m^\gamma \Delta^m f = f^T \Delta f \tag{15}$$

where $\Delta = \sum_{m=1}^M \alpha_m^\gamma \Delta^m = \sum_{m=1}^M \alpha_m^\gamma (I - \Theta^m)$. Similar to UMHG construction, regularization term $R(f)$ in MMHG can be calculated as $R(f) = \sum_{i=1}^N (f(v_i) - y(v_i))^2 = ||f - y||_F$. In this way, the optimization objective function $\Phi(\alpha, f)$ of MMHG is formulated as

$$\begin{aligned}\Phi(\alpha, f) &= \arg\min_{\alpha, f} \quad \Omega(f) + \lambda R(f) \\ &= \arg\min_{\alpha, f} \quad f^T \Delta f + \lambda ||f - y||_F. \end{aligned} \tag{16}$$

Since the combination weights should be guaranteed to be nonnegative and the sum of them should be guaranteed to be 1, the objective function should be optimized subjecting to the following condition:

$$\sum_{m=1}^{M} \alpha_m = 1, 0 \leq \alpha_m \leq 1, \lambda > 0. \tag{17}$$

D. Search and Weight Learning

Different users might use different query images to search the same landmark and thus query images can have highly diverse visual contents. Accurate computing of query specific combination weights can play very important role in the UMHG that can better characterize the visual contents of query. In this paper, we model search and weight learning in an unified computational framework, which iteratively calculates

the similarity scores of database images and adaptively learns query specific combination weights.

The whole CBVLS is performed on the constructed MMHG. At the stage of online search, a query image is uploaded from user and low-level features of it are obtained by the same feature extraction processes performed on database images. There are two cases need to be dealt with separately: 1) query images are already in the database and 2) query images are outside the database. When query images are already in database, inner structure of hypergraph built in the offline process can be preserved without any adjustment. In this case, image search results can be easily obtained by ranking the calculated similarity scores directly. In contrast, when the query images are not in database, the hypergraph should be adjusted accordingly since new vertex and hyperedge are added. However, the adjustment process will bring additional computations, which may generate negative effects on the efficiency of online learning. Inspired by the success of query expansion in graph learning [26], we apply it here to deal with the problem of out of sample. More specifically, the positions in the initial similarity score vector, which correspond to k nearest images of query are set to 1, and the remaining positions are set to 0. Formally

$$y(v_{im}) = \begin{cases} 1 & dist(x_{im}, x_{qm}) < k\_dist(v_{qm}) \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

With the initial similarity score vector f for query image, the estimated similarity scores of all database images in f are obtained by solving the (16) subjecting to the conditions in (17). Fortunately, the optimization function is convex with respect to both of them, the optimization parameters can be converged to certain values. Since there are two parameter sets that are involved in the optimization objective function, they cannot be solved at one time. In this paper, we propose an alternating optimization approach to separately optimize the parameters, namely, the combination weight vector α and similarity score vector of f. Specifically, we optimize each parameter set by fixing the other.

We calculate the partial derivatives of $\Phi(\alpha, f)$ with respective to α and f as

$$\begin{aligned} \frac{\partial \Phi(\alpha, f)}{\partial f} &= 2 \sum_{m=1}^{M} \alpha_m^{\gamma} \left(\Delta^m f + \lambda (f - y) \right) \\ \frac{\partial \Phi(\alpha, f)}{\partial \alpha_m} &= \gamma \alpha_m^{\gamma - 1} \left(f^T \Delta^m f + \lambda ||f - y||_F \right) \end{aligned} \tag{19}$$

where $\gamma > 0$ is the smooth factor, which reflects the effect of the smoothness difference of hypergraphs.

By fixing α, the problem is transformed to UMHG-based search. By substituting the Laplacian matrix of UMHG to that of MMHG, we can obtain the solution of f. Formally, its solution can be given as

$$f = \left(I + \frac{1}{\lambda} \Delta \right)^{-1} y. \tag{20}$$

Following the rules proposed in [47], (20) can be efficiently solved via iterated computing. Detailed steps are presented in Algorithm 1. Details of proof are presented in the Appendix.

Algorithm 1 Iterative Computation for f

Input:

Initial similarity score vector of database images, $\bar{y}$.

Output:

$f = y^t$

1: Initialize $y^1 = \bar{y}$, $t = 1$.
2: **repeat**
3: $t = t + 1$.
4: Update y via: $y^t = \frac{1}{1+\lambda}(I - \Delta) y^{t-1} + \frac{\lambda}{1+\lambda} \bar{y}$.
5: **until** $|y^t - y^{t-1}| < \theta$
6: **return** y^t

By fixing f, the (16) is transformed to

$$\begin{aligned} &\arg\min_{\alpha} \sum_{m=1}^{M} \alpha_m f^T \Delta^m f + \lambda ||f - y||_F \\ &\text{s.t.} \quad \sum_{m=1}^{M} \alpha_m = 1, 0 \le \alpha_m \le 1, \lambda > 0. \end{aligned} \tag{21}$$

The solution of α can be derived as

$$\alpha_m = \frac{\left(\frac{1}{f^T \Delta^m f + \lambda ||f - y||_F} \right)^{\frac{1}{\gamma - 1}}}{\sum_{m=1}^{M} \left(\frac{1}{f^T \Delta^m f + \lambda ||f - y||_F} \right)^{\frac{1}{\gamma - 1}}}. \tag{22}$$

The complete procedure of MMHG-based CBVLS can be shown in Algorithm 2. The computational cost of MMHG-based CBVLS consists of two major parts. One is for offline MMHG construction and the other is for online CBVLS. It can be easily derived that the computational cost of MMHG construction is $O(MN^2)$, where M is the number of visual modalities, N is the number of database images. MMHG construction is offline process, thus its calculation has no impact on the efficiency of online CBVLS. The process of initial score vector setting in online CBVLS can be completed with time complexity $O(N)$. Search and weight learning costs $O(MTN)$ (T is the number of iterations), which is linear with the number of database images. Furthermore, both the calculation of f and $\{\alpha_m\}_{m=1}^{M}$ consist of many processes of sparse matrix-vector multiplication (note that, Δ is sparse matrix), which can be efficiently stored with sparse matrix compression, and accelerated with parallel implementation [48]. Therefore, the online search approach proposed in this paper can be applied to large-scale landmark search.

V. Experiments

This section presents experimental results and analysis. First, experimental configuration including experimental dataset and testing method is introduced. Second, we present comparative study to show the superior performance of MMHG compared with state-of-the-art approaches. Third, various factors that directly influence the performance of MMHG are discussed in detail. Fourth, parameter experiments are performed to show the robustness of the MMHG when parameters are varied. Finally, we apply MMHG to the task of general image search and evaluate its performance.

Algorithm 2 MMHG-Based CBVLS

Input:

Query image Q and database images $I_1, I_2, \ldots, I_N$.

Output:

Landmark search results.

Offline MMHG Construction

Extract low-level features $\{x_{11}, x_{12}, \ldots, x_{1M}, x_{21}, x_{22}, \ldots, x_{2M}, \ldots, x_{N1}, x_{N2}, \ldots, x_{NM}\}$ for all database images.
Consider each image in database as a vertex. Find k most similar images for each of them. $k+1$ images construct the hyperedge of UMHG in each modality.
Calculate incidence matrix, vertex degree matrix, hyperedge degree matrix, and hyperedge weight matrix for each UMHG.
Output incidence matrices $\{H_m\}_{m=1}^{M}$, vertex degree matrices $\{D_{v_m}\}_{m=1}^{M}$, hyperedge degree matrices $\{D_{e_m}\}_{m=1}^{M}$, hyperedge weight matrices $\{D_{w_m}\}_{m=1}^{M}$ for online CBVLS.

Online CBVLS

Extract low-level features $\{x_{q1}, x_{q2}, \ldots, x_{qM}\}$ for query image.
Find k nearest database images for query in raw feature space.
Initialize the similarity score vector according to (18).
Calculate Laplacian matrices $\{\Delta^m\}_{m=1}^{M}$ for MMHG.
for $t = 1$ to T **do**
    Update weight vector α according to (22).
    Based on the obtained α, update similarity score vector f according to Algorithm 1.
end for
Rank the similarity score vector f in descending order.
return Ranked order of database images.

A. Experimental Dataset and Testing Method

To facilitate experimental study, we develop a test collection called landmark-25, which contains real-world landmark images by crawling the images from Flickr. For each landmark, candidate images are first obtained by retrieving images from Flickr with relevant keywords and the provided API. Images with low relevance and quality are then removed manually. Finally, landmark-25 has 25 landmark categories, including the images photographed from different viewpoints, under different lighting conditions, and for different beauty spots. To make evaluation fair and robust, the same number of images (200 images) from each landmark category are randomly selected to comprise database images. Twenty images are randomly selected as query images. This dataset is challenging because visual contents of images in a landmark category have high visual diversity. Typical images are shown in Fig. 4. In evaluation, images in the same category are considered as relevant, and vice versa.

To measure the performance of MMHG-based CBVLS, we use standard evaluation metric precision-scope, which has been also employed in [31] and [49] for performance evaluation. For given NQ query images, the average search precision is

Fig. 4. Typical images sampled from landmark-25. Landmark names and their abbreviations are listed under the sample images.

defined as

$$\text{Precision} = \frac{1}{NQ}\sum_{i=1}^{NQ}\frac{R_i}{\text{Scope}} \tag{23}$$

where NQ denotes the number of query images (in this paper, $NQ = 500$), R_i denotes the number of relevant images in returned results for query image i, scope denotes the number of returned images. Scope is varied to observe the performance variations. It should be noted that, since query images are distributed with high visual diversity, search precision in our case can also demonstrate the robustness of the search approaches on dealing with different types of queries.

B. Comparative Experiments

In this section, comparative experiments are conducted to compare the performance of MMHG against state-of-the-art approaches. The performance of MMHG is obtained when $k = 10$, $\gamma = 1.1$, $\lambda = 0.3$, and $T = 10$. Details of the approaches used for comparison are as follows.

1) *Similarity-Based Search With EW (SSEW) [3]:* First, similarities between low-level features of query image and that of database images are calculated in each visual modality. Then, they are integrated into a unified one with equal combination weights. Finally, database images are ranked by comparing the weighted sum of similarities that are calculated in all modalities.
2) *Similarity-Based Search With Proper Weights (SSPW) [50]:* SSPW is similar with SSEW on most of execution procedures, except for the way of combination weight generation. In our implementation, the optimal weights are learned by brute force search.
3) *Manifold-Based Search (MS) [51]:* MS can be regarded as a typical graph-based search. MS leverages simple graph to model the relations among images. In implementation, MMHG Laplacian is substituted with graph Laplacian.
4) *Multifeature Learning via Hierarchical Regression (MLHR) [6]:* MLHR explores discriminative information contained in multiple features of both labeled

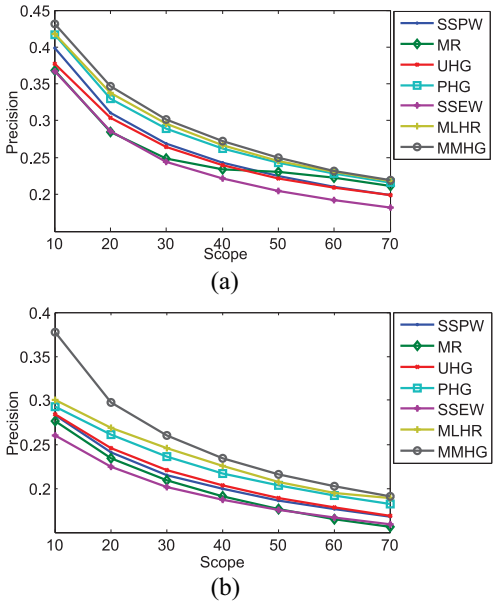

Fig. 5. Performance comparisons with state-of-the-art approaches on landmark-25. (a) Query images are selected from the database. (b) Query images are not selected from the database.

and unlabeled images in framework of semi-supervised graph learning. The optimal combination weights are learned via hierarchical regression. MLHR can be regarded as a state-of-the-art graph learning approach. In implementation, images are ranked with classification scores.

5) *Unified Hypergraph (UHG) [52]:* UHG constructs hypergraph by integrating hyperedges in MMHGs into an UHG, based on which search process is performed.
6) *PHG [31]:* PHG constructs PHG to model the relations among landmark images. It assigns each vertex to hyperedge in probabilistic way. Multimodal features are combined to describe the affinity relations among vertices within each hyperedge.

In this experiment, number of returned results (search scope) is varied from 10 to 70 to compare how different schemes perform. The main experimental results are presented in Fig. 5 and a few interesting observations are gained.

1) MMHG consistently outperforms all the competitors used for comparison on all search scopes. The main reason is that MMHG can represent diverse visual contents of landmark images, perform well on capturing high-order relations among landmark images and the distinctiveness of query images. Detailed discussions are given in Section V-C to illustrate the contribution of each part of MMHG on the final performance.
2) Among the approaches used for comparison, MLHR achieves the best performance, PHG obtains the second best results, and SSEW achieves the worst performance. The good performance of MLHR can be mainly attributed to LRGA, while the good performance of PHG may benefit from the PHG modeling for search process. However, both MLHR and PHG equally deal with multiple features in the process of global alignment or hypergraph construction, which both fail to take advantages of each features and capture the distinctiveness of queries. Therefore, they achieve worse performance than MMHG. For example, when search scope is 10, the performance gap between MMHG and MLHR is 1.26% in Fig. 5(a), while 7.72% is achieved if query images are not from the database [as shown in Fig. 5(b)].
3) As shown, under both settings, SSEW performs worse than SSPW and it even achieves similar performance with UHG on several search scopes. This experimental phenomenon indicates that capturing distinctiveness of features can effectively leverage their discriminative information. It also reveals the fact that the involved features in this paper have different discriminative capabilities on describing the visual contents of landmark images.

TABLE I
DETAILED SEARCH PRECISIONS (%) ACHIEVED ON LANDMARK CATEGORIES BY THE PROPOSED MMHG. SEARCH SCOPE IS SET TO 20. CASE I DENOTES THE SEARCH SCENARIO WHEN QUERY IMAGES ARE INSIDE THE DATABASE. CASE II DENOTES THE SEARCH SCENARIO WHEN QUERY IMAGES ARE OUTSIDE THE DATABASE

Categories	Case I	Case II
Acropolis of Athens	24.50	22.25
Befreiungshalle	39.50	34.50
Big Ben	27.75	23.50
Brooklyn Bridge	36.00	31.50
Buckingham Palace	18.25	16.25
Chrysler Building	25.50	24.25
Citadel of Qaitbay	43.50	37.75
City Wall	27.25	23.25
Dolmabahe Palace	29.75	24.00
Ellis Island Immigration Museum	39.50	31.75
Empire State Building	21.00	17.50
Forbidden City	15.25	13.25
Great Wall	44.25	45.75
Mosta Dome	47.50	37.50
Pantheon	34.50	26.00
St. Patrick's Cathedral	61.50	55.00
Saint Peter Square	32.25	27.00
Golden Gate Bridge	45.75	38.75
St Paul's Cathedral	20.25	17.00
Sultan Ahmed Mosque	48.75	44.25
London Eye	27.50	21.50
Statue Of Liberty	65.50	55.00
Tiananmen Square	36.25	32.75
Tower Bridge	23.00	17.00
Yonghe Temple	32.00	27.00
Mean	34.67	29.77

Table I presents the detailed search precisions achieved on all landmark categories by our proposed MMHG. It can be easily observed from this table that MMHG performs rather differently on landmark categories. For example, in case I (query images are inside database), the search precisions on these landmark categories are widely different. The gap is even more than 50% between the highest precision 65.50% on Statue of Liberty and the lowest precision 15.25% on Forbidden City. The reason to explain this phenomenon is that different landmark categories have different visual diversities. Visual contents of images in some landmark categories could

be highly diverse. Indeed, diverse visual contents of images in a landmark category lead to the bigger challenge in design of corresponding CBVLS algorithm and lower search precision.

C. Discussion

In this section, we present comprehensive analysis and discussion about how various factors influence the performance of MMHG. In particular, we investigate the effects of hypergraph learning, effects of multimodal feature fusion, and effects of weight learning on the overall system performance. Due to limited space here, we only present the experimental results achieved when queries are inside the database. Similar results can also be obtained when queries are outside the database.

1) Effects of Hypergraph Learning: In MMHG, hypergraph is core component, which is employed to model the latent high-order relations among landmark images. This experiment is conducted to validate the effects of hypergraph learning. More specifically, we compare the performance of hypergraph learning with distance learning and graph learning in each visual modality. Details of them are as follows.

1) *Similarity Learning [3]:* Raw feature similarity is used to describe the relations among landmark images. Search results are obtained by ranking the similarities that correspond to the database images in descending order. Similarities between query image and database images are calculated as the similarities between their corresponding visual features. Formally, similarity between query image q and database image i in modality m is calculated as

$$sim(q,i) = exp(-||x_{qm} - x_{im}||_F). \tag{24}$$

2) *Graph Learning [47]:* Graph is used to model the relations among landmark images. Final similarity scores of database images in modality m are calculated by solving the following formula:

$$f = \left(I + \frac{1}{\lambda}\Delta_{GL}^{m}\right)^{-1} y \tag{25}$$

where Δ_{GL}^{m} denotes graph Laplacian, $\Delta_{GL}^{m} = I - \Theta_{GL}^{m} = I - D^{m-1/2} A_{GL}^{m} D^{m-1/2}$. A_{GL}^{m} is affinity matrix, which is calculated as

$$A_{GL}^{m}(i,j) = \begin{cases} exp\left(-\frac{dis(x_{im},x_{jm})}{\sum_{j=1}^{N} dist(x_{im},x_{jm})}\right) \\ dist\left(x_{im},x_{jm}\right) \le k\_dist(x_{im}) \\ 0 \text{ otherwise.} \end{cases} \tag{26}$$

D^m is a diagonal matrix, whose diagonal element is $D_{ii}^{m} = \sum_{j=1}^{m} A_{GL}^{m}(i,j)$.

Fig. 6 presents the performance achieved by three image relation modeling approaches in each single modality when search scope is 20. We can easily draw two conclusions from this figure.

1) Search performance of similarity learning can be improved by graph learning. This is because graph can exploit the relations among database images for searching, while distance learning only leverages the relations between query image and database images.

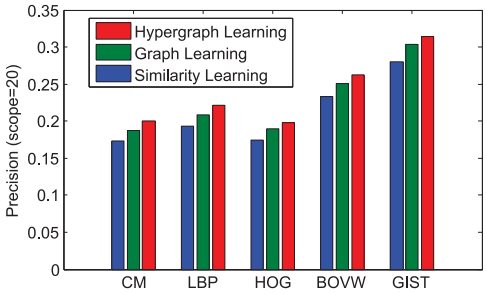

Fig. 6. Performance improvement with hypergraph learning in each visual modality. Search scope is set to 20. Best viewed in color.

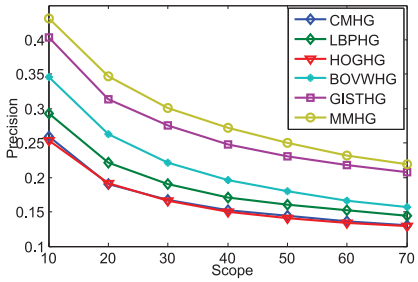

Fig. 7. Performance achieved by different UMHGs. Hypergraphs built in different modalities are complement to each other, and the combination of them can achieve better performance.

2) Hypergraph learning can further improve the performance of graph learning. The reason is that hypergraph can capture high-order relations among landmark images, which cannot be modeled by simple graph.

These experimental results clearly validate the effectiveness of hypergraph learning on describing the high-order relations of images in CBVLS.

2) Effects of Multimodal Feature Fusion: In our approach, discriminative information from multiple visual modalities are integrated into an unified MMHG. In fact, landmark images potentially contain rich heterogeneous information on facets of color, texture, shape, and appearance. These latent information can be characterized by visual features extracted from their corresponding modalities. Heterogenous features may complement each other, and the combination of them may make contributions on improving system performance. In this section, experiment is conducted to investigate the above possibility on MMHG-based CBVLS task. Performance of MMHG and UMHGs are compared directly.

Fig. 7 summarizes the main experimental results obtained by different UMHGs. In these results, "XXHG" denotes that hypergraph is constructed with feature "XX." For example, CMHG denotes that hypergraph is built in color modality with CM feature. From this figure, we can clearly find that MMHG performs better than other UMHGs. Among UMHGs, HOGHG obtains the worst performance. The reason is that landmark images are generally photographed from various viewpoints, which makes images in a landmark category have large visual diversity on facets of shape distribution. In addition, we can find that GISTHG performs better than other UMHGs used for comparison. This is because GIST can capture global information distribution of a particular scene, and it

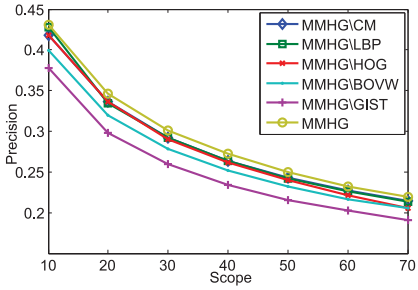

Fig. 8. Performance loss when different features are removed from the construction process of MMHG.

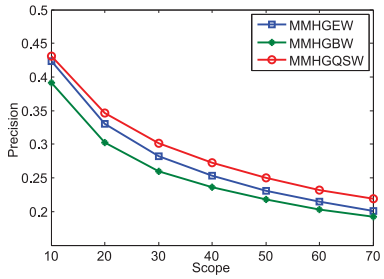

Fig. 9. Performance achieved by different weight generation mechanisms. Our approach can capture query specific combination weights and achieve the best performance.

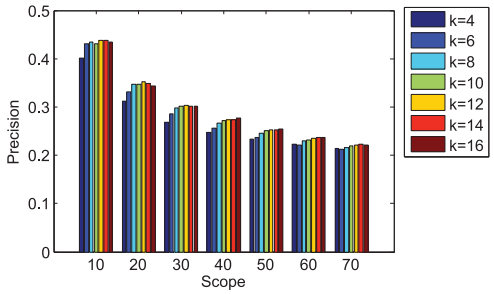

Fig. 10. Performance variations with hyperedge size k. Best viewed in color.

has higher discriminative ability. From the above experimental results, we can draw a conclusion that integrating multiple heterogeneous UMHGs from multiple modalities into an unified MMHG can bring performance improvement in CBVLS.

We further conduct experiments to investigate the contribution of each feature on the final performance of MMHG. Fig. 8 shows the main experimental results when different features are removed from the construction process of MMHG. In these results, "MMHG\XX" denotes the feature XX is removed from MMHG. For example, MMHG\CM denotes CM feature is removed from MMHG construction. In other words, MMHG is constructed with LBP, HOG, BOVW, and GIST. From this figure, we can easily observe the following.

1) Any feature configurations with feature removing generate more or less performance loss. It reveals that the features employed in this paper all contribute to the final performance.
2) Different feature configurations generate different performance loss.

For example, MMHG\GIST brings the maximum performance loss, while MMHG\BOVW brings the second performance loss. MMHG\CM, MMHG\LBP, and MMHG\HOG all bring small performance loss. This phenomenon can be observed and it is because the performance loss is related to the discriminative ability of features. The removal of high discriminative feature brings much more performance loss.

3) Effects of Weight Learning: Weight learning in our approach is used to capture query specific feature combination weights. We maintain that this process is very essential for query images of landmarks are also presented with diverse visual appearances, and learning proper combination weights can capture the characteristics of different queries. To validate our assumption, we compare the proposed weight learning approach with other two weight generation mechanisms. Details are as follows.

1) *MMHG + EW (MMHGEW):* Multiple UMHGs are combined with EW (weight = 0.2). In addition, these combination weights are the same for all queries.
2) *MMHG + Brute Weights (MMHGBW):* The optimal combination weights are obtained via brute searching range of 0 to 1 with step size 0.1. The optimal combination weights for UMHGs are 0.2, 0.2, 0.1, 0.4, and 0.1, respectively. Similar to MMHGEW, these combination weights are the same for all queries.
3) *MMHG + Query Specific Weights (MMHGQSW):* Query specific weights are learned using the approach proposed in this paper to combine UMHGs.

Fig. 9 illustrates search results of different weight learning mechanisms. As shown, our approach consistently outperforms other weight learning mechanisms used for comparison on all search scopes. These presented results demonstrate the effectiveness of our approach on capturing the distinctiveness of the landmark query images. In addition, it is worthwhile to note that MMHGEW performs better than MMHGBW. This experimental phenomenon demonstrates that setting improper weights for all queries even obtain lower search accuracy than exploiting simple weights. In a summary, from the above presented experimental results, we can draw a conclusion that capturing query specific combination weights can improve the performance of MMHG-based CBVLS.

D. Parameter Sensitivity Experiment

In this part of experimental study, we investigate the performance variations with hyperedge size k and the involved parameters T, γ, λ in MMHG.

1) Performance Variations With Hyperedge Size k: In this paper, k denotes the size of hyperedge. At the point of theoretic analysis, with larger k, a hyperedge can contain more images, and more computations will be caused. Also, it will bring negative effects that irrelevant images may be included and noises may be generated accordingly. In contrast, with smaller k, more relevant images may be removed from the hyperedge and descriptive accuracy may be degraded simultaneously.

This experiment is conducted to demonstrate the performance variations with hyperedge size. Fig. 10 presents the performance variations with hyperedge size. It shows that

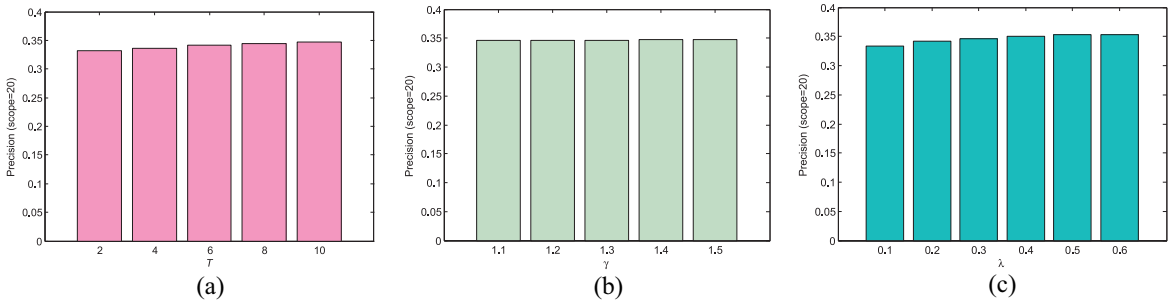

Fig. 11. Performance variations with the involved parameters T, γ, and λ in MMHG when search scope is 20. (a) Search precision increases steadily when T is ranged from 2 to 10. (b) Search precision is stable when γ varies from 1.1 to 1.5. (c) Search precision first increases when λ varies from 0.1 to 0.5, and it becomes steady when the λ is more than 0.5.

search precision increases steadily when the k varies from 4 to 10 on most of scopes, they become stable when hyperedge size is larger than 10. Therefore, we set the hyperedge size to 10 in all our experiments. These presented results reveal that hyperedge learning can perform well on removing noises in larger hyperedges, which are brought in the process of MMHG construction.

2) Performance Test With Various Parameters in MMHG: In this experiment, we fix one of the three parameters and observe the search performance when the other two parameters are varied. Fig. 11 shows performance variations with the involved parameters T, γ, λ in MMHG. The best performance of MMHG is achieved when $\gamma = 1.1$, $\lambda = 0.3$, and $T = 10$. We can obtain the following observations from this figure. For parameter T, search precision increases steadily when T is ranged from 2 to 10. The gap between the lowest and the highest precision (only 0.0141) is small when T varies. For parameter γ, search precision is stable when γ varies from 1.1 to 1.5. For parameter λ, search precision first increases when λ varies from 0.1 to 0.5, and it becomes stable when λ is more than 0.5. Besides, we can easily find that the gap between the lowest and the highest precision (only 0.0194) is also small when λ varies. From the above experimental results, we can clearly find that MMHG is robust to the variations of parameters.

E. Application to General Image Search

We also conduct experimental study using Corel5K [53] to demonstrate the effectiveness of MMHG on task of general image search. Corel5K is collected from Flickr, which consists of 50 categories with 100 images in each category. In our experiment, ten images are randomly selected from each image category to comprise query images, and the remaining is determined as database images. Note that, we explore Corel5K here because it has been used as the benchmark for manifold ranking [49], hypergraph learning [31]. Fig. 12 shows the main experimental results. It clearly shows that our approach consistently outperforms all the competitors, which indicates that MMHG can still perform well on task of general image search. In addition, we find that, on larger scopes, graph-based learning approach (MR and MLHR) can achieve comparable performance compared with hypergraph-based approaches (UHG and PHG). This experimental result

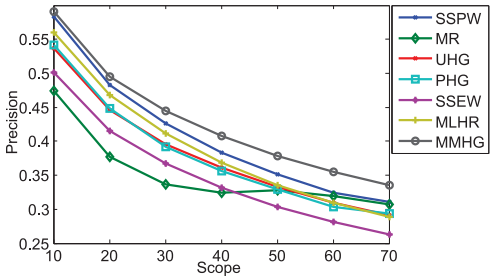

Fig. 12. Performance comparisons with state-of-the-art approaches on Corel5K.

demonstrates that graph can be highly effective and comprehensive for representing the relations among general images.

VI. Conclusion

High visual diversity of landmark images poses big challenges in design of CBVLS system. Several different image views describe the same landmark, and some different landmarks may have several close views. These complex high-order relations among images cannot be represented well by traditional similarity-based search or graph-based search models. This paper explores MMHG to model the high-order relations among landmark images, to describe the diverse visual contents of landmark images, and capture the distinctiveness of queries. For capturing query specific combination weights, search and weight learning are formulated in an unified computational module, which iteratively calculates the best combination weights and similarity scores of database images. Experiments on real-world landmark dataset demonstrate the effectiveness of the proposed approach.

The research opens up three promising directions for future investigation in future study.

1) Integrating discriminative information from textual modality into the current MMHG. Textual feature contains rich semantics that cannot be represented by low-level visual features. Effective combination of textual hypergraph and visual hypergraph could boost the performance further.
2) On efficiency side, integrating MMHG with hashing to further accelerate search process. Hashing is a hot topic in recent literature, which performs well on feature

indexing [25]. MMHG-based hashing will not only have strong describing capability, but also have high search efficiency.

3) Extending the MMHG to solve other tasks, such as landmark classification, where we will face similar challenges.

APPENDIX

We proof the convergency of Algorithm 1

$$y^t = \left(\frac{\lambda}{1+\lambda}\right)\sum_{i=0}^{t-1}\left(\frac{1}{1+\lambda}(I-\Delta)\right)^i \bar{y} + \left(\frac{1}{1+\lambda}(I-\Delta)\right)^t \bar{y}.$$

Since the eigenvalues of $I-\Delta$ are $[1,-1]$, we obtain that

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}\left(\frac{1}{1+\lambda}(I-\Delta)\right)^i \bar{y} = \left(I-\frac{1}{1+\lambda}(I-\Delta)\right)^{-1} = \frac{1+\lambda}{\lambda}\left(I+\frac{1}{\lambda}\Delta\right)^{-1}\lim_{t\to\infty}\left(\frac{1}{1+\lambda}(I-\Delta)\right)^t \bar{y} = 0.$$

Therefore, we can derive that

$$\begin{aligned} y = \lim_{t\to\infty} y^t &= \left(\frac{\lambda}{1+\lambda}\right)\lim_{t\to\infty}\sum_{i=0}^{t-1}\left(\frac{1}{1+\lambda}(I-\Delta)\right)^i \bar{y} \\ &\quad + \lim_{t\to\infty}\left(\frac{1}{1+\lambda}(I-\Delta)\right)^t \bar{y} \\ &= \left(\frac{\lambda}{1+\lambda}\right)\left(\frac{1+\lambda}{\lambda}\right)\left(I+\frac{1}{\lambda}\Delta\right)^{-1}\bar{y} + 0 \\ &= \left(I+\frac{1}{\lambda}\Delta\right)^{-1}\bar{y}. \end{aligned}$$

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive and helpful suggestions.

REFERENCES

[1] J. Shen, Z. Cheng, J. Shen, T. Mei, and X. Gao, "The evolution of research on multimedia travel guide search and recommender systems," in *Advances in Multimedia Modeling*. Cham, Switzerland: Springer, 2014, pp. 227–238.

[2] Q. Fang, J. Sang, and C. Xu, "GIANT: Geo-informative attributes for location recognition and exploration," in *Proc. ACM Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 13–22.

[3] Z. Cheng, J. Ren, J. Shen, and H. Miao, "Building a large scale test collection for effective benchmarking of mobile landmark search," in *Advances in Multimedia Modeling*. Berlin, Germany: Springer, 2013, pp. 36–46.

[4] E. Gavves and C. G. Snoek, "Landmark image retrieval using visual synonyms," in *Proc. ACM Int. Conf. Multimedia*, Florence, Italy, 2010, pp. 1123–1126.

[5] C. Deng, R. Ji, D. Tao, X. Gao, and X. Li, "Weakly supervised multi-graph learning for robust image reranking," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 785–795, Apr. 2014.

[6] Y. Yang *et al.*, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 572–581, Apr. 2013.

[7] X.-C. Xu, X.-S. Xu, Y. Wang, and X. Wang, "A heterogenous automatic feedback semi-supervised method for image reranking," in *Proc. ACM. Int. Conf. Inf. Knowl. Manage.*, San Francisco, CA, USA, 2013, pp. 999–1008.

[8] 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.

[9] X. Wu, A. G. Hauptmann, and C. Ngo, "Measuring novelty and redundancy with multiple modalities in cross-lingual broadcast news," *Comput. Vis. Image Und.*, vol. 110, no. 3, pp. 418–431, 2008.

[10] J. Shen, "Stochastic modeling western paintings for effective classification," *Pattern Recognit.*, vol. 42, no. 2, pp. 293–301, 2009.

[11] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, and A. G. Hauptmann, "Event detection using multi-level relevance labels and multiple features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 97–104.

[12] Y. Ren, M. Yu, X.-J. Wang, L. Zhang, and W.-Y. Ma, "Diversifying landmark image search results by learning interested views from community photos," in *Proc. Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 1289–1292.

[13] J. Ye *et al.*, "DLMSearch: Diversified landmark search by photo," in *Proc. ACM Int. Conf. Multimedia*, Nara, Japan, 2012, pp. 905–908.

[14] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2012.

[15] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nice, France, 2003, pp. 1470–1477.

[16] Q. Hao *et al.*, "3D visual phrases for landmark recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 3594–3601.

[17] X. Xiao, C. Xu, J. Wang, and M. Xu, "Enhanced 3-D modeling for landmark image classification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1246–1258, Aug. 2012.

[18] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly, "Structure from motion for scenes with large duplicate structures," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 3137–3144.

[19] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. ACM Int. Conf. World Wide Web*, 2008, pp. 297–306.

[20] T. Chen and K. Yap, "Discriminative BoW framework for mobile landmark recognition," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 695–706, May 2014.

[21] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multi-task spectral clustering by exploring inter-task correlation," *IEEE Trans. Cybern.*, to be published.

[22] P. Qian, F. Chung, S. Wang, and Z. Deng, "Fast graph-based relaxed clustering for large data sets using minimal enclosing ball," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 672–687, Jun. 2012.

[23] X. Geng, D.-C. Zhan, and Z.-H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 6, pp. 1098–1107, Dec. 2005.

[24] Y. Zhang, K. Huang, X. Hou, and C. Liu, "Learning locality preserving graph from data," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2088–2098, Nov. 2014.

[25] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[26] Y. Yang *et al.*, "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.

[27] S. Jones and L. Shao, "A multigraph representation for improved unsupervised/semi-supervised learning of human actions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 820–826.

[28] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn.*, vol. 25, no. 12, pp. 2303–2308, Dec. 2014.

[29] Y. Huang, Q. Liu, and D. N. Metaxas, "Video object segmentation by hypergraph cut," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1738–1745.

[30] L. Zhang *et al.*, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

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

[32] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[33] C. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Int. Conf. Multimedia*, Singapore, 2005, pp. 399–402.

[34] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. ACM Int. Conf. Multimedia*, New York, NY, USA, 2004, pp. 572–579.
[35] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[36] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2009, pp. 606–613.
[37] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. Int. Conf. Mach. Learn.*, Kyoto, Japan, 2004, p. 6.
[38] Z. Ma *et al.*, "Complex event detection via multi-source video attributes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2627–2633.
[39] S. Wang *et al.*, "Semi-supervised multiple feature analysis for action recognition," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 289–298, Feb. 2014.
[40] Z. Xu, Y. Yang, I. Tsang, N. Sebe, and A. G. Hauptmann, "Feature weighting via optimal thresholding for video analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 3440–3447.
[41] F. Mindru, T. Tuytelaars, L. V. Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Comput. Vis. Image Und.*, vol. 94, nos. 1–3, pp. 3–27, 2004.
[42] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.
[43] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 886–893.
[44] S. Wei, D. Xu, X. Li, and Y. Zhao, "Joint optimization toward effective and efficient image search," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2216–2227, Dec. 2013.
[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[46] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[47] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2003, pp. 321–328.
[48] X. Feng, H. Jin, R. Zheng, Z. Shao, and L. Zhu, "A segment-based sparse matrix-vector multiplication on CUDA," *Concurr. Comput. Pract. Exp.*, vol. 26, no. 1, pp. 271–286, 2014.
[49] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Generalized manifold-ranking-based image retrieval," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3170–3177, Oct. 2006.
[50] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, 2008, pp. 1–8.
[51] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.
[52] J. Xu, V. Singh, Z. Guan, and B. Manjunath, "Unified hypergraph for image ranking in a multimodal context," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, 2012, pp. 2333–2336.
[53] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, 2002, pp. 97–112.