Pattern Recognition Letters 33 (2012) 239–246

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Feature grouping and local soft match for mobile visual search

Xianglong Liu *, Bo Lang, Yi Xu, Bo Cheng

State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 24 April 2011
Available online 25 October 2011
Communicated by S. Wang

Keywords:
Mobile visual search
Mobile image
Feature grouping
Spatial information
Local soft match

ABSTRACT

More powerful mobile devices stimulate mobile visual search to become a popular and unique image retrieval application. A number of challenges come up with such application, resulting from appearance variations in mobile images. Performance of state-of-the-art image retrieval systems is improved using bag-of-words approaches. However, for visual search by mobile images with large variations, there are at least two critical issues unsolved: (1) the loss of features discriminative power due to quantization; and (2) the underuse of spatial relationships among visual words. To address both issues, this paper presents a novel visual search method based on feature grouping and local soft match, which considers properties of mobile images and couples visual and spatial information consistently. First features of the query image are grouped using both matched visual features and their spatial relationships; and then grouped features are softly matched to alleviate quantization loss. An efficient score scheme is devised to utilize inverted file index and compared with vocabulary-guided pyramid kernels. Finally experiments on Stanford mobile visual search database and a collected database with more than one million images show that the proposed method achieves promising improvement over the approach with a vocabulary tree, especially when large variations exist in query images.

1. Introduction

Recently as mobile phones have evolved into powerful electronic products in daily life, there emerges a class of image search applications whose query images (named mobile images) captured from mobile devices like camera mobile phones. In daily life when people encounter objects (books, arts, etc.) that they are interested in, they would like to get information about these objects. Because mobile phones have become an important part of life, it would be an easy and useful way to take photos of these objects using mobile phones and then search the related information only by submitting these photos to the visual search engine. There are many applications for such an image retrieval system, for example Google goggles (Google, 2009), Nokia Point and Find (Nokia, 2006), Richo iCandy (Erol et al., 2008) and other applications like CD search (Nister and Stewenius, 2006) and street search.

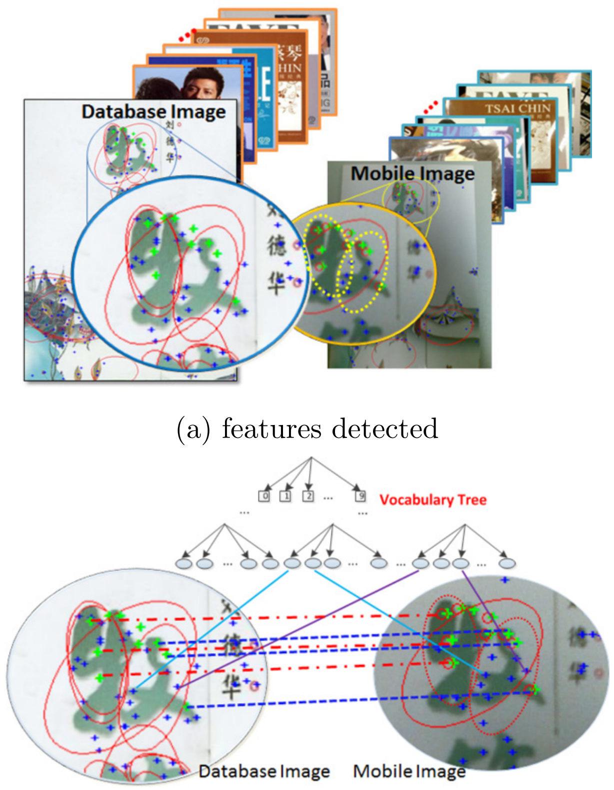
In the literature image search has been very extensively investigated. State-of-the-art large scale image retrieval systems (Sivic and Zisserman, 2003; Nister and Stewenius, 2006; Wu et al., 2009, 2010) achieve efficiency by quantizing local features like Scale-Invariant Feature Transform (SIFT) (Lowe, 2003) into visual words, and then applying scalable textual indexing and retrieval schemes (Salton and Buckley, 1988). However, in this paper we concentrate on visual search by mobile images (taken by the mobile phone) in a large scale image database where usually the target image is the original image captured from. It differs from traditional image retrieval, due to image appearance variations caused by background clutter, foreground occlusion, and differences in viewpoint, orientation, scale and light conditions. Fig. 1(a) illustrates examples of images taken by mobile phones.

For visual search by mobile images with large variations, there exist at least two issues unsolved from visual and spatial aspects: the discriminative power loss and the spatial relationship underuse. Researchers have developed techniques like soft assignment (Philbin et al., 2008), multi-tree scheme (Wu et al., 2010), feature selection (Turcot and Lowe, 2009) and query expansion (Chum et al., 2007). Then for the second issue, methods in (Philbin et al., 2007; Lazebnik et al., 2006) significantly improved retrieval precision using full geometric verification, in practice which is computationally expensive and thus can only be applied to part of the top retrieved images. Moreover, most of previous research take these two issues into consideration independently, and till now there has been very few works to couple both spatial and visual information together to alleviate these issues. Zhang et al. (2011) encode spatial information through the geometry-preserving visual phrases (GVP). Wu et al. (2009) have attempted by exploiting the geometric constraints using bundled features grouped by Maximally Stable Extremal Regions (MSERs) (Matas et al., 2002). However, due to reasons like rotations, both methods would be unsuitable for mobile visual search (Liu et al., 2011).

Motivated by our observations on mobile images, in this paper we propose a novel method based on feature grouping and local soft match. It first groups local visual features (SIFT) using the

* Corresponding author. Tel.: +86 1082338094; fax: +86 1082316736.
E-mail address: xlliu@nlsde.buaa.edu.cn (X. Liu).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.10.002



(a) features detected

(b) grouping features using matched points

Fig. 1. (a) Red eclipses are MSERs extracted by MSER detector while yellow ones in the mobile image are corresponding regions undetected. (b) Blue "·" and green "+", respectively mark SIFT points and exactly matched SIFT points. Points that fit affine transformation well are labeled by red circles near "+" and connected by dot lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exactly matched visual words and their geometric relationships to improve features discriminative ability. Then a soft match scheme is provided for the grouped features to alleviate the loss of feature quantization (Liu et al., 2011). By feature grouping and soft match, visual and spatial information are coupled consistently. For efficient retrieval, we design an efficient index and score mechanism using Term Frequency Inverse Document Frequency (TF-IDF). The decomposition of the score indicates a better discriminative power and can be related to vocabulary-guided (VG) pyramid kernel.

Experiments on Stanford mobile visual search (SMVS) database (Chandrasekhar et al., 2011) and a collected database of more than one million images, show that our method outperforms the method using vocabulary tree (Nister and Stewenius, 2006), especially when large variations exist, while the time is increased slightly.

Specifically the main contributions of this work are as follows:

- We present a simple but effective feature grouping mechanism to improve feature discriminative power by consistently coupling visual and spatial information. Experiments prove that this mechanism helps to improve performance of mobile visual search, especially when large variations exist in mobile query images.
- Based on feature grouping, a novel local soft match scheme is first proposed to utilize neighbor corresponding features in word space and thus to alleviate quantization loss. The soft match hierarchically weights word matches at different levels of the vocabulary, with the similar idea of VG pyramid kernel but incorporating spatial information to guide the match.
- An efficient soft match score based on TF-IDF is devised, and an inverted file index is designed to support fast score calculation. Experiments demonstrates the scalability of this method.
- We first build a large dataset for mobile visual search and hope that it can help advance research in this field.

The rest of this paper is organized as follows. Section 2 describes feature grouping based on visual and spatial consistency. Local soft match including efficient score and index for mobile visual search is given in Section 3. In Section 4 the comparison between soft match and VG pyramid kernel is studied. Section 5 discusses experimental results. Finally we conclude in Section 6.

2. Feature group based on visual and spatial consistency

Grouped features have been proven to be more discriminative than individual local features (Wu et al., 2009; Zhang et al., 2011). If we can detect the corresponding regions, then features can be grouped and matched to improve the discriminative ability. In this section, we first introduce features we will use: SIFT and MSER. Then we will propose a simple but effective scheme to group SIFT features in query image according to MSERs detected in database images and exactly matched SIFT points between them.

2.1. Point features

The SIFT feature is one of the most popular and robust point features (Lowe, 2003), which is invariant to image variations like scale and rotation. Since SIFT features are of high dimension, to match them using similarity will be time consuming. The efficient way is bag-of-words which quantizes SIFT features into some visual words using the vocabulary tree (Sivic and Zisserman, 2003; Nister and Stewenius, 2006; Wu et al., 2009).

In this paper, we extract SIFT features of each image including both database images and query images, and quantize them using the vocabulary tree as Nister and Stewenius (2006) did. The vocabulary tree, built by hierarchical k-means clustering, defines a hierarchical quantization. Features of all the training data are extracted and clustered recursively by an k-means process into k (we set $k = 10$) groups and the tree finally goes up to the maximum level L (we use $L = 6$). For feature quantization, each feature is simply propagated down to the L level of the tree and represented by corresponding leaf nodes (visual words) in the vocabulary tree.

2.2. Features group

To enhance the discriminative ability of SIFT, region features like the Maximally Stable Extremal Region (MSER) (Matas et al., 2002) can be used to group SIFT features (Wu et al., 2009). MSER performs better on detection of affine-covariant stable elliptical regions than other region features like Harris-affine and Hessian-affine (Mikolajczyk et al., 2005). Wu et al. (2009) bundle features of both query and database images by MSERs, however, Liu et al. (2011) show that it would fail in mobile visual search for the reasons like unrepeatable visual features (Fig. 1(a)) and no rotation assumption. Our observation, that usually corresponding regions of query images and original database images have more common SIFT features exactly matched, motivates us to detect the corresponding regions in mobile images using the information of both these matched SIFT points and MSERs well detected in database images, which are free of variations.

Fig. 1(b) highlights the well matched SIFT points and shows that the matched points fit the affine transformation very well. It indicates a straightforward way to detect the corresponding region: since the local region is usually small, we can assume the region in query image is obtained by affine transformation of MSER in the original database image. We randomly select three pairs of matched points $\boldsymbol{x} = (x,y)^T$ and $\boldsymbol{u} = (u,v)^T$ satisfying simple geometric constraints (length ratio standard variation $\leqslant 0.05$), and estimate the affine transformation $\boldsymbol{A}$ from these points by solving linear equations as Lowe (2003) did. After estimation, we can detect the corresponding region in the query image by $\boldsymbol{u}' = \boldsymbol{A}x'$: affine transformation of the corresponding MSER (only three key points $\boldsymbol{x}'$: the center, major and minor axis endpoints) in the original image.

Denote $S = \{p_i\}$ the SIFT features and $R = \{r_k\}$, $k > 0$ both the MSER in the database image I and the detected region in the input mobile image Q. We define feature groups G_k to be: $G_k = \{p_i | p_i \in r_k, p_i \in S\}$, where $p_i \in r_k$ means that the point feature p_i falls inside the region r_k. For each G_k, if its corresponding feature group G'_k in query image Q can be detected through the above process, we denote $G_k \triangleright Q$. In practice the ellipse of the MSER is enlarged by factor 1.5 when computing $p_i \in r_k$ (Wu et al., 2009), and r_k is discarded if it is empty or its ellipse spans more than half the width or height of the image. Furthermore, SIFT features that do not belong to any G_k are treated to fall into the same region r_0 and form G_0. Here, G_0 might dissatisfy $G_0 \triangleright Q$ because it is not a MSER of I.

Since the feature group contains multiple SIFT features and the spatial corresponding information, we believe that they will be more discriminative than a single SIFT feature, which will be verified by our experiments. These feature groups also allow us to exploit relationships between the neighbor visual words in the corresponding regions.

3. Local soft match and score

After feature groups are detected using both the visual and spatial information, relationship between corresponding features in these groups can be exploited and utilized to alleviate quantization loss.

3.1. Soft match

Due to different variations mentioned before, features of the mobile image are usually changed compared with corresponding ones of the original database image. Thus after feature quantization, the visual words of corresponding features may be different as shown in Fig. 1(b). However, our observation indicates that these visual words are usually neighbors in the vocabulary tree. If these neighbor words in detected corresponding regions can be fully utilized, then we can further improve the performance of mobile visual search. In this paper, we propose a local soft match mechanism that softly matches features in the corresponding regions to alleviate the quantization loss, and thus improve the match accuracy. Compared to Philbin et al. (2008), this method works in word space and saves distance computation between high dimensional features and several cluster centers.

Let $G_k = \{p_i\}$, $(k > 0)$ feature group in database image I and $G'_k = \{q_j\}$ its corresponding feature group in query image Q. Point features p_i and q_j are quantized and represented by visual words in our visual vocabulary W. Words in W are sequentially numbered from the leftmost leaf node to the rightmost one. If $\|p_i - q_j\| \leqslant D$, which means they are close in W, then we call that they are neighbors and denote $\mathcal{N}^D_{p_i}(q_j)$. Here $\|\cdot\|$ is the word number difference. Note that when $D = 0$, namely $p_i = q_j$, the neighbor words are exactly matched.

$$\mathcal{N}^D_{p_i}(q_j) = \begin{cases} 1, & \text{if } \|p_i - q_j\| \leqslant D; \\ 0, & \text{if } \|p_i - q_j\| > D. \end{cases} \tag{1}$$

We use exponential weighting function to measure the importance of soft match. Intuitively, the closer the words are, the higher the probability that they are corresponding features is, and thus the more weight the match between them will gain. The weight decays as the distance of two words in vocabulary tree increases:

$$\omega^D_{p_i,q_j} = \mathcal{N}^D_{p_i}(q_j) e^{-\|p_i - q_j\|}. \tag{2}$$

3.2. Soft match score

Then in retrieval, the TF-IDF score (Salton and Buckley, 1988) is used to measure the similarity. For each visual word p_i in the feature group G_k of database image I: if $G_k \triangleright Q$, then visual words of G_k and corresponding G'_k detected in Q can be both exactly and softly matched; Otherwise, only exact match is operated on words of G_k and Q. Now, we define a matching score $M_Q(G_k)$ for feature group G_k. The matched features are scored:

$$M_Q(G_k) = \begin{cases} \lambda_{G_k,G'_k} \sum\limits_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}, & \text{if } G_k \triangleright Q; \\ \sum\limits_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}, & \text{otherwise}, \end{cases} \tag{3}$$

where $\upsilon_{\cdot} = \text{tf}_{\cdot}\text{idf}_{\cdot}$, $\text{tf}_{\cdot}$ is the word frequency and $\text{idf}_{\cdot}$ is the inverse document frequency. We give a higher score for spatial matched regions with more common visual words using the term λ_{G_k,G'_k}:

$$\lambda_{G_k,G'_k} = \ln \sum_{p_i \in G_k, q_j \in G'_k} \mathcal{N}^0_{p_i}(q_j). \tag{4}$$

Finally, a database image I is scored S for the query image Q:

$$S_Q(I) = \sum_{G_k} M_Q(G_k). \tag{5}$$

The score actually combines both spatial and visual match, and achieves a consistency between them by the region detection and soft match. Regions with many common and neighbor words will be scored higher than regions with fewer matched words.

3.3. Efficient soft match score

In retrieval to utilize the inverted file index in Section 3.4 and compute the soft match score $S_Q(I)$ efficiently, we first denote

$$S^{\mathcal{W}}_Q = \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \notin G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} \tag{6}$$

and

$$S^{\mathcal{N}}_Q = \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} (\lambda_{G_k,G'_k} - \mathcal{N}^0_{p_i}(q_j)) \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}. \tag{7}$$

Then $S_Q(I)$ can be rewritten in a more comprehensible form (see Appendix A):

$$S_Q(I) = \sum_{\substack{p_i \in I \\ q_j \in Q}} N_{p_i} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - S^{\mathcal{W}}_Q + S^{\mathcal{N}}_Q, \tag{8}$$

where $N_{p_i} = \sum_{G_k} I_{(p_i \in G_k)}$ denotes the number of feature groups that p_i belongs to.

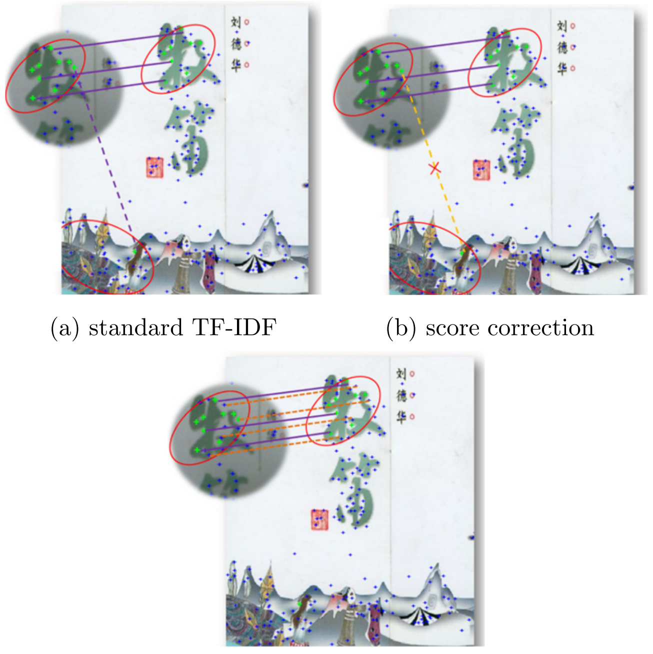
Note that the three terms in the matching score play different roles in similarity measurement: standard TF-IDF score weighted by N_{p_i} (lines in Fig. 2(a)), score correction of wrong matches (dotted lines in Fig. 2(b)), and the score for soft match of local grouped features (lines in Fig. 2(c)). Multiple occurrence in several regions means a strong geometric correspondence, and therefore the first term gives higher score for words appearing in multiple regions. In the second term, part of wrong matches in G_k satisfying $G_k \triangleright Q$ are removed by feature grouping. For the third term, because $G_k \triangleright Q$ holds, there exist at least three exactly matched points, namely $\sum_{p_i \in G_k, q_j \in G'_k} \mathcal{N}^0_{p_i}(q_j) \geqslant 3$, and thus $\lambda_{G_k,G'_k} > 1$ (in practice we multiplies λ with a scale factor larger than 1.0). With $\mathcal{N}^0_{p_i}(q_j) \leqslant 1, S^{\mathcal{N}}_Q \geqslant 0$ holds. Therefore for database images I that have more corresponding regions to the query image Q, more wrong matches would be eliminated and additive positive score $S^{\mathcal{N}}_Q$ will be added (dotted lines in Fig. 2(c)), which means that the last two terms will be very helpful for search by mobile images and especially partial ones.

3.4. Index

We use an inverted file index (Salton and Buckley, 1988) for large-scale indexing and retrieval, since it has been proved efficient for both textual and image retrieval (Sivic and Zisserman, 2003; Nister and Stewenius, 2006; Wu et al., 2009). Fig. 3 shows the structure of our index. Each visual word has a list in the index containing images (ImgID, 32 bit) and MSERs (RID, 8 bit) in which the visual word appears. In addition, for each occurrence of a visual word, the word frequency (tf, 8 bit) is also stored as traditional inverted file index does. This format supports at most 2^{32} images and 256 MSERs per image. Actually all images contains less than 256 regions in this paper.

The traditional text index would contain the location (x,y) of each word within the document, while in our index we stored positions of visual words for each region in database images. Also another table records central point coordinates (x,y), lengths a, b of major and minor axis, angles θ, and visual words pointers for MSERs of each database image with structure shown in Fig. 3(b).

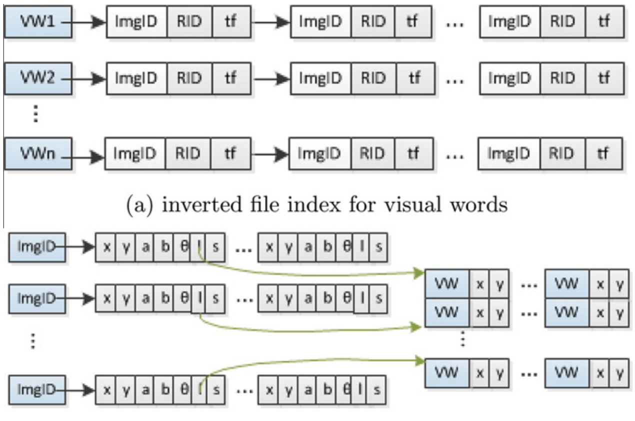
After index has been built, the mobile image retrieval can be solved by traversing the inverted file index once. The first part of the score (Eq. (8)) can be calculated efficiently by traversing the inverted file index as standard textual retrieval does. During this process, the corresponding regions will be detected by affine transformation estimated from common words in same MSERs. According to the geometric constraints between these corresponding regions, wrong matches are removed and soft match is conducted, namely the last two parts $S^{\mathcal{W}}_Q$ and $S^{\mathcal{N}}_Q$ can also be obtained after index traversing. Finally database images are ranked by their scores. We can see that the index can be utilized efficiently by the score scheme in previous section.



(a) standard TF-IDF (b) score correction

(c) soft match score

Fig. 2. Score based on local soft match.



(a) inverted file index for visual words

(b) region index

Fig. 3. Index structure.

4. Comparison with pyramid kernel

Vocabulary-guided (VG) pyramid match (Graumana and Darrell, 2006; Pu et al., 2007) is a fast kernel function which approximates the optimal partial match between any two feature sets in linear time. The basic idea of VG pyramid kernel is to partition the given feature space into a pyramid, according to which points set are then encoded as multi-resolution histograms. For feature space of d dimension, after partitioning, there are L levels of bins over the feature space $\{H_0, H_1, \ldots, H_{L-1}\}$. For two feature sets X and Y, first let $H_i(X)$ and $n_{ij}(X)$, respectively denote the histogram of X in level i and number of points in X falling into the jth bin of H_i. Then VG pyramid kernel is in the following form:

$$\mathcal{K}(X,Y) = \sum_{i=0}^{L-1} \sum_{j=1}^{k^i} (w_{ij} - p_{ij}) \min(n_{ij}(X), n_{ij}(Y)), \tag{9}$$

where k is the number of sub-clusters for each level. From this equation, we can note that VG pyramid kernel is actually a hierarchical weighted score for the pyramid.

Our soft match based on TF-IDF has the similar idea to that of VG pyramid match. The vocabulary tree for soft match can be viewed as a pyramid, and instead of min operation in VG pyramid kernel we use TF-IDF. TF-IDF is also positive definite (Bishop, 2006) and has been proved to be an useful kernel in both textual and image retrieval.

The first two parts of soft match score $S_Q(I)$ in Eq. (8), as Eq. (A.4) in Appendix A shows, are the bottom level (leaf nodes) match because only the exact match is operated on words:

$$\sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}, \tag{10}$$

where $\omega^0_{p_i,q_j} \geqslant 0$ will always hold.

The last part, namely $S^{\mathcal{N}}_Q$, actually account for match at top levels of the pyramid and hierarchically scores different levels by soft match between neighbor words in spatial corresponding regions:

$$\sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \left[I_{(p_i=q_j)} (\lambda_{G_k,G'_k} - 1) \upsilon_{p_i}\upsilon_{q_j} + \sum_{d=1}^{D} I_{(\|p_i-q_j\|=d)} \lambda_{G_k,G'_k} e^d \upsilon_{p_i}\upsilon_{q_j} \right], \quad (11)$$

where $I_{(\cdot)}$ is an indicator function. We have prove that $\lambda_{G_k,G'_k} > 1$, and thus both $I_{(p_i=q_j)}(\lambda_{G_k,G'_k} - 1) \geqslant 0$ and $I_{(\|p_i-q_j\|=d)}\lambda_{G_k,G'_k} e^d \geqslant 0$ hold. For each $d > 0$, $I_{(\|p_i-q_j\|=d)}\lambda_{G_k,G'_k} e^d \upsilon_{p_i}\upsilon_{q_j}$ is the similarity contribution of d-neighbor words to top levels over level $(L - \lceil \log_k d \rceil - 1)$.

Therefore, both soft match and VG pyramid kernel weight feature matches at different levels and then take the sum of all hierarchically weighted match score as the similarity of two features sets.

The difference is that the spatial information is lost in VG pyramid kernel, while in soft match, the spatial relationship is used to guide both the match and the weight at different levels as discussed in Section 3.3: wrong matches between bottom children are eliminated and large weight is assigned to match between neighbors. Such scheme couples both visual and spatial information to improve the discriminative power of the local features.

5. Experiments

We evaluate the proposed method from different aspects by performing queries on a reference and a collected databases.

5.1. Data sets and measurement

To evaluate the performance on different categories, we conduct experiments on Stanford mobile visual search dataset (SMVS) (Chandrasekhar et al., 2011) which has been released recently.

For the performance with respect to database size, we crawled one million images including posters and CD covers of 7000 popular singers in Google Music (http://www.google.cn/music/artistlibrary) to form our basic dataset, and build three smaller datasets (M = 5 K, 30 K, and 100 K) by sampling the basic dataset. We manually take photos of sampled CD covers using camera phones (CHT9000 and Nokia 2700c) with background cluttering, foreground blocking, and different light conditions, viewpoints, scale and rotations. n = 100 representative mobile images are selected and labeled as our queries in our experiments. Each of these mobile images corresponds to only one original image in the dataset. Fig. 1(a) illustrates typical examples. The dataset can be downloaded from http://www.nlsde.buaa.edu.cn/xlliu/mvs_images.zip.

In both datasets, SIFT features and MSERs of each image are extracted using the Open Source SIFT Library (Hess, 2010) and MSER functions in OpenCV (OpenCV, 1999). The maximum number of SIFT points per image is 3000, while for MSER, each region is enlarged by factor 1.5 and will be discarded if too small ($\leqslant$10% of the width or height) or too large ($\geqslant$50% of the width or height). Then a vocabulary tree with one million words is used to quantize SIFT features as mentioned before.

In this paper, for mobile image search we concern whether the original database image is retrieved and ranked on the top, namely the rank of the correct answer, so we use mean reciprocal rank (MRR) as our evaluation metric following Voorhees (1999). For each query image, its reciprocal rank is calculated and then averaged for all queries. The MRR is defined as follows:

$$MRR@\gamma = \frac{1}{n} \sum_i I_{(rank_i \leqslant \gamma)} \frac{1}{rank_i}, \quad (12)$$

where n is the query number and $rank_i$ stands for the position of the original database image in the retrieved list. Because we only concern top results, we truncate the list at different positions including γ = 1, 3, 5 and 10, and check whether the original database image is contained in the list before these positions.

5.2. Evaluation

Liu et al. (2011) show that bundling method (Wu et al., 2009) would fail in mobile visual search (MRR performance are lower than 20% for both datasets), so in this paper we only comprehensively compare the proposed method with the recognition method with vocabulary tree ("voctree") (Nister and Stewenius, 2006), which has been proven effective in image retrieval. Both methods are first running without re-ranking using geometric verification. We use a vocabulary of 1 M visual words following Nister and Stewenius (2006) and the maximum distance D of nearest neighbors in the soft match is set to 10 (Liu et al., 2011).

5.2.1. Varying categories

To evaluate the performance of the proposed method for different applications in mobile visual search, experiments on different categories are conducted on Stanford mobile visual search dataset (SMVS). There are totaly 1191 distinct database images across 8 image categories in SMVS and 3264 query images taken by camera phones like iPhone4 and Nokia N95. These images are collected under widely varying light conditions, foreground and background clutter (Chandrasekhar et al., 2011). The results are shown in Table 1 where the proposed method achieve better performance than "voctree" for most categories, especially indoor ones like books, CD, DVD, and video frames, while both the proposed and "voctree" might be not suitable for arts, landmark and prints search, where the landmark category is the most challenging. This conclusion is consistent with that of Chandrasekhar et al. (2011). One possible reason might be that, in images like landmarks the dynamic environments bring much noise, which results in lots of wrong matches between query and database images, and thus degrades the retrieval performance.

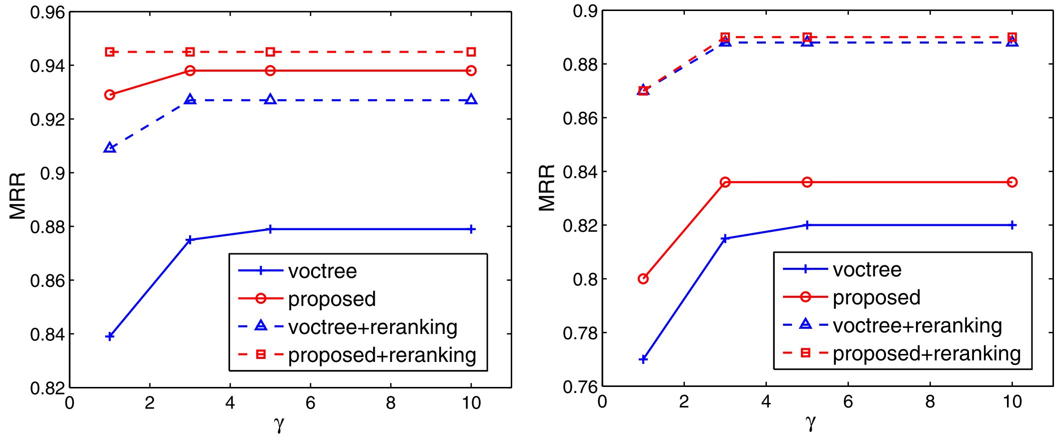
5.2.2. Low quality query images

In SMVS, most the query images in SMVS are in high resolution. However, in practice a number of query images are captured by

Table 1
Experiment results on SMVS dataset with n query images and M database images for each category.

Category (M/n)	$MRR@1$		$MRR@3$		$MRR@5$		$MRR@10$	
	Voctree	Proposed	Voctree	Proposed	Voctree	Proposed	Voctree	Proposed
Books (100/400)	0.862	**0.890**	0.895	**0.914**	0.898	**0.916**	0.899	**0.917**
Cards (100/400)	0.537	**0.540**	0.646	**0.647**	0.654	**0.655**	0.663	**0.665**
CD (100/400)	0.663	**0.685**	0.699	**0.723**	0.710	**0.733**	0.721	**0.743**
DVD (100/400)	0.700	**0.738**	0.744	**0.775**	0.753	**0.784**	0.765	**0.793**
Landmarks (500/500)	**0.112**	0.102	**0.158**	0.149	**0.170**	0.165	**0.178**	0.171
Arts (91/364)	0.470	**0.478**	0.547	**0.551**	0.561	**0.565**	0.574	**0.578**
text (100/400)	0.290	**0.295**	0.320	**0.322**	0.332	**0.335**	0.341	**0.343**
Frames (100/400)	0.825	**0.843**	0.860	**0.873**	0.866	**0.878**	0.870	**0.882**

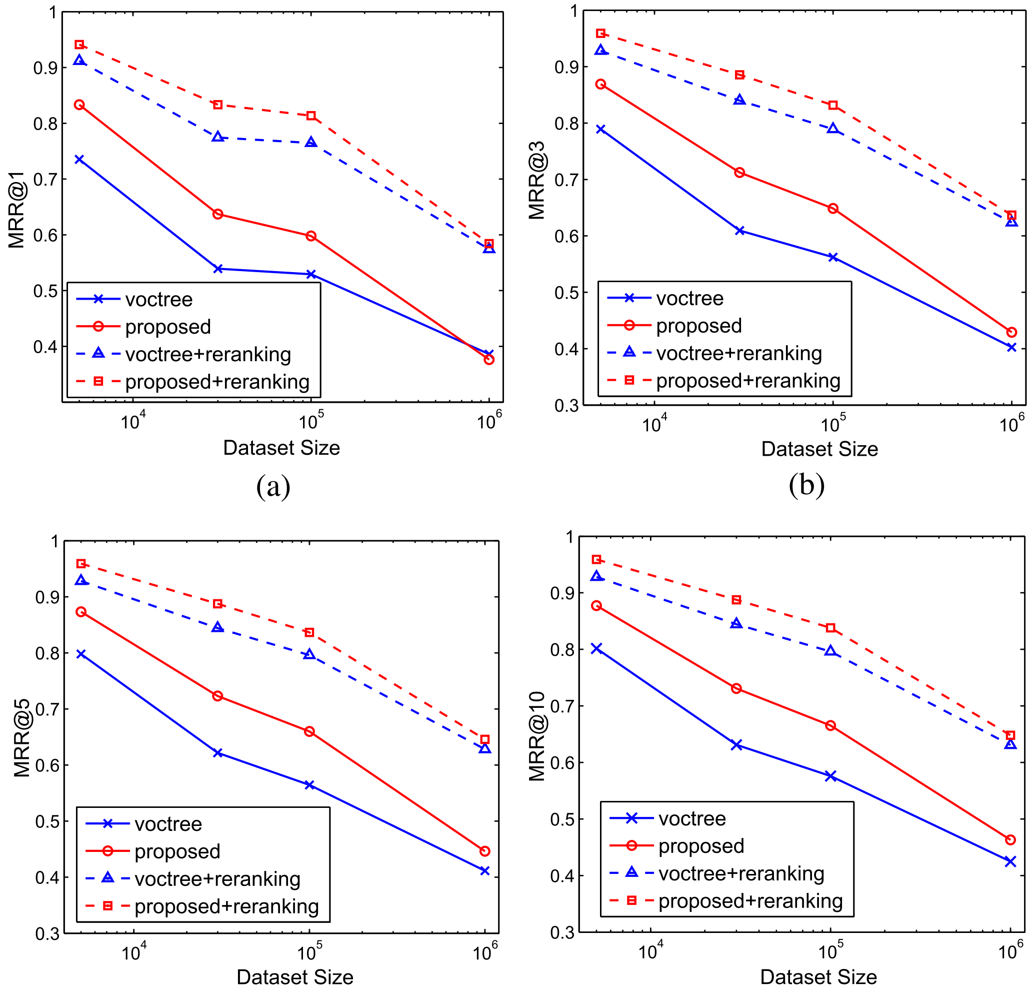
Bold values indicate better performance.



(a) books ($M = 100/n = 56$)

(b) cards ($M = 100/n = 100$)

Fig. 4. Performance comparison on low quality query images of SMVS.



(a)

(b)

(c)

(d)

Fig. 5. Performance comparison on collected dataset of different size.

low-end camera phones. Also in some cases limited by the transmission bandwidth in mobile applications, query images should be of low resolution. So it is important that the visual search algorithms can well support search by low quality mobile images. To evaluate such performance of the proposed method, we choose query images with resolutions below 800 × 800 from SMVS. These images contains much noise like large foreground blocking, background cluttering, and light and viewpoint variations. The details of this subset and experimental results are shown in Fig. 4. Compared with results in Table 1, a larger performance improvement (5.2% on average) is achieved on low quality query images, which means that our method can work better on query images with large variations and noises.

Experiments on collected dataset also verify this conclusion. Our query images in collected dataset are of much lower resolutions (most 640 × 480), contain greater variations and are usually partial (Fig. 1(a)). Fig. 5 shows that on datasets of different size, the proposed method also archives a larger improvement (12.0% on average) to “voctree”. We believe that this because that the grouped features and their soft match in our method encode more discriminative information in low quality images and serve as a complement to the traditional visual TF-IDF by achieving visual and spatial consistency.

5.2.3. Different database size

To evaluate the scalability and efficiency of the proposed method, the collected database and three smaller datasets (5 K, 30 K and 100 K) are built. All of these datasets are much larger than SMVS.

Performance comparison between both methods shown in Fig. 5 leads to two main observations. First, on the 5 K, 30 K, 100 K, and 1 M datasets, performance of both methods degrades as the size of datasets increases; Second, the proposed method significantly outperforms “voctree” with respect to different database size. On 1 M dataset, MRR@10 of the proposed method reaches 46.3% (a 8.94% improvement) as shown by the curve labeled “proposed” in Fig. 5(d).

We perform experiments on a desktop with a single CPU of 2.5 GHz and 16G memory. For one image query on 1 M dataset, the average query time of the proposed method is 1.34 s while 0.87 s for “voctree”. As expected, more time is spent in the proposed method due to feature matching in corresponding regions. But it indicates that the proposed approach takes no much more query time but achieves higher retrieval accuracy than “voctree”.

5.2.4. Re-ranking

We now investigate re-ranking the top-ranked results using spatial constraints. In image retrieval system, re-ranking can substantially improve the retrieval performance by filtering out false positive images based on spatial verification (Wu et al., 2009; Philbin et al., 2007; Lazebnik et al., 2006). The standard estimation algorithm is RANSAC (Fischler and Bolles, 1981): generate transformation hypotheses using a minimal number of correspondences and then evaluate each hypothesis based on the number of inliers among all features under that hypothesis. Because re-ranking is time-consuming, we only re-rank the top 30 candidate images on SMVS and 100 on collocated dataset.

As comparison shown in Figs. 4 and 5, spatial re-ranking significantly improves the retrieval quality of the system. After re-ranking for both method, the proposed method (“proposed + reranking”) still outperforms “voctree” (“voctree + reranking”). For book category in Fig. 4(a), even our method without re-ranking achieves better performance than ”voctree” with re-ranking. When combined re-ranking, our method achieves MRR@5 of 0.945, a 12.6% improvement over “voctree” and a 4.0% improvement over “voctree” with re-ranking (for $\gamma = 1, 3, 10$, we obtain similar results). This is because only images in top retrieved list are re-ranked, whereas our approach can fundamentally improve the matching quality by coupling spatial and visual information, and thus bring more correctly matched images into the list.

On the whole, the proposed method based on feature grouping and soft match provides a favorable performance: (1) a significant improvement on both datasets with or without re-ranking; (2) a promising performance on different categories, especially indoor ones; (3) robustness to low quality query images; and (4) scalability up to one million images.

6. Conclusion

In this paper we have proposed a novel mobile visual search method based on feature grouping and local soft match. With respect to properties of mobile images, features of query image are first grouped using matched ones and their spatial corresponding information; Then feature groups are softly matched. An efficient scheme for soft match score schemes is proposed to utilize the inverted file index and analysis shows that it possesses good properties. The local soft match exploits spatial relationships between features and combines them with visual matching information to improve features discriminative power. Experimental results make us believe that the performance may be improved further by exploiting better schemes that can utilize both visual and spatial information consistently.

Acknowledgments

We thank Wei Su of Beihang University and Hao Su of Stanford University for discussion on spatial information and data collection. We also would like to thank Lusong Li of National University of Singapore for his valuable comments. This work is supported by the National High Technology Research and Development Program (2009AA043303) and National Major Project of China “Advanced Unstructured Data Repository” (2010ZX01042-002-001-00).

Appendix A

Given query image Q, the score $S_Q(I)$ for database image I can be calculated by summing over all feature groups G_k in I:

$$S_Q(I) = \sum_{G_k} M_Q(G_k) \tag{A.1}$$

$$= \sum_{G_k \triangleright Q} \lambda_{G_k,G'_k} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} \tag{A.2}$$

$$= \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$$

$$+ \sum_{G_k \triangleright Q} \left[\lambda_{G_k,G'_k} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} \right] \tag{A.3}$$

$$= \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$$

$$+ \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} (\lambda_{G_k,G'_k} - \mathcal{N}^0_{p_i}(q_j)) \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}. \tag{A.4}$$

Let $S^{\mathcal{N}}_Q$ denote $\sum_{G_k \triangleright Q} \sum_{p_i \in G_k, q_j \in G'_k} (\lambda_{G_k,G'_k} - \mathcal{N}^0_{p_i}(q_j)) \omega^D_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$, then $S_Q(I)$ can be rewritten as:

$$S_Q(I) = \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + S^{\mathcal{N}}_Q \tag{A.5}$$

$$= \sum_{G_k \triangleright Q} \left[\sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - \sum_{\substack{p_i \in G_k \\ q_j \notin G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} \right]$$

$$+ \sum_{G_k \not\triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + S^{\mathcal{N}}_Q \tag{A.6}$$

$$= \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + \sum_{G_k \ntriangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$$

$$- \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \notin G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} + S^{\mathcal{N}}_Q. \tag{A.7}$$

After substituting $S^{\mathcal{W}}_Q$ for $\sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \notin G'_k}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$, the score can be decomposed into three parts as follows:

$$\begin{aligned} S_Q(I) = & \sum_{G_k \triangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} \\ & + \sum_{G_k \ntriangleright Q} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - S^{\mathcal{W}}_Q + S^{\mathcal{N}}_Q \end{aligned} \tag{A.8}$$

$$= \sum_{G_k} \sum_{\substack{p_i \in G_k \\ q_j \in Q}} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - S^{\mathcal{W}}_Q + S^{\mathcal{N}}_Q. \tag{A.9}$$

Since $\{p_i | p_i \in I\} = \bigcup_k \{p_i | p_i \in G_k\}$, then $\sum_{G_k} \sum_{p_i \in G_k, q_j \in Q} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} = \sum_{p_i \in I, q_j \in Q} N_{p_i} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j}$, where $N_{p_i} = \sum_{G_k} I_{(p_i \in G_k)}$ denotes the number of feature groups that p_i belongs to. Therefore, we derive the final result:

$$S_Q(I) = \sum_{\substack{p_i \in I \\ q_j \in Q}} N_{p_i} \omega^0_{p_i,q_j} \upsilon_{p_i} \upsilon_{q_j} - S^{\mathcal{W}}_Q + S^{\mathcal{N}}_Q. \tag{A.10}$$

References

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Chandrasekhar, V., Chen, D., Tsai, S., Cheung, N., Chen, H., Takacs, G., Reznik, Y., Vedantham, R., Grzeszczuk, R., Bach, J., Girod B., 2011. The stanford mobile visual search data set. In: ACM Conf. on Multimedia Systems (MMSys), pp. 117–122.

Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A., 2007. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. In: Internat. Conf. on Computer Vision (CVPR), pp. 1–8.

Erol, B., Antunez, E., Hull, J., 2008. HOTPAPER: multimedia interaction with paper using mobile phones. In: ACM Internat. Conf. on Multimedia (MM), pp. 983–984.

Fischler, M., Bolles, R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Comm ACM, 24:381–395.

Google, 2009. <http://www.google.com/mobile/goggles/>.

Graumana, K., Darrell, T., 2006. Approximate Correspondences in High Dimensions. Adv. Neural Inform. Process. Systems (NIPS), 505–512.

Hess, R., 2010. An open source SIFT Library. In: ACM Internat. Conf. on Multimedia (MM).

Lazebnik, S., Schmid, C., Ponce J., 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2169–2178.

Liu, X., Lou, Y., Yu, W., Lang, B., 2011. Search by mobile image based on visual and spatial consistency. In: IEEE Internat. Conf. on Multimedia & Expo (ICME).

Lowe, D., 2003. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vision (IJCV) 20, 91–110.

Matas, J., Chum, O., Urban, M., Pajdla T., 2002. Robust wide baseline stereo from maximally stable extremal regions. In: The British Machine Vision Conf. (BMVC), pp. 384–393.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L., 2005. A comparison of affine region detectors. Internat. J. Comput. Vision (IJCV) 65, 43–72.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2161-2168.

Nokia, 2006. <http://www.pointandfind.nokia.com>.

OpenCV, 1999. <http://www.opencv.willowgarage.com>.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-8.

Pu, W., Liu, N., Yan, S., Yan, J., Xie, K., Chen, Z., 2007. Local Word Bag Model for Text Categorization. In: IEEE International Conference on Data Mining (ICDM), pp. 625–630.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage.: Internat. J. 24 (5), 513–523.

Sivic, J., Zisserman, A., 2003. Video Google: a text retrieval approach to object matching in video. In: International Conference on Computer VisionInternational Conference on Computer Vision (CVPR), pp. 1470–1477.

Turcot, P., Lowe, D., 2009. Better matching with fewer features: The selection of useful features in large database recognition problems. In: ICCV Workshop on Emergent Issues in Large Amounts of Visual Data (WS-LAVD), pp. 2109–2116.

Voorhees, E., 1999. TREC-8 Question Answering Track Report. In: Proc. of the 8th Text Retrieval Conference, pp. 77–82.

Wu, Z., Ke, Q., Isard, M., Sun J., 2009. Bundling features for large-scale partial-duplicate web image search. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 25–32.

Wu, Z., Ke, Q., Sun J., 2010. A multi-sample, multi-tree approach to bag-of-words image representation for image retrieval. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1992–1999.

Zhang, Y., Jia, Z., Chen, T., 2011. Image retrieval with geometry-preserving visual phrases. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).