LARGE SCALE PARTIAL-DUPLICATE IMAGE RETRIEVAL WITH BI-SPACE QUANTIZATION AND GEOMETRIC CONSISTENCY

Wengang Zhou[1], Houqiang Li[1], Yijuan Lu[2], Qi Tian[3]

Dept. of EEIS, University of Science and Technology of China[1], Hefei, P.R. China
Dept. of Computer Science, Texas State University at San Marcos[2], Texas, TX 78666
Dept. of Computer Science, University of Texas at San Antonio[3], Texas, TX 78249
zhwg@mail.ustc.edu.cn[1], lihq@ustc.edu.cn[1], yl12@txstate.edu[2], qitian@cs.utsa.edu[3]

ABSTRACT

The state-of-the-art image retrieval approaches represent image with a high dimensional vector of visual words by quantizing local features, such as SIFT, solely in descriptor space. The resulting visual words usually suffer from the dilemma of discrimination and ambiguity. Besides, geometric relationships among visual words are usually ignored or only used for post-processing such as re-ranking. In this paper, to improve the discriminative power and reduce the ambiguity of visual word, we propose a novel bi-space quantization strategy. Local features are quantized to visual words first in descriptor space and then in orientation space. Moreover, geometric consistency constraints are embedded into the relevance formulation. Experiments in web image search with a database of one million images show that our approach achieves an improvement of 65.4% over the baseline bag-of-words approach.

Index Terms— Image retrieval, large scale, quantization, spatial embedding

1. INTRODUCTION

Given a query image, our target is to search for its nearly duplicated or partially duplicated versions in a large corpus, such as million scale or even larger scale, of web images.

In image-based object retrieval, the main challenge is image variations due to 3D view-point change, illumination change, or object-class variability [8]. Web image retrieval is different from that, and the target images are usually obtained by editing the original 2D image with changes in scale, cropping, and partial occlusion, *etc*. In most cases, modifications of the original web images are often substantial and cannot be described by a single 2D transformation. Users usually take different parts from the original image and paste them into the target images with modifications, resulting in a partially duplicated image that differs from the original not only in appearance, but also in 2D spatial layout, such as those shown in Fig. 5.

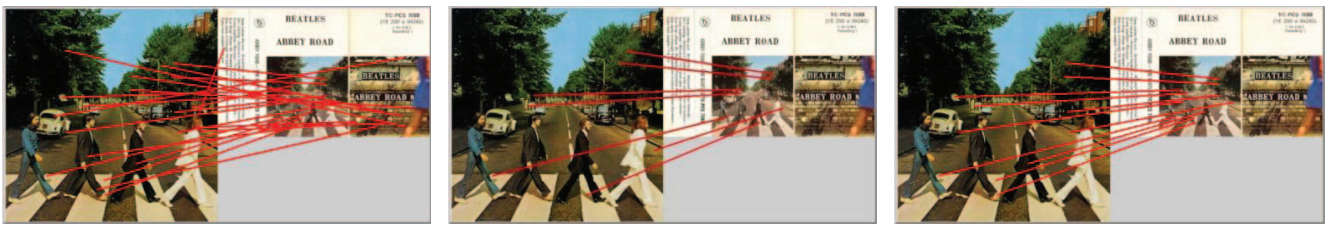
With the introduction of local features [1] for invariant representation of images and the idea of feature quantization with bag-of-words approach, large scale image retrieval systems have been greatly boosted. Besides, popular text based retrieval schemes, such as indexing with inverted file, are also leveraged for image retrieval, so as to obtain efficient indexing and fast retrieval response.

The state-of-the-art large scale image retrieval systems quantize local SIFT descriptors to visual words and then apply scalable textual indexing and retrieval schemes [2][3][4][5][7][8]. The resulting visual words usually suffer from the dilemma of discrimination and ambiguity, as shown in Fig. 1 (a) and (b). On the one hand, if the size of the visual word codebook is large enough, the ambiguity is mitigated and different local descriptors can be easily distinguished from each other. However, similar descriptors with noise pollution may be quantized to different visual words. On the other hand, with small size of code book, the variation of similar descriptors is diluted. But different descriptors may not be discriminated from each other.

Geometric verification [2][4] is very popular as a post-processing step to improve retrieval precision. But full geometric verification is computationally expensive. In practice, it is only applied to a small portion of top-ranked candidate images [6][8].

In this paper, we propose a novel scheme to quantize local features in bi-spaces, *i.e.* descriptor space and orientation space. The resulting visual words are more discriminative and less ambiguous, as shown in Fig. 1(c). Moreover, geometric constraints, including scale difference and 2D spatial layout, are imposed for formulating the relevance between images. Experiments in web image retrieval, with a database of one million images, reveal that our approach achieves an improvement of 65.4% over the baseline state-of-the-art bag-of-words approach.

Our work is closely related to [4], as both exploit the orientation information to improve retrieval performance. However, the orientation information is used in totally different perspectives. In [4], the orientation difference of matched features is used to filter false matches, while in our approach, orientation is applied for quantization. The advantage of our approach is that the error caused by orientation inconsistency will not be transferred to the later stage to confuse the checking of positive matching.



(a) (b) (c)

Figure 1. Matching of SIFT features with and without orientation quantization for nearly duplicated images. Each red line crossing two images denotes a match of a pair of local features, with the end of the line located at the key point position of local feature. (a) and (b) denote the matching without orientation quantization with a codebook of 140K and 1.4M visual words, respectively. (c) denotes the matching after orientation quantization with a codebook of 140K visual words .

2. OUR APPROACH

Our approach mainly consists of feature vector quantization and spatial embedding for relevance formulation. The quantization of local feature (SIFT) is performed in two independent spaces, i.e. descriptor space and orientation space, in a sequential manner. After quantization, spatial consistency of matched features is checked to formulate the relevance between query and database images.

2.1. Vector quantization of features

Intuitively, we can quantize a SIFT feature first in the descriptor space and then in the orientation space, or in an inverse sequence. However, with soft quantization in mind, considering that the descriptor space is 128-dimensional while the orientation space is one-dimensional, it is more preferable to first quantize SIFT feature in the descriptor space in a hard mode and then in the orientation space in a soft manner.

2.1.1 Descriptor quantization

For descriptor quantization, bag-of-words approach [2] is adopted. The descriptor quantizer $q(\cdot)$ is defined to map a descriptor $x \in R^d$ to an integer index. The quantizer is often obtained by performing *k*-means clustering on a sampling set and the resulting cluster centroids are defined as visual words. The quantizer $q(x)$ is then the index of the centroid closest to the descriptor x. To perform the quantization more efficiently, a hierarchical vocabulary tree [3] is used and the resulting leaf nodes are regarded as visual words.

2.1.2 Orientation quantization

Our orientation quantization is based on an assumption that the partially duplicated target image does not undergo much affine transformation distortion with local features' orientation changing in much inconsistent manner. Under such assumption, we can further impose an orientation constraint that the query image shares similar spatial configuration with the target images. The orientation constraint can further be relaxed by rotating the query image by some angles to generate new queries, as discussed in detail in section 2.4. The retrieval results of all rotated queries can be aggregated to obtain the final results.

For each leaf node, quantization is further performed in the orientation space. To mitigate the quantization error, a soft strategy is applied. In index construction stage, all database features in each leaf node are sorted by their orientation value in ascending or descending order. Assume that the quantization number of orientation space is t, when a query feature with orientation value of v is given, we first find the nearest leaf node with the vocabulary tree. Then, database features in the leaf node with circular orientation distance to v less than π/t are considered as valid match. The significance of the orientation space quantization is that the rate of false positive match of local features will be greatly reduced.

2.2. Geometric constraints: scale and spatial consistency

2.2.1. Scale constraints

Generally, if a targeting duplicated image undergoes scale changes, the scale differences of each pair of matched features should be the same. Based on such observation, false matches can be filtered by checking the scale consistency [4]. For each pair of matched images, a histogram of scale differences between each pair of matched features is constructed. The peak of the histogram can be determined and the matched pair with scale difference far away from the histogram peak will be discarded.

2.2.2. Spatial constraints

With orientation assumption, the spatial consistency can be loosely checked by the spatial layout of matched features. Let $p = \{p_i\}$ and $q = \{q_i\}$ $(i = 1 \sim m)$ be the m matched features belonging to the target image and the query image, respectively. In [8], two geometric inconsistency terms, $M^X(q;p)$ and $M^Y(q;p)$ denoting the geometric inconsistency order in X- and Y-coordinates respectively, are defined for local bundling feature. Here, we adopt the same terms for checking the geometric consistency of matched features in the whole image, instead of only a MSER region. The philosophy behind our idea is that with quantization in both descriptor space and orientation space,

the matched features are regarded as positive with a high probability. If two images contain many such matched features with spatial consistency, then they will be duplicates with a high probability. Consequently, the spatial consistency factor is defined as

$$f = \frac{m}{\max(1, \max(M^X(q;p), M^Y(q;p)))} \quad (1)$$

2.3. Definition of scoring

The relevance of a database image to the query image is determined by their matched features and the geometric consistency of the matching. We formulate the relevance definition as a voting problem. Each visual word in the query image votes on its matched images. And the voting value is weighted by the spatial consistency factor and the number of local features in the corresponding target image. Suppose a query image and a target image share m visual words with consistency of scale difference, the spatial consistency factor is f and the SIFT number of the candidate target image is N, we define the relevance as follows,

$$v = m \cdot f / \log(N) \quad (2)$$

2.4. Relaxing of orientation constraints

To relax the orientation constraints in section 2.1.2, we can rotate the query image by a few pre-defined angles so that all possible orientation changes from the original image will be covered. Each rotated version is used as query and the aggregation results are returned as the final retrieval results. In fact, the query image does not need to be rotated, since the SIFT features of each rotated query share the same descriptors as original query but differ in orientation value. Therefore, we just need to change the orientation value and compute the new spatial location for each query feature. After that, quantization is performed. It should be noted that the quantization in descriptor space needs to be performed only once. For each candidate target image, assume the relevance to the *i-th* rotated query is v^i, the aggregation of all rotated versions is defined as the maximum one, $v = \max(v^i \mid i = 1 \sim t)$. Finally, all images are sorted according to their aggregated relevance value.

3. EXPERIMENTAL RESULTS

We build our basic dataset by crawling one million images that are most frequently clicked in a popular commercial image-search engine. Then, we collected and manually labeled 737 partially duplicated web images from 10 groups and the images in each group are partial duplicates of each other. There are no exact or very near-exact duplicates in these images. Fig. 5 shows some typical examples. We add these labeled images into the basic dataset to construct an evaluation dataset. Since the basic dataset contains partial duplicates of our ground truth dataset, for evaluation purpose we identify and remove these partially duplicated images from the basic dataset by querying the database with every image from the ground-truth dataset, and manually checking the returned images sharing any common visual words with the query images.

To evaluate the performance with respect to the size of dataset, we also build three smaller datasets (50K, 200K, and 500K) by sampling the basic dataset. In our evaluation, 50 representative query images are selected from the ground truth dataset. Following [6], we use mean average precision (mAP) as our evaluation metric.

Impact of orientation quantization size. To study the impact of orientation quantization, we experiment with different step sizes on different sizes of image dataset. The performance of mAP for different orientation quantization sizes are shown in Fig. 2. For each dataset, when the quantization size increases, the performance first increases and then keeps stable with a little drop, while the time cost increases linearly. In our experiment, we select the orientation quantization size as 7.

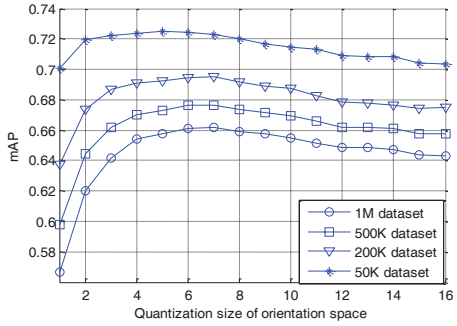


Figure 2. Performance (mAP) of our approach on different image datasets for different orientation quantization size.

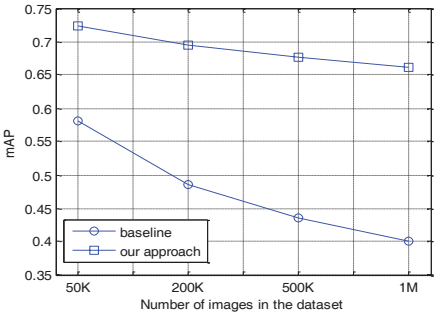


Figure 3. Performance comparison of our approach and baseline under different size of dataset with a vocabulary tree of 140K visual words.

Comparison with baseline. We use vocabulary tree approach [3] as the baseline for comparison. Fig. 3 illustrates the performance of the baseline and our approach (orientation quantization size: 7) with the same visual codebook (in descriptor space) of 140K visual words on different databases. When the size of the database increases,

the mean average precision (mAP) drops for both approaches. The baseline's mAP decreases steeper than that of our approach, reflecting the discriminative power of orientation quantization. On the 1M dataset, the mAP for our approach is 67.7%, much better than the 40% performance of the baseline.

Impact of visual word number. We test the performance of our approach with different visual codebook size under the same orientation quantization setting on different size of dataset. From Fig. 4, it can be seen that when the codebook size increases from 140K to 1.4M, the performance will drop. This is due to the fact that too fine quantization of descriptor space will be sensitive to small variation, such as noise, and false negative rate of matched features will increase. This phenomenon can also be observed in Fig. 1.

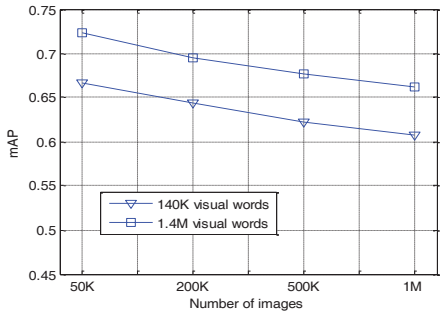


Figure 4. Performance of our approach with different size of visual word codebook.

Runtime. We perform the experiments on a server with 3.0G Hz CPU and 32G memory. Table 1 shows the average time for one image query. The time for SIFT feature extraction is not included. Compared with the baseline, our approach is more time-consuming. The time cost of our approach is mainly due to the multi-query of orientation rotation of the original query. In our experiment, the orientation quantization number is 7, which means 6 more queries are involved. In fact, it has been observed in our experiment that most target images share the same orientation configuration to the query. If such constraint is imposed such that only the original query is adopted, then retrieval time cost will be comparable with the baseline, with some small drop in performance.

4. CONCLUSION

We propose to perform SIFT feature quantization in bi-spaces and impose spatial consistency constraints for large scale partially duplicated web image retrieval. Quantization in both descriptor space and orientation space can better help to discriminate local feature with less ambiguity. After quantization, the geometric consistency is embedded for image relevance formulation. In the future, we will consider soft quantization in descriptor space and reduce the quantization loss further. To enhance the quantization, hamming embedding [4] can also be exploited in the descriptor quantization, such that the loss in descriptor quantization can be further reduced.

Our approach is based on an assumption that the partial-duplicated target image does not undergo much affine transformation distortion that incurs the orientation changes of image local features in a much inconsistent manner. The tolerance for affine distortion will be explored in the future.

	Baseline	Our approach
Time cost	0.42s	4.80s

Table 1. Average time cost per query image (not including the time for feature extraction) on 1M database.

Figure 5. Example of retrieval results on the one-million-image dataset. Queries are show on the left of the arrows and typical retrieved images (selected from those before the first false positive) are shown on the right.

Acknowledgement

The work is supported in part by NSFC No. 60632040, 60672161, in part by 863 Program No. 2006AA01Z317, in part by DHS Grant N0014-07-1-0151, and in part by the start-up funding from Texas State University.

5. RERERENCE

[1] D. Lowe, "Distinctive image features from scale-invariant key points," *IJCV*, 60(2):91-110, 2004.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," In Proc. *ICCV*, 2003.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," In Proc. *CVPR*, pages 2161-2168, 2006.

[4] H. Jegou, et al, "Hamming embedding and weak geometric consistency for large scale image search," In Proc. *ECCV,* 2008.

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," In Proc. ICCV, 2007.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," In Proc. CVPR, 2007.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," In Proc. *CVPR*, 2008.

[8] Z. Wu, et al, "Bundling Features for Large Scale Partial-Duplicate Web Image Search," in Proc. *CVPR*, 2009.