Pattern Recognition 44 (2011) 2263–2273

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Latent visual context learning for web image applications

Wengang Zhou [a], Qi Tian [b,*], Yijuan Lu [c], Linjun Yang [d], Houqiang Li [a,*]

[a] University of Science and Technology of China, Hefei, China
[b] University of Texas at San Antonio, San Antonio, TX, USA
[c] Texas State University, TX, USA
[d] Microsoft Research Asia, Beijing, China

ARTICLE INFO

Available online 14 August 2010

Keywords:
Visual context
Image re-ranking
Canonical image selection
Set coverage

ABSTRACT

Recently, image representation based on bag-of-visual-words (BoW) model has been popularly applied in image and vision domains. In BoW, a visual codebook of visual words is defined, usually by clustering local features, to represent any novel image with the occurrence of its contained visual words. Given a set of images, we argue that the significance of each image is determined by the significance of its contained visual words. Traditionally, the significances of visual words are defined by term frequency-inverse document frequency (tf-idf), which cannot necessarily capture the intrinsic visual context. In this paper, we propose a new scheme of latent visual context learning (LVCL). The visual context among images and visual words is formulated from latent semantic context and visual link graph analysis. With LVCL, the importance of visual words and images will be distinguished from each other, which will facilitate image level applications, such as image re-ranking and canonical image selection.

We validate our approach on text-query based search results returned by Google Image. Experimental results demonstrate the effectiveness and potentials of our LVCL in applications of image re-ranking and canonical image selection, over the state-of-the-art approaches.

1. Introduction

The last decade has witnessed the proliferation of images on the web, which casts a great challenge on exploration of gigantic amount of images. Currently, image retrieval has become a popular service in many search engines, such as Bing, Google and Yahoo! However, most of them are mainly based on textual information. This is partly due to the fact that text-based search techniques are successful and mature, while image content information is difficult or expensive to exploit. Recently, a novel image identification based commercial search engine Tineye [1] has been attracting more and more attention, with real-time response for a billion scale database. However, it is designed only for duplicated image search.

For text-based image search, two problems in terms of relevance and diversity are usually encountered. With the ignorance of visual content information, the returned image search results usually suffer from poor relevance, such as noisy and cluttered irrelevant images. To address this issue, image re-ranking has become an active research topic in multimedia community. The goal of image re-ranking is to refine the text-based image search results according to their visual content consistency, such that the relevant images are moved to the top while irrelevant images sink to the bottom, resulting in better relevance, especially for top-ranked images. Further, although image re-ranking will enhance the relevance of top-ranked images to the text query, image redundancy, such as duplicated or very similar images, is still prevalent. If canonical images, a subset of photos that best summarize a photo collection, can be selected so as to compress the image redundancy, then the user experience of image exploration will be greatly improved.

In this paper, we propose a new scheme of latent visual context learning (LVCL) to explore the visual context among images and visual words to address the above two problems. With bag-of-visual-words model (BoW) [12] for image representation, visual context expresses the intrinsic relationships among images and visual words, which is related with many different aspects, such as statistical distribution of visual words in images, the co-occurrence of visual words, the geometric relationships of visual words, the latent semantics underlying image and visual words, and so on. Traditionally, visual words are weighted by term frequency-inverse document frequency (tf-idf), which can be regarded as only statistical visual context. As in LVCL, we focus on mining those visual contexts from latent semantic analysis and visual link graph analysis in a latent manner. The discovered visual context information will be used to discover the significance of visual words and images for web image applications.

* Corresponding authors.
E-mail addresses: qitian@cs.utsa.edu (Q. Tian), lihq@ustc.edu.cn (H. Li).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.08.016

In BoW, with a visual codebook and a collection of images from text query, we argue that the significance of each image is determined by the significance of its contained visual words, while the significances of different visual words are related by the corresponding concept of the text query. Therefore, we first focus on visual word level and then proceed to image level to explore the visual context. The pair-wise similarities for visual words and images are formulated, respectively. And then, by means of graph analysis, the significance for visual words and images will be distinguished, respectively.

The LVCL for visual words is graph based. We construct a visual graph for visual words, where the node denotes visual word and the edge that links two nodes are weighted by their similarity. The visual similarity is formulated with latent semantic context and visual word link context. To explore the implicit semantic context, we adopt latent semantic analysis [4] to discover the implicit semantic similarity for pair-wise visual words. To analyze the visual link context, we construct a four-layer visual word link graph, with MSER [3] as a latent layer to impose local spatial constraint for visual words. Then based on the visual word graph, similar to [8], PageRank [9] is used to explore the visual significance for all visual words.

As for LVCL for images, we construct a three-layer visual link graph for images. In the image link graph, visual word works as a latent layer and the image links are weighted by the significance of visual words. The visual context in the image graph is analyzed by exploring the visual transition relationship between image and visual words. After obtaining the visual graphs for images, PageRank [9] is also used to explore the visual significance for all images.

The application of LVCL to image re-ranking is intuitive. According to their discovered visual content significance, we obtain the content based ranking of images, which can further be fused with the text based ranking of the initial image search results returned via text query. As for canonical image selection, we define the representative photos as those that contain many most important and distinctive visual words and propose to select the canonical images via weighted set coverage strategy—greedily discover images that contain those important visual words from a candidate image pool. In order to construct a good candidate image pool and filter some noisy images, we also adopt the image re-ranking results and select the top ones for canonical image selection.

As a summary, the main contributions of this paper include:

(1) Propose a novel scheme to explore latent visual context from both latent semantic and visual link graphs. The significance of visual words and images for an image collection will be discovered by means of graph analysis.
(2) Apply LVCL for image re-ranking and canonical image selection. The resulting content based re-ranking of images can be fused with text based image ranking information. Canonical image selection is performed by a set coverage strategy with LVCL results.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 gives a brief overview of our framework of latent visual context learning for web images applications. In Sections 4 and 5, we will discuss LVCL for visual words and images, respectively. Section 6 provides the applications of LVCL in image re-ranking and canonical image selection. Section 7 presents experimental results to evaluate our approach. Finally, conclusions are made in Section 8.

2. Related work

Our work is related to several research topics, including visual re-ranking, topic model, link graph analysis and MSER detection, and canonical image selection. The related literature is briefly reviewed below.

In literature, there are many works about visual re-ranking based on different schemes, such as clustering-based [15,21], classification-based [19] and the graph-based [8,16–18,20,22,24]. Generally, an assumption is made that visually similar images should be ranked close to each other. In all visual re-ranking methods, an essential problem is how to measure the visual similarity. Currently, the similarity is mainly estimated based on low-level visual features: global features [17,18,24], such as color moments and Gabor feature, and local features [8,20,22,24], such as scale invariant feature transform (SIFT) [2]. Global features work well for cases such as natural scene images, while local features do good job in rigid canonical object images. As the state-of-the-art approach, VisualRank [8] builds an image graph and intuitively determines the pair-wise image similarity by the number of shared SIFT features and computes the image rank value directly through an iterative procedure similar to PageRank [9]. In fact, there is an underlying assumption in VisualRank that all matched local features in image are equally important. In fact, given an image set returned by text-based image search engine, some local features are expected to be more discriminative than others. Therefore, it is preferred to give these discriminative features a larger weight. In our approach, we investigate the visual word significance by analyzing both latent semantic context and visual word link context, to infer image significance.

Recently, many research efforts have been made to select canonical images, a subset of photos that best summarize the image collections. Jaffe et al. [25] used a pure metadata-driven approach to select canonical views for a region with both geo-locations and several heuristics on metadata. Since no visual content information is considered, the selected views are sensitive to noise. Some other research works use clustering-based approaches to discover canonical images. Raguram and Lazebnik [26] proposed to compute iconic views for a collection of images of any concept with joint clustering based on visual and textual features to extract subsets of images. Iconic views were chosen from images with the highest visual quality from each subset. Kennedy et al. [29] leveraged both metadata and visual features to form a hybrid approach for canonical view selection. They discovered landmark related images and performed k-means clustering based on color and texture features. Then canonical views were selected as top-ranked images from top-ranked clusters. These clustering based methods unavoidably suffer from the issue of determining the number of clusters, which greatly constrains their flexibility. Simon et al. [28] proposed a pure vision-based approach. The original photo collection was partitioned into several non-overlapping subsets. Then, a greedy k-means clustering was adopted to select canonical views from each subset. However, like other clustering methods, the greedy k-means is still sensitive to outliers, which are prevalent in web image collections. Yang et al. [27] proposed a canonical view selection method based on online search results. They analyzed the distribution of visual words [13] and computed a coverage score for each photograph. 200 distinctive visual words were selected by wc-tf-idf measurement and a greedy scheme was proposed to iteratively select a few canonical views. However, wc-tf-idf cannot necessarily capture the visual word latent similarity relationship, and its effectiveness may be weakened.

Our work is based on bag of visual words (BoW) model [12–14]. BoW is originally derived from natural language processing and popularly applied in image and vision domains. Generally, the bag-of-visual-words representation defines a visual codebook by clustering local features [7,13], such as SIFT, into k groups. Each resulting cluster centroid is considered as a visual word and all k visual words constitute a visual codebook. With bag-of-visual-word model, each

local feature from an image is assigned to the closest visual word. Consequently, an image is represented as a histogram of the assignment of all local features to visual words and a k-dimensional vector is obtained, which is subsequently normalized. Traditionally, the significances of visual words are defined by term frequency-inverse document frequency (tf-idf). However, with the ignorance of the intrinsic context among visual words such as semantic context and spatial context, tf-idf cannot necessarily distinguish the significance of visual words. Our proposed latent visual context learning will naturally address such drawback.

Based on BoW, many topic models, such as latent semantic analysis (LSA) [4], Probabilistic latent semantic analysis (pLSA) [5] and latent Dirichlet analysis (LDA) [6] can be used to analyze the topics within images. As generative data models, pLSA and LDA are based on statistical foundation and work with the number of latent topics determined beforehand. LSA, instead, is based on singular value decomposition (SVD) to explore the higher-order semantic structure in an implicit manner without knowing the latent topic number. In this paper, LSA is adopted to explore the underlying implicit semantic context in conjunction with visual words and to generate visual word similarity in latent semantic sense.

Our work is partly motivated by Cai et al. [10], where the webpage is segmented into different unique blocks and the webpage link analysis is converted to multi-layer link graph analysis, so as to better explore the semantic topics of the web page. As for our problem, we regard the context relationship between image and visual word as visual hyperlinks and construct visual word link graph and image link graph, respectively. Then, the visual significance discovery for visual words and images is fulfilled by analyzing the corresponding visual graphs. In [31,32], link analysis techniques are also used for recognition of object categories. Without quantizing local features to visual words, it constructs a visual similarity network (VSN), where the nodes are features extracted from all images and the edges link features that have been matched across images. The weight of the graph edge reflects the correspondence consistency and is obtained with a spectral technique [30]. Both Kim's link analysis approach and our method are based on graph link analysis, but the link relationship definitions are greatly different from each other. Although good performance is achieved in object categorization, it ignores the local geometric constraints of features in images. Besides, without feature quantization, it will not suffer quantization error, but the time cost is expensive.

Our work is also related to [11], as both make use of MSER region [3] to impose local geometric constraint. The difference is that, [11] exploits MSER to bundle local features to improve the discriminative power of visual word, while in our approach MSER is adopted as an intermediate to formulate the visual hyperlink context among visual words. As a latent layer in our visual word hyperlink graph, MSER region is used such that local features will be related with only those sharing the same MSER region in the image.

As for canonical image selection, our work is related to [27] in the adoption of set coverage. In [27], a predefined number of distinctive visual words were selected by *wc-tf-idf* for a greedy scheme to iteratively select a few canonical views. As for our approach, important visual words are adaptively chosen with LVCL. Besides, the weighting schemes of visual words and images for set coverage are all derived from LVCL results.

3. Framework overview

Fig. 1 illustrates the framework of our latent visual context learning for web image applications. We first collect images with text query from text-based image search engine, such as Google Image. Then, two kinds of features, SIFT [2] and MSER [3], and extracted for each image. SIFT features are quantized to visual words with a predefined visual codebook. After that, we begin to perform latent visual context learning (LVCL).

LVCL mainly consists of two parts: visual word context analysis and image context analysis. Visual word context analysis is carried out from two perspectives, namely, latent semantic topic context and visual word link graph. Then visual word rank is performed to discover visual word significance, which is adopted for image context learning with image link graph analysis.

Finally, with the obtained visual context information of visual words and images, we will apply it to web image applications, including image re-ranking and canonical image selection.

4. Visual word context analysis

Visual word context is learned through visual word graph analysis. We construct a visual word graph, where the node denotes visual word and the edge linking two nodes is weighted by their similarity. Then the problem becomes how to define the pair-wise visual word similarity. We propose to formulate it from two perspectives: latent semantic topic and visual word link graph. After that, random walk [8,9] can be employed to discover the significance of the graph nodes.

4.1. Visual word similarity decomposition

Visual similarity is related with human psychological cognition, which is a very complex process for simulation. In this paper, we approach it from two kinds of visual context. The first one is related with latent semantics analysis, and the second one is about visual link graph.

Before re-ranking, it is necessary to explore the pair-wise relationships for visual words, in other words, the similarity between visual words. We propose to formulate the similarity definition for visual word pair (i, j) as follows:

$$W^{VW}(i,j) = \alpha W_s^{VW}(i,j) + (1-\alpha) W_g^{VW}(i,j) \quad (1)$$

where W_s^{VW} is formulated with latent semantics related visual context, W_g^{VW} is defined with visual link graph related context and α is a weighting factor with range $0 < \alpha < 1$. In the following subsections, we will explain the formulation of the above two decomposed similarity components for visual words.

4.2. Latent semantic similarity analysis

Since our image collection is returned from text query retrieval, it is reasonable to assume that there exist some topics among these images. The latent semantic context beneath the visual word–image relationship can be explored by means of semantic model. Generally, image topic is difficult to be represented explicitly. Also, this is not necessary for our case. LSA, originally proposed for text indexing and retrieval and proved to be powerful for discovering the implicit higher-order context in the association of terms with documents [4], can serve for our task.

According to LSA, given a visual word–image matrix M_0 with size $m \times n$, each column of which is a normalized histogram of visual word occurrence in the corresponding image, it can be decomposed into the product of three other matrixes by singular value decomposition (SVD) as follows:

$$M_0 = T_0 S_0 D_0^T \quad (2)$$

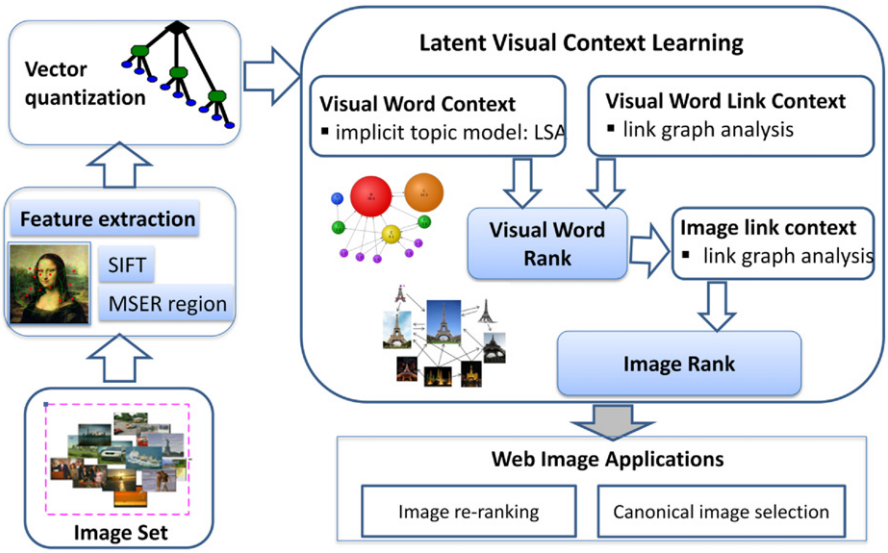

Fig. 1. The proposed framework of latent visual context learning for web image applications.

where T_0 and D_0 are column-orthonormalized matrices and S_0 is a diagonal matrix with all diagonal elements positive and in decreasing order. The sizes of T_0, S_0 and D_0 are $m \times k$, $k \times k$ and $n \times k$, respectively. The beauty of SVD is that it provides a simple strategy for optimally approximate fit using smaller matrices. To maintain the real data structure and at the same time ignore the sampling error or unimportant details, only the top $t(t < k)$ largest diagonal elements in S_0 are kept while the remaining smaller ones are set to zero. This is equivalent to delete the zero rows and columns of S_0 to obtain a compact matrix S and delete the corresponding columns of T_0 and D_0 to yield T and D, respectively. Consequently, a reduced matrix M is defined as follows:

$$M = T\,S\,D^T \tag{3}$$

which is the rank-t model with the best possible least squares-fit to M_0 [4]. Geometrically, the rows of the reduced matrices, i.e., T and D, can be regarded as coordinates of points representing the images and visual words in a t-dimensional space.

The amount of dimension reduction of S_0, i.e., the value of t, is a significant issue and is usually determined by operational criterion. If t is too large, M will be sensitive to noise. On the other hand, if t is too small, the latent structure may not be kept. Therefore, some trade-off should be made. In our implementation, we empirically find best results with $t = \min(i | S_0(i,i) > S_0(0,0)/30)$.

Assume row-normalizing M yields M^R. Then the dot product between two row vectors of M^R reflects the extent to which two terms have a similar pattern of occurrence across the set of images [4]. Therefore, the pair-wise row vector distance of Y^R can be defined as

$$U = M^R (M^R)^T \tag{4}$$

Each entry in U is a cosine distance between two normalized vectors, with value ranging from -1 and 1. We keep those similar visual word pairs and discard those dissimilar ones. We define the pair-wise visual word similarity in the sense of latent semantic context as follows:

$$W_s^{VW}(i,j) = h(U(i,j)) \tag{5}$$

where $h(\cdot)$is a non-decreasing function.

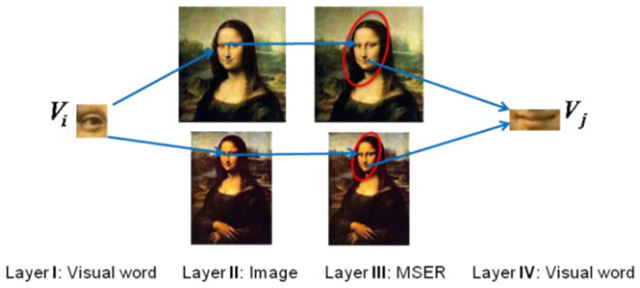

Fig. 2. An illustration of four layers between two nodes in the visual word link graph. The red ellipses in layer 3 denote detected MSER region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Visual word link graph

Visual word is a kind of visual concept atom. They are intrinsically related through image and works as visual concept carrier. Usually, visual concept is composed of a set of visual atoms under some geometric constraints. Therefore, it is necessary to incorporate the local geometric relationship among visual words to analyze the visual link context between visual words. MSER region [3] can serve for this task.

Usually, the MSER detector generates a relatively small number of regions per image with high repeatability. Such MSER region intrinsically imposes local geometric constraints for its contained visual words, which works as geometric context.

Consequently, a four-layer visual word graph is constructed. As illustrated in Fig. 2, there are two intermediate layers, i.e., image layer and MSER layer. Visual words do not transit to each other directly. Instead, a visual word V_i first transits to an image that contains V_i, then further to the MSER region in the image, and finally to another visual word V_j that shares the same MSER region. The intuition behind is that if a user is viewing an image, he or she is most likely attracted by some local features, and other local features within the neighborhood may also be of interest. Such transition behavior naturally reflects the co-occurrence context of visual words.

Based on the visual word graph, we essentially define a propagation probability matrix W on the edges of the graph as follows:

$$P(V_j|V_i)=\sum_{k=1}^{N}\sum_{t=1}^{N_k}P(V_j|M_{k,t})P(M_{k,t}|I_k)P(I_k|V_i) \quad (6)$$

where N denotes the total number of images to be ranked, N_k denotes the number of MSER regions in the kth image, $M_{k,t}$ denotes the set of visual words in the tth MSER region in the kth image, $P(V_j|M_{k,t})$ is defined as the normalized term-frequency of the visual word V_j in $M_{k,t}$; $P(M_{k,t}|I_k)$ is defined as the normalized MSER-frequency of $M_{k,t}$ in image I_k. $P(I_k|V_i)$ is defined as the inverse image frequency of visual word V_i for image I_k: $P(I_k|V_i)=1(/N(V_i))$, where $N(V_i)$is the total number of images that contain visual word V_i.

Further, we simply define the visual link similarity for visual words as follows:

$$W_g^{VW}(i,j)=P(V_j|V_i) \quad (7)$$

4.4. Visual word rank

After obtaining W_s^{VW}andW_g^{VW}, W^{VW} is calculated with Eq. (1). Similar to VisualRank [8], the idea of PageRank [9] can be adopted to discover the visual word significance. Consequently, the significance of visual word R is iteratively defined as follows:

$$R=d\,W^*\,R+(1-d)p,\ \text{where}\ p=\left[\frac{1}{n}\right]_{n\times 1} \quad (8)$$

where W_* is the column-normalized version of the transposition of W^{VW}, p is a distracting vector for random walk behavior [8], and d is a constant damping factor. Usually, $d>0.8$ is chosen. The iteration of Eq. (8) is considered converged when the change of R is small enough or a maximal iteration number, such as 100, is achieved.

5. Image context analysis

Since images are represented by visual words, image visual context can also be deduced from the significance of visual words. We explore image context through image link graph with visual word as a latent layer, and formulate the weight of the graph edge with the visual word significance. Then, image significance is also discovered with random walk re-ranking method [8,9].

5.1. Image link graph

Generally, images are related through intermediate medium of visual words, which work as visual hyperlinks. And, different visual words will cast different votes to the image that contains them, according to their significance. These context relationships can be represented with a graph of three layer as illustrated in Fig. 3. Instead of direct transition, an image first transmits to a contained visual word, and then to another image that shares the same visual word. Visual word plays a role of latent layer in the image link graph.

Intuitively, it is more preferred to generate a middle-layer image graph, similar to the visual word graph in Section 4.3, with visual word set in MSER region as a latent layer. However, this involves extracting representative MSER visual word sets by clustering potential samples of these sets. Due to the imperfect repeatability of local features and quantization error induced by BoW, the potential variants of similar MSER visual word sets is numerous, and summarizing item-set pattern for MSER visual word set will be an unrealistic task. Therefore, we avoid the MSER layer and keep only three layers.

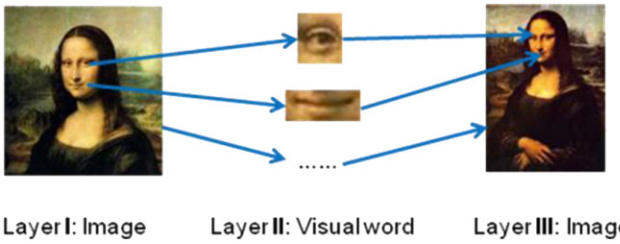

Fig. 3. An illustration of three layers between two nodes in the image link graph.

Based on the above discussion, the image transition probability from image I_i to image I_j can be defined as follows:

$$P(I_j|I_i)=\frac{1}{N_i}\sum_{V_k\in I_j,I_i}P(I_j|V_k)P(V_k|I_i)f(R_k) \quad (9)$$

where $P(I_j|V_k)$ is defined as the inverse image frequency of visual word V_k in image I_j, $P(V_k|I_i)$ denotes the normalized term frequency of visual work V_k in image I_i, R_k is the significance value of visual word V_k obtained with Eq. (8), $f(\cdot)$ is a non-decreasing function and N_i is a normalization factor such that the sum of transition probability for the ith image to any other image is one.

In Eq. (9), the item $f(R_k)$ weights visual link that passes through visual word V_k. Therefore, images with many significant visual words will enjoy high probability to be propagated to. There are many choices for $f(\cdot)$, such as $f(x)=e^x$,$f(x)=x^r$. In our approach, we define $f(\cdot)$ by Eq. (10) and determine the value of r through experimental study.

$$f(x)=x^r \quad (10)$$

It should be noted that the image transition probability matrix is asymmetric. In fact, the symmetry property is not required. For instance, consider the case in which one image is a cropped version of another image. It is obvious that the alternate conditional probability is unequal.

5.2. Image rank

In Section 5.1, the image transition probability is obtained. However, it does not necessarily define the image pair-wise similarity, since the more features regardless of importance an image contains, the larger probability it is propagated from other images. Therefore, a regularization term should be included for penalizing images with too many features. We formulate the image visual similarity heuristically as follows:

$$W^{Img}(i,j)=Prob(I_j|I_i)\tau(I_j) \quad (11)$$

where $\tau(I_j)$ is the regularization term for the jth image, defined as

$$\tau(I_j)=\frac{1}{\sqrt{N(I_j)+\overline{N}/n}} \quad (12)$$

where $N(I_j)$ denotes the number of features in the jth image, $\overline{N}$ is the average local feature number per image, n is a constant. In our experiment, $n=10$. In Eq. (12), the term $\overline{N}/n$ works as a residue to prevent over-weighting those images with too few local features.

Similar to visual word rank discussed in Section 4.4, we can also adopt the idea of PageRank to explore the image significance. In PageRank, each graph node is a web page, while in our task an image corresponds to a node in the graph. And the key lies in the

definition of the similarity between each pair of nodes in the graph. Consequently, the significance of image S^{Img} is iteratively defined as follows:

$$S^{Img} = d\,U\,S^{Img} + (1-d)p, \text{ where } p = \left[\frac{1}{n}\right]_{n\times 1} \quad (13)$$

where U is the column-normalized version of the transposition of W^{Img} defined in Eq. (11). The convergence condition is the same as that of Eq. (8), as discussed in Section 4.4.

6. Web image applications

In this paper, based on the results from latent visual context learning as discussed from the above two sections, we focus on two applications for web images: image re-ranking and canonical image selection.

6.1. Image re-ranking

Image re-ranking is to re-order images based on the relevance of images to the given text concept. Generally, the image relevance can be fused with two kinds of clues. The first one is the original ranking order of the image search results, which is mainly based on textual metadata of images. The second one is from the image significance based on visual content, as that discussed in Section 5.2. Consequently, we define the image rank value as follows:

$$g(I_i) = \lambda O^{Content}(I_i) + (1-\lambda)O^{Text}(I_i) \quad (14)$$

where $O^{Content}(I_i)$ is the re-ranking order of image I_i according to the visual significance S^{Img} obtained with Eq. 13, $O^{Text}(I_i)$ denotes the ranking order of image I_i in the original search results from text based image search engine and λ is a weighting factor with range $0<\lambda<1$.

Finally, all images are re-ordered according to the rank value $g(I_i)$. It should be noted that for a constant λ, some images with different $O^{Content}(I_i)$ and $O^{Text}(I_i)$ may also get the same rank value. In such cases, images with greater value of $O^{Content}(I_i)$ will be given more priority to rank to the top.

6.2. Canonical image selection

With the selected distinctive visual words and good candidate images, we can perform canonical image selection. Ideally, the canonical images should be representative to the query, and also exhibit a diverse set of views. In [27], three criteria are proposed, including: similarity, coverage, and orthogonality. In this paper, we define the representative photos as those that contain many most important and distinctive visual words and adopt a weighted set coverage (WSC) scheme to select multiple search-related canonical images, considering orthogonality and coverage.

We denote the candidate image pool generated in Section 5 as $S=\{I_1,I_2,...,I_m\}$ and the set of distinctive visual word elements contained in these selected images as $X=\{v_1,v_2,...,v_n\}$. Each visual word v_i has a weight $w(v_i)$, and each image I_i has an incremental weight c_i and an importance weight u_i. u_i is defined to be equal to $VR^{img}(I_i)$, the significance value of I_i, obtained from Eq. (13). The definitions of $w(v_i)$ and c_i are described as follows:

$$w(v_i) = \begin{cases} \sqrt{VR^{vw}(v_i)} \text{ if } VR^{vw}(v_i) > \bar{s} \\ 0 \text{ else} \end{cases} \quad (15)$$

where $\bar{s}$ is the average of all visual word significance value $VR^{vw}(v_i)$, (i=1,2,...,n), obtained from Eq. (8)

$$c_i = \sqrt{\frac{\#(VW(I_i \backslash G) \cap X)}{\#(VW(I_i))}} \quad (16)$$

where G is a subset of S, representing the selected canonical images. $VW(I_i \backslash G)$ denotes the set of distinctive visual words only covered by image I_i but not covered by any image in G, and $\#(VW(I_i))$ denotes the total number of visual words covered by I_i.

Further, we define TW_i as the total sum of weight of the distinctive visual words in $VW(I_i \backslash G)$:

$$TW_i = \sum_{v_i \in VW(I_i \backslash G)} w(v_i) \quad (17)$$

Finally, our weighted set coverage is performed as Algorithm 1.

Algorithm 1: Weighted Set Coverage (G, S, X)
Input: S, a set of images; X, a set of visual word elements.
Output: G, a subset of S with maximum coverage of good visual words.
Procedure:
1. $G=\Phi$.
2. Repeat
(1) Select $I_i \in S$, that maximize $TW_i\, c_i\, u_i$;
(2) $G \leftarrow G \cup \{I_i\}$, $S \leftarrow S \backslash \{I_i\}$
(3) If all Good visual words are covered by G, stop.

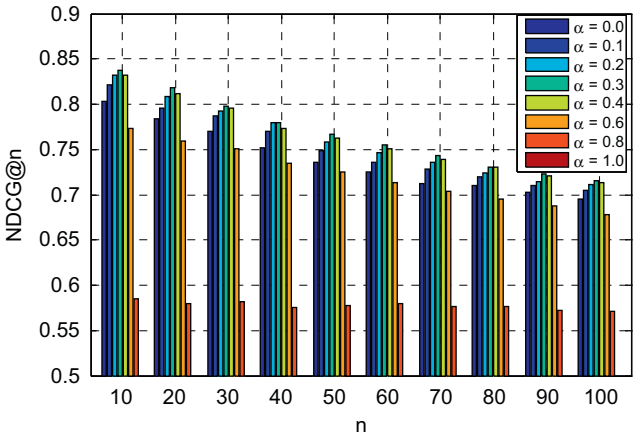

Fig. 4. Performance (NDCG) for different values of α. Best performance is obtained with $\alpha=0.3$.

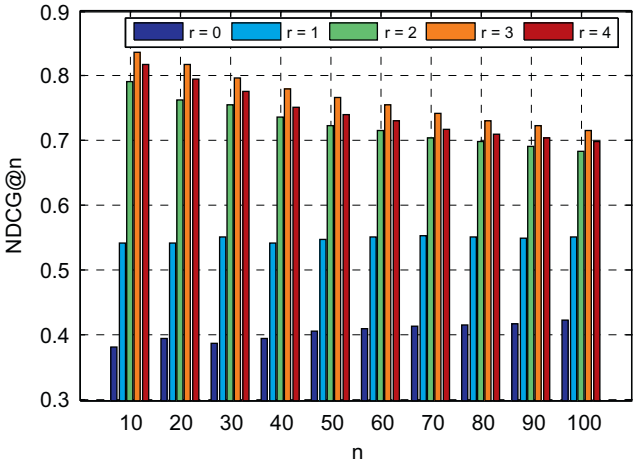

Fig. 5. Performance (NDCG) for different values of r with $\alpha=0.3$. Best performance is obtained with $r=3$.

7. Experiments

To validate the effectiveness of our latent visual context learning (LVCL) scheme, we conduct experiments with images collected from the Web. Thirty famous landmark related text queries are gathered and 500 full size images for each query are downloaded through Google Image. Typical queries include "Coliseum", "Lincoln Memorial", "Eiffel Tower", "Statue of Liberty", and so on. Each image is labeled with ground truth according to its relevance to the corresponding text query on four levels, i.e., "Excellent", "Good", "Fair", "Irrelevant".

Landmark queries are selected for several considerations. First of all, they are popular in Web queries. Second, they contain canonical objects of diversity and fit themselves well to the type of local features used in our study.

For each image, we extract the widely used SIFT features, with a standard implementation [2]. The DOG detectors are used for key point detection and a 128-dimensional orientation histogram is extracted to describe the local patch around the key points. Before feature extraction, images are scaled to have a maximum axis size of 400. From our study, the average SIFT feature number for a single image is 680. For SIFT quantization, a hierarchical visual vocabulary tree [13] with 4 levels and 10 branches for each non-leaf node is adopted. With 1 million samples out of 10 million SIFT features from an independent image dataset for clustering a visual codebook of 10 thousand visual words is obtained.

7.1. Image re-ranking evaluation

To evaluate the performance of image re-ranking results, we adopt normalized discounted cumulative gain (NDCG) [23,24] which is widely used in information retrieval evaluation involving more than two relevance levels. Given a ranking list, the NDCG score at position n is defined as

$$NDCG@n = Z_n \sum_{i=1}^{n} \frac{2^{r(i)}-1}{\log(1+i)} \tag{18}$$

where r(I) is the relevance score of ith image in the ranking list, Z_n is the normalization factor which is chosen such that NDCG@n for the perfect ranking list is 1.

To demonstrate the full potential of LVCL, we set $\lambda=1$ in Eq. (14). For our re-ranking algorithm, there two parameter α in Eq. (1) and r in Eq. (10) to be determined. The re-ranking performance for different values of α is illustrated in Fig. 4, with r set a constant 3. The effect of r will be studied latter. From Fig. 4, it can be observed that latent semantic similarity and visual word hyperlink similarity complement each other and a trade-off is achieved when $\alpha=0.3$.

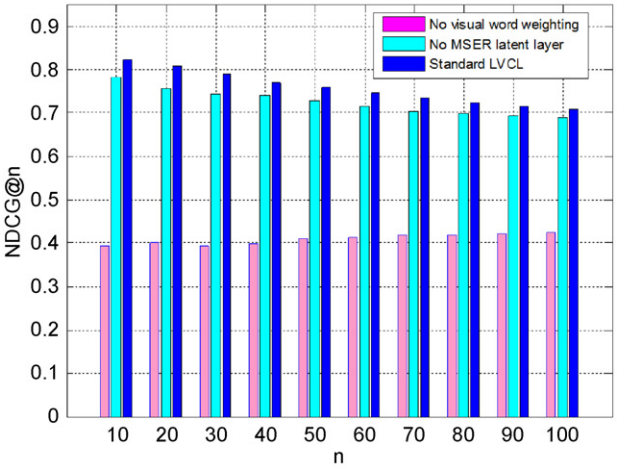

Fig. 6. Performance comparison with and without visual word weighting and MSER layer. The standard LVCL includes both visual word weighting and MSER latent layer.

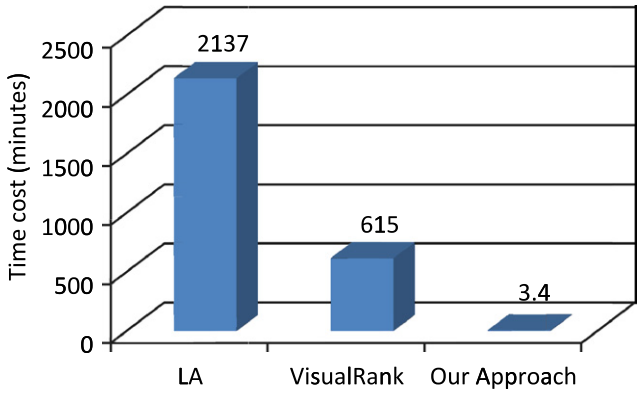

Fig. 8. Average time cost comparison between VisualRank and our LVCL approach for re-ranking 500 full size images per query. Feature extraction time is not included.

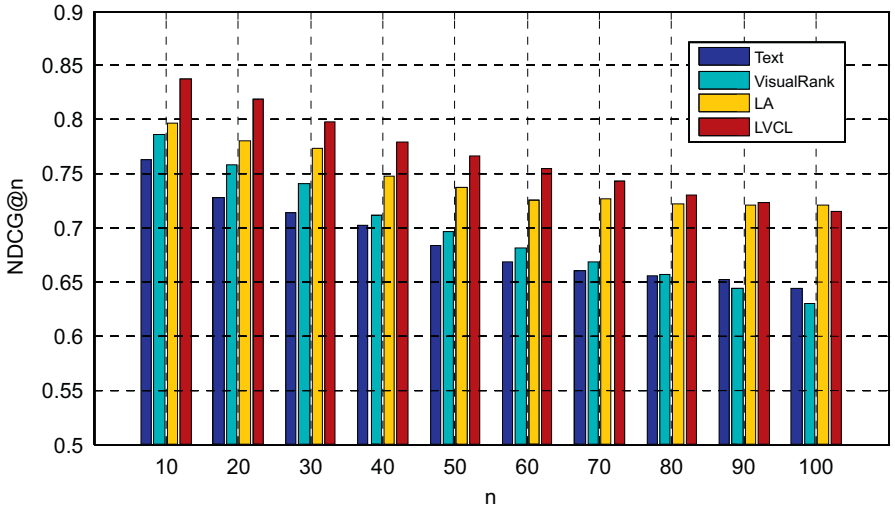

Fig. 7. Performance comparison of re-ranking results between VisualRank, link analysis (LA) and LVCL. The baseline is text-based search results.

To study the effect of r, we test the whole dataset with different values of it. The corresponding performance is illustrated in Fig. 5. It can be observed that, when r increases, the corresponding NDCG first rises and then decreases, with the best performance obtained at $r=3$.

One of the key issues of our approach is to weight the image graph with visual word significance value. To demonstrate the necessity of the visual word weight for image graph, we set $f(x)=1$ in Eq. (9) and keep the other components unchanged. Then we compare the results with and without visual word significance, as shown in Fig. 6. It can be observed that, without weighting the visual word layer, the performance will suffer from dramatic decrease.

To demonstrate the necessity of MSER layer for our visual word graph, we replace the original four-layer visual word graph with a three-layer graph, ignoring the MSER layer, and formulate the visual word transition probability as follows:

$$P(V_j|V_i)=\sum_{k=1}^{M}P(V_j|I_k)P(I_k|V_i) \tag{19}$$

where $P(V_j|I_k)$ denotes the term frequency of V_j in image I_k and $P(I_k|V_i)$ is the same as the previous definition. And other components of the framework stay the same. Then, the altered version is evaluated with the same landmark dataset and the results are compared with that of the standard LVCL approach, as

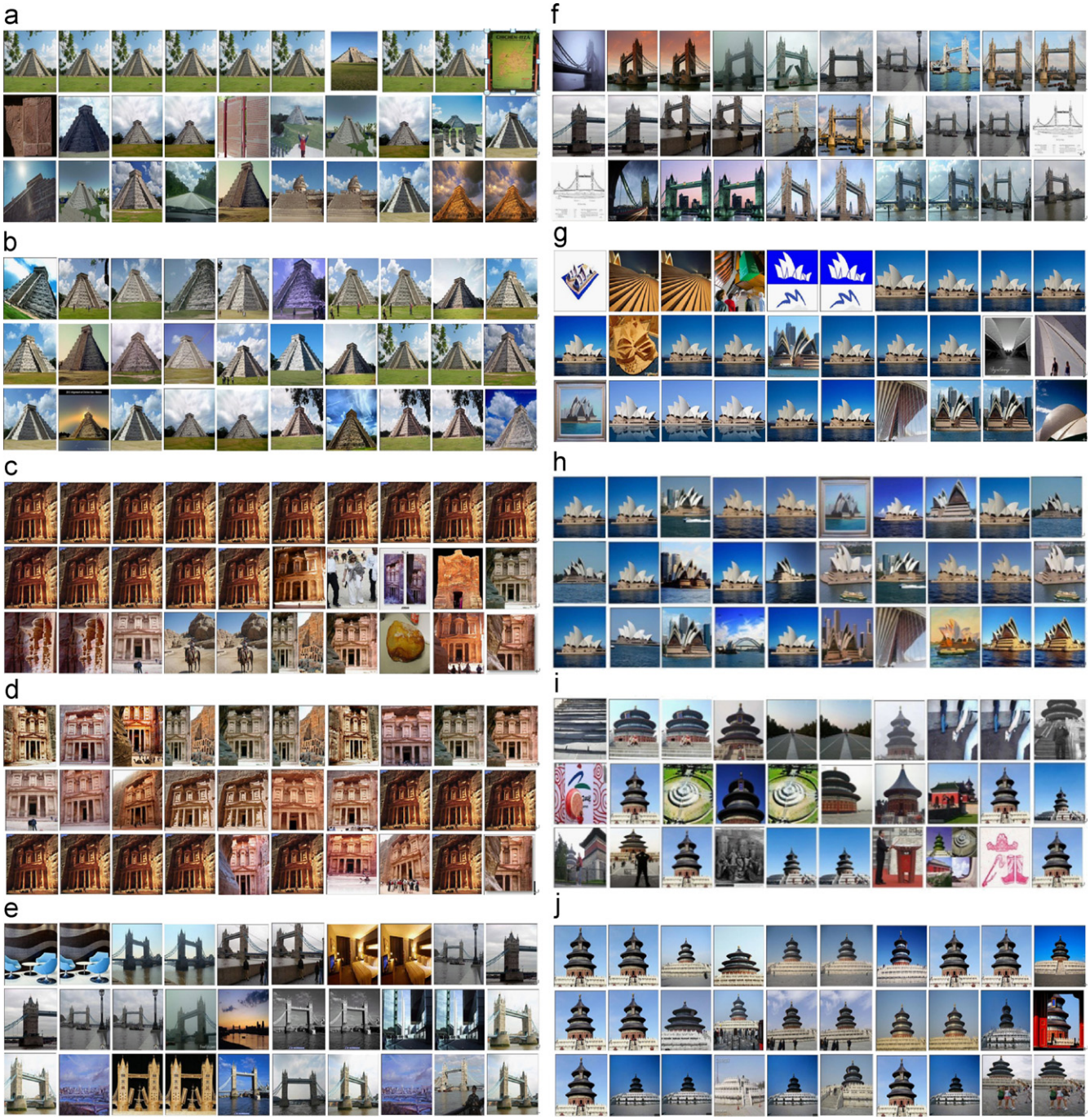

Fig. 9. Top 30 re-ranking results of VisualRank (a),(c), (e), (g), (i) and LVCL (b), (d), (f), (h), (j) for image set obtained from Google Image with text query "Chichen Itza", "Petra in Jordan", "Tower Bridge", "Sydney Opera House" and "Temple of Heaven", respectively.

Fig. 10. Comparison of top five representative images for three queries, "Colosseum", "The Sphinx", and "Mount Rushmore National Memorial", obtained by WSC. Each row corresponds to one query. Left column denotes the results from image re-ranking based on LVCL, and right column is the canonical image selection results.

illustrated in Fig. 6, from which we can observe that with the latent MSER layer for LVCL, improved performance can be achieved.

Fig. 7 illustrates the comparison of VisualRank, link analysis (LA) [31] and LVCL, with the text search results as the baseline. It can be observed that LVCL yields better re-ranking results than VisualRank. Compared with link analysis (LA) [31] approach, LVCL achieves better performance for highly ranked images. When n increases to about 90 and higher, LA will surpass LVCL and works the best. In addition, 5 examples of re-rankings results are also given. In Fig. 9, the top 30 returned results of both VisualRank and our approach are shown for each query, including "Chichen Itza", "Petra in Jordan", "Tower Bridge", "Sydney Opera House" and "Temple of Heaven". It can be observed from these instances that, besides the improved relevance, the image consistency to query is also enhanced for our approach.

Our experiments are performed on a server with 2.0 GHz CPU and 16 G memory. Without considering the time for feature extraction, it takes an average of 3.4 min to re-rank 500 full size images per query with our algorithm, while for VisualRank the average time cost is 614.9 min per query, about two orders of magnitude more than our approach, as illustrated in Fig. 8. Link analysis is the most time consuming and costs 2137 min per query. In fact, the expensive computing of VisualRank is caused by the image pair-wise similarity computation, while for link analysis [31] approach, without feature quantization, the pair-wise correspondence discovery of images by spectral technique [30] is extremely time-consuming. As for our approach, the most time-consuming part is the SIFT quantization and SVD decomposition of a sparse matrix for LSA.

7.2. Canonical image selection

In our experiments, we select the top-ranked 150 images from our image re-ranking results to construct the candidate pool for canonical image selection. Two baseline methods are used for comparison. The first one is the top-K retrieved images from the text-based search results. The second baseline employs affinity propagation (AP) clustering on the same top 150 images. Clusters are ranked by the inter-cluster distance, and the cluster exemplars are selected as canonical images.

Fig. 10 shows three sample results of image re-ranking and canonical image selection based on LVCL, respectively. It can be seen that, although good relevance is guaranteed, the redundancy of duplicated images is also prevalent. With canonical image selection, redundancy is suppressed to achieve good diversity.

To quantitatively evaluate the performance, motivated by [26], we implement a subjective evaluation of the top five results for each method and ask 5 users to answer the following three evaluation questions:

Table 1
Comparison of variant representative image selection methods by subjective evaluation.

Method	Representative	Redundant	1st place voted (%)
Top-K	4.60	2.14	21.3
AP	4.47	2.33	16.0
WSC	4.72	1.14	62.7

(1) Representative: How many photos are representative to the query (0–5)?
(2) Redundant: How many photos are redundant (0–5)?
(3) 1st place voted: Which canonical image set (among the three methods) is the most satisfactory?

The results for each question are averaged over all users and 30 queries.

Table 1 shows a summary of the subjective evaluation results. As for representative score, WSC shares very similar performance to the other two baselines. However, WSC performs much better in terms of redundancy and 1st place voted scores. The consideration of orthogonality and coverage helps WSC obtain the smallest redundant score. It shows that our method can select canonical images, which are not only representatives of the collected photos, but also exhibit a diverse set of views with minimal redundancy.

8. Conclusion and discussion

We propose a novel framework of latent visual context learning (LVCL) for web image applications of image re-ranking and canonical image selection. We explore the visual context structure by analyzing visual link graphs and latent semantics. The significance information for visual words and images are discovered by means of visual graph analysis, respectively. Experiment with landmark Web image dataset demonstrates the superiority of the proposed approach over traditional approaches in image re-ranking and canonical image selection.

In this paper, we select the local feature and MSER region for image representation. Therefore, it is well suited for re-ranking image set of rigid canonical objects, such as landmarks. It may also work well for most product images. However, for cases such as natural scene images, where the image content cannot be satisfactorily characterized by local features, our approach may fail to work.

In this work, singular value decomposition (SVD) is adopted when performing LSA. Due to the computational complexity, it is not applicable to re-rank a large amount of images, say, over one million. However, in practice, it is reasonable to assume the top

text-search results are in good quality. Therefore, it will be sufficient to re-rank only top, such as 500, images, with acceptable computational time cost.

In the future, we will incorporate global feature into our framework to deal with cases where local features are insufficient to work. Moreover, discovering more representative regional visual word set with MSER or some other regional constraints is the next research direction. Further, more comprehensive experiments will be performed on some other public datasets.

Acknowledgement

This work was supported in part by NSFC under contract No. 60632040 and 60672161, Program for New Century Excellent Talents in University (NCET), Research Enhancement Program (REP) and start-up funding from the Texas State University.

References

[1] 〈http://www.tineye.com〉.
[2] D. Lowe, Distinctive image features from scale-invariant key points, IJCV 60 (2) (2004) 91–110.
[3] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: Proceedings of the BMVC, 2002.
[4] S. Deerwester, S.T. Dumais, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.
[5] T. Hofmann., Probabilistic latent semantic indexing, in: ACM SIGIR, 1999.
[6] D.M. Blei, A.Y. Ng, M.I. Jordan, J. Lafferty, Latent dirichlet allocation, Journal of Machine Learning Research (2003).
[7] B.J. Frey, D. Dueck., Clustering by passing messages between data points, Science 315 (2007) 972–976.
[8] Y. Jing, S. Baluja., VisualRank: applying PageRank to large-scale image search, IEEE Transactions on PAMI 30 (2008) 1877–1890.
[9] S. Brin, L. Page, The anatomy of a large-scale hypertextual (Web) search engine, in: Proceedings of the Seventh International World Wide Web Conference, 1998.
[10] D. Cai, X. He, J. Wen, W. Ma, Block-level link analysis, in: Proceedings of the ACM SIGIR, 2004.
[11] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large-scale partial-duplicate web image search, in: Proceedings of the CVPR, 2009.
[12] J. Sivic, A. Zisserman, Video Google: a text retrieval approach to object matching in videos, in: Proceedings of the ICCV, 2003.
[13] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proceedings of the CVPR, 2006, pp. 2161—2168.
[14] Xiao Zhang, Zhiwei Li, Lei Zhang, Wei-Ying Ma, Heung-Yeung Shum, Efficient indexing for large scale visual search, in: Proceedings of the ICCV, 2009.
[15] W.H. Hsu, L.S. Kennedy, S.F. Chang, Video search re-ranking via information bottleneck principle, in: Proceedings of the ACM Multimedia, 2006, pp. 35-44.
[16] J. Liu, W. Lai, X. Hua, Y. Huang, S. Li, Video search re-ranking via multi-graph propagation, in: Proceedings of the ACM Multimedia, 2007, pp. 208–217.
[17] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, X. Hua, Bayesian video search reranking, in: Proceedings of the ACM Multimedia, 2008.
[18] H. Zitouni, S. Sevil, D. Ozkan, P. Duygulu, Re-ranking of web image search results using a graph algorithm, in: Proceedings of the ICPR, 2008, pp. 1–4.
[19] R. Yan, A.G. Hauptmann, R. Jin, Multimedia search with pseudo-relevance feedback, in: Proceedings of the CIVR, 2003.
[20] W.H. Hsu, L.S. Kennedy, S.F. Chang, Video search re-ranking through random walk over document-level context graph, in: Proceedings of the ACM Multimedia, 2007, pp. 971–980.
[21] N. Ben-Haim, B. Babenko, S. Belongie, Improving web-based image search via content based clustering, in: Proceedings of the SLAM, 2006, pp. 106–111.
[22] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrase for image applications, in: Proceedings of the ACM Multimedia, 2009.
[23] K. Jarvelin, J. Kekalainen, IR evaluation methods for retrieving highly relevant documents, in: Proceedings of the ACM SIGIR, 2000.
[24] L. Wang, L. Yang, X. Tian, Query aware visual similarity propagation for image search reranking, in: Proceedings of the ACM Multimedia, 2009.
[25] A. Jaffe, M. Naaman, Generating summaries and visualization for large collections of geo-referenced photographs, in: Proceedings of the ACM MIR, 2006, pp. 89–98.
[26] R. Raguram, S. Lazebnik. Computing Iconic summaries for general visual concepts, in: Proceedings of the CVPR, 2008.
[27] Y.-H. Yang, P.-T. Wu, et al. ContextSeer: context search and recommendation at query time for shared consumer photos, in: Proceedings of the ACM Multimedia, 2008.
[28] I. Simon, N. Snavely, et al., Scene summarization for online image collections, in: Proceedings of the ICCV, 2007, pp. 1–8.
[29] L.S. Kennedy, M. Naaman, Generating diverse and representative image search results for landmarks, in: Proceedings of the WWW, 2008, pp. 297–306.
[30] Marius Leordeanu, Martial Hebert, A spectral technique for correspondence problems using pairwise constraints, in: Proceedings of the ICCV, 2005.
[31] Gunhee Kim, Christos Faloutsos, Martial Hebert, Unsupervised modeling of object categories using link analysis techniques, in: Proceedings of the CVPR, 2008.
[32] Gunhee Kim, Christos Faloutsos, Martial Hebert, Unsupervised modeling and recognition of object categories with combination of visual contents and geometric similarity links, in: Proceedings of the ACM MIR, 2008.

Wengang Zhou received his B.S. degree in Electronic Information Engineering from Wuhan University, Wuhan, China, in 2006. He is currently a Ph.D. student in signal and information processing in EEIS Department, University of Science and Technology of China. His research interests include Multimedia and Computer Vision. He has done some work in Partial Differential Equations (PDE) for bio-medical image processing. His current research is focused on large-scale multimedia information retrieval.

Qi Tian (SM'04) received his Ph.D. degree in electrical and computer engineering from the University of Illinois, Urbana-Champaign in 2002. He is currently an Associate Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA).

Dr. Tian's research interests include multimedia information retrieval and computer vision. He has published over 100 refereed journal and conference papers. His research projects were funded by ARO, DHS, HP Lab, SALSI, CIAS, and CAS. He has been serving as Program Chairs, Session Chairs, Organization Committee Members and TPC for over 120 IEEE and ACM Conferences including ACM Multimedia, SIGIR, ICCV, ICASSP, etc. He is the Guest co-Editors of IEEE Transactions on Multimedia, Journal of Computer Vision and Image Understanding, ACM Transactions on Intelligent Systems and Technology, and EURASIP Journal on Advances in Signal Processing and is the Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and in the Editorial Board of Journal of Multimedia. He is a Senior Member of IEEE and Member of ACM.

Yijuan Lu is an Assistant Professor in the Department of Computer Science, Texas State University. She received her Ph.D. in CS in 2008 from the University of Texas at San Antonio. During 2006, 2007, 2008, she was a summer Intern Researcher at FXPAL lab, Web Search & Mining Group, Microsoft Research Asia (MSRA), National Resource for Biomedical Supercomputing (NRBSC) at the Pittsburgh Supercomputing Center (PSC), Pittsburgh. She was the Intern Researcher at Media Technologies Lab, Hewlett-Packard Laboratories (HP) 2008, and research fellow of Multimodal Information Access and Synthesis (MIAS) Center at University of Illinois at Urbana-Champaign (UIUC) 2007.

Her current research interests include Multimedia Information Retrieval, Computer Vision, Machine Learning, Data Mining, and Bioinformatics. She has published extensively and serves as reviewers at top conferences and journals. She is the 2007 Best Paper Candidate in Retrieval Track of Pacific-rim Conference on Multimedia (PCM) and the recipient of 2007 Prestigious HEB Dissertation Fellowship, 2007 Star of Tomorrow Internship Program of MSRA, She is a member of IEEE and ACM.

Linjun Yang received his B.S. and M.S. degrees from East China Normal University and Fudan University, Shanghai, China, in 2001 and 2006, respectively. Since 2001, he has been with Microsoft Research Asia, Beijing, China, where he is currently an Associate Researcher in the Media Computing Group. His current interests are in the broad areas of multimedia information retrieval with focus on multimedia search ranking and large-scale Web multimedia mining. He has authored or coauthored more than 30 publications in these areas and has more than 10 filed patents or pending applications. He is a member of ACM and IEEE.

Houqiang Li received his B.S., M.Eng., and Ph.D degree from University of Science and Technology of China (USTC) in 1992, 1997, and 2000, respectively, all in electronic engineering. He is currently a professor at the Department of Electronic Engineering and Information Science (EEIS), USTC. His research interests include video coding and communications, image/video analysis, and computer vision. His research has been supported by National Natural Science Foundation of China (NSFC), State High-Tech Development Plan of China (863 program), Microsoft, Nokia, and Huawei. He has authored or co-authored over 60 papers in journals and conferences. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and in the Editorial Board of Journal of Multimedia. He has served on technical/program committees, organizing committees, and as program co-chair, track/session chair for over 10 international conferences.