Large Scale Image Search with Geometric Coding

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[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
Bag-of-Visual-Words model is popular in large-scale image search. However, traditional Bag-of-Visual-Words model does not capture the geometric context among local features in images. To fully explore geometric context of all visual words in images, efficient global geometric verification methods are demanded. In this paper, we propose a novel geometric coding algorithm to encode the spatial context among local features of an image for large scale partial duplicate image retrieval. Our approach is not only computationally efficient, but also can effectively detect duplicate images with rotation, scale changes, occlusion, and background clutter with low computational cost. Experiments show the promising results of our approach.

Categories and Subject Descriptors
I.2.10 [Vision and Scene Understanding]: VISION

General Terms
Algorithms, Experimentation, Verification.

Keywords
Image retrieval, partial-duplicate, large scale, rotation-invariant, geometric square coding, geometric fan coding.

1. INTRODUCTION
With the emergence of Tineye [1] and Google Image Search [2], partial-duplicate image search has been attracting more and more attention in recent years. Partial-duplicate images are referred as those images, part of which are usually cropped from the same original image, but edited with modification in color, scale, rotation, partial occlusion, *etc*. Fig. 1 shows some instances of partial-duplicate Web images. Partial-duplicate image search can be widely used in many applications, such as image/video copyright violation detection, tracking the appearance of an image online and duplicate image annotation.

Large scale image retrieval [5~17] with local features has been significantly improved based on Bag-of-Visual-Words (BOW) model. BOW model achieves scalability for large scale image retrieval by quantizing local features to visual words. Popular local features include SIFT [4], MSER [23], *etc*. Local feature quantization makes image representation very compact. However, it also reduces the discriminative power of local descriptors. And the unavoidable quantization error will cause false matches of local features between images and decrease retrieval accuracy.

Figure 1. Examples of partial-duplicate Web images.

To reduce the quantization error, some approaches improve the discrimination power of local features, such as soft-quantization [12, 15], Hamming Embedding [6]. Some other approaches focus on utilizing geometric information in images to improve retrieval precision in a pre-processing or post-processing way.

The motivation of pre-processing approaches is to encode spatial context of local features into image representation. In [17], a spatial-bag-of-features scheme is used to encode geometric information of objects within an image and generate ordered bag-of-features for image search. Due to the large amount of local features in images, it is hard for the pre-processing approaches to fully encode various spatial relationships.

To avoid these problems, post-processing approaches use geometric consistency to filter those false matches. In [3], the locally spatial consistency of some spatially nearest neighbors is used to filter false matches. However, such loose geometric constraint is sensitive to the image noise from background clutter. Bundled-feature [16] is to assemble features in local MSER [20] regions to increase the discriminative power of local features. Geometric min-hashing [14] constructs repeatable hash keys with loosely local geometric information for more discriminative description. All of the above post-processing approaches only verify spatial consistency of features in local areas instead of the entire image. Although they are computationally efficient, they cannot capture the spatial relationship between all features, which makes it hard to detect all false matches.

To capture geometric relationships of all features in a whole image, global geometric verification method such as RANSAC

*Area Chair: Nicu Sebe.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA..

[11, 18, 19] is often used for this task. RANSAC can greatly improve retrieval precision. However, it is computationally expensive. It is usually applied on the subset of the top-ranked candidate images, which may not be sufficient in large scale image retrieval systems. The spatial coding approach [8] is another global geometric-verification method to remove false matches based on spatial maps. The drawback of spatial coding is that it requires that the duplicated patches in the query and the matched image share the same or very similar spatial configuration and it cannot handle rotation very efficiently.

In this paper, our motivation is to design an efficient global geometric verification approach, which can achieve rotation-invariant and is insensitive to background clutter. We propose two coding schemes, *i.e.*, geometric square coding and geometric fan coding, to encode the geometric relationships of local features for global spatial verification. Our approach can efficiently and effectively address images with free rotation changes.

2. OUR APPROACH

2.1 Geometric Coding

The spatial context among local features of an image is critical in identifying duplicate image patches. After SIFT quantization, SIFT matches between two images can be obtained. However, the matching results are usually polluted by some false matches. To refine the matching results effectively and efficiently, we propose the geometric coding scheme.

The key idea of geometric coding is to encode the geometric context of local SIFT features for geometric consistency verification. Our geometric coding is composed of two types of coding strategies, *i.e.*, geometric square coding and geometric fan coding. The difference between the two strategies lies in how the image plane is divided. Before encoding, the image has to be divided with a certain criterion that can address both rotation-invariance and scale-invariance. We design the criterion via the intrinsic invariance merit of SIFT feature.

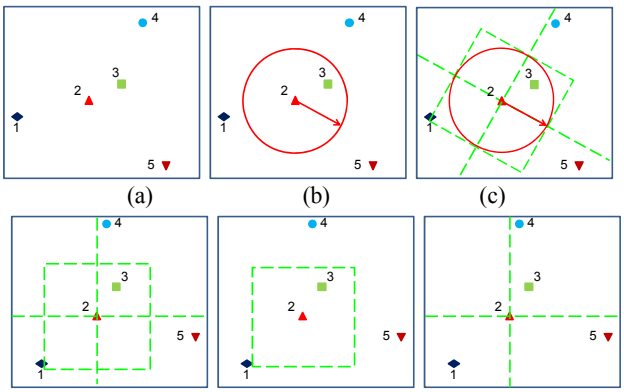

Figure 2. Illustration of image plane division. (a) Five SIFT features in image; (b) Key point of feature 2 displayed as vector indicating scale, orientation, and location (red arrow); (c) Image plane division with lines and square (green dashed lines) with feature 2 as reference point; (d) Image plane rotation from (c); (e) and (f): Image subdivisions from (d).

Fig. 2 gives a toy example of image plane division with feature 2 as reference point. Fig. 2(b) illustrates an arrow originated from feature 2, which corresponds to a vector indicating the scale and orientation of the SIFT feature. With feature 2 as origin and direction of the arrow as major direction, two lines horizontal and vertical to the arrow are constructed. Besides, centered at feature 2, a square is also drawn along these two lines, as shown in Fig. 2(c). For comparison convenience, we rotate the locations of all features to align the arrow to be horizontal, as shown in Fig. 2(d). After that, the image plane division with the two lines and the square can be decomposed into two kinds of sub-divisions, as shown in Fig. 2(e) and (f), which are used for geometric square coding and geometric fan coding, respectively.

2.1.1 Geometric Square Coding

Geometric square coding (GSC) encodes the geometric context in axial direction of reference features. In GSC, with each SIFT feature as reference center, the image plane is divided by regular squares. A square coding map, called *S*-map, is constructed by checking whether other features are inside or outside of the square.

To achieve rotation-invariant representation, before checking relative position, we have to adjust the location of each SIFT feature according to the SIFT orientation of each reference feature. For instance, given an image I with M features $\{f_i(x_i, y_i)\}$, $(i = 1,2,\ldots,M)$, with feature $f_i(x_i, y_i)$ as reference point, the adjusted position $f_j^{(i)}(x_j^{(i)}, y_j^{(i)})$ of $f_j(x_j, y_j)$ is formulated as follows:

$$\begin{pmatrix} x_j^{(i)} \\ y_j^{(i)} \end{pmatrix} = \begin{pmatrix} \cos(\phi_i) & -\sin(\phi_i) \\ \sin(\phi_i) & \cos(\phi_i) \end{pmatrix} \cdot \begin{pmatrix} x_j \\ y_j \end{pmatrix}, \quad 0 \le i, j < M \qquad (1)$$

where ϕ_i is a rotation angle equal to the SIFT orientation of the reference feature f_i.

S-map describes whether other features are inside or outside of a square defined by the reference feature. For image I, its *S*-map is defined as:

$$Smap(i,j) = \begin{cases} 1 & \text{if } \max\left(\left|x_j^{(i)} - x_i^{(i)}\right|, \left|y_j^{(i)} - y_i^{(i)}\right|\right) < s_i \\ 0 & otherwise \end{cases} \qquad (2)$$

where s_i is a half-square-length proportional to SIFT scale of feature f_i : $s_i = \alpha \cdot scl_i$, α is a constant.

To more strictly describe the relative positions, we advance to general squared maps. For each feature, n squares are drawn, with an equally incremental step of the half side length on the image plane. Then, the image plane is divided into $(n+1)$ non-overlapping parts. Correspondingly, according to the image plane division, a generalized geo-map should encode the relative spatial positions of feature pairs. The general *S-map* is defined as follows,

$$GS(i,j) = \frac{\max\left(\left|x_j^{(i)} - x_i^{(i)}\right|, \left|y_j^{(i)} - y_i^{(i)}\right|\right)}{s_i} \qquad (3)$$

where s_i is the same as that in Eq. (2), $k = 1,2,\ldots,r$.

2.1.2 *Geometric Fan Coding*

Geometric square coding (GSC) encodes the geometric context perpendicular to axial direction of reference features. In geometric

fan coding, we take each SIFT feature as reference point and divide the image plane into some regular fan regions. Then two fan coding maps, *i.e.*, *H*-map and *V*-map, are constructed by checking which fan region other features fall into.

Based on the adjusted new positions of SIFT feature in Eq. (1), two binary geometric maps, called *H*-map and *V*-map, are generated. *H*-map and *V*-map describe the relative spatial positions between each feature pair along the horizontal and vertical directions, respectively. They are formulated as follows,

$$H(i,j)=\begin{cases}0 & \text{if } x_j^{(i)} \le x_i^{(i)} \\ 1 & \text{if } x_j^{(i)} > x_i^{(i)}\end{cases} \tag{4}$$

$$V(i,j)=\begin{cases}0 & \text{if } y_j^{\,i} \le y_i^{\,i} \\ 1 & \text{if } y_j^{\,i} > y_i^{\,i}\end{cases} \tag{5}$$

We can put forward the geometric fan coding to more general formulations to impose stricter geometric constraints. The image plane is divided into $4 \cdot r$ parts, with each quadrant evenly divided into r fan regions. We decompose the division into r independent sub-divisions, each dividing the image plane into four quadrants. Each sub-division is encoded independently and their combination leads to the final fan coding maps.

The general fan coding maps *GH* and *GV* are both 3-D and defined as follows. With feature f_i as reference, the location of feature f_j is rotated counterclockwise by $\theta_i^{(k)}=\frac{k\cdot\pi}{2\cdot r}+\phi_i$ degree $(k=0,1,...,r-1)$ according to the image origin point, yielding the new location $f_j^{(i,k)}(x_j^{(i,k)},y_j^{(i,k)})$. ϕ_i is the SIFT orientation angle of f_i, as used in Eq. (1). Then *GH* and *GV* are formulated as,

$$GH(i,j,k)=\begin{cases}0 & \text{if } x_j^{(i,k)} \le x_i^{(i,k)} \\ 1 & \text{if } x_j^{(i,k)} > x_i^{(i,k)}\end{cases} \tag{6}$$

$$GV(i,j,k)=\begin{cases}0 & \text{if } y_j^{(i,k)} \le y_i^{(i,k)} \\ 1 & \text{if } y_j^{(i,k)} > y_i^{(i,k)}\end{cases} \tag{7}$$

2.2 Spatial Verification

Denote that a query image I_q and a matched image I_m are found to share N pairs of matched features through SIFT quantization. Then the corresponding sub-geo-maps of these matched features for both I_q and I_m can be generated and denoted as (GS_q,GH_q,GV_q) and (GS_m,GH_m,GV_m) by Eq. (3), Eq. (6) and Eq. (7), respectively. After that, the comparison of geometric maps is performed as follows. We do logical Exclusive-OR (XOR) operation on GH_q and GH_m, GV_q and GV_m, respectively:

$$V_H = GH_q \oplus GH_m\,; \qquad V_V = GV_q \oplus GV_m \tag{8}$$

If some false matches exist, the entries of these false matches on GH_q and GH_m may be inconsistent, and so are that on GV_q and GV_m. These inconsistencies will cause the corresponding exclusive-OR result of V_H and V_V to be 1. We define the inconsistency from geometric fan coding as follows,

$$F_H(i,j)=\bigcup_{k=1}^{r} V_H(i,j,k)\;; \quad F_V(i,j)=\bigcup_{k=1}^{r} V_V(i,j,k) \tag{9}$$

The inconsistency from geometric square coding is defined as:

$$F_S(i,j)=\left|GS_q(i,j)-GS_m(i,j)\right| \tag{10}$$

Consequently, by checking F_H, F_V and F_S, the false matches can be identified and removed. Denote:

$$T(i,j)=\begin{cases}1 & \text{if } F_S(i,j)>\tau \text{ and } F_H(i,j)+F_V(i,j)>\beta \\ 0 & \text{otherwise}\end{cases} \tag{11}$$

where β and τ are constant integers.

Ideally, if all N matched pairs are true positives, T will be zero for all entries. If false matches exist, the entries of these matches on those geometric coding maps may be inconsistent. We can iteratively remove such match that causes the largest inconsistency, until all remained matches are consistent to each other.

3. EXPERIMENTS

We construct our basic dataset by crawling one million images from the Web. To build the ground truth dataset, we collected and manually labeled 1104 partial-duplicate Web images of 33 groups from the Web. These ground-truth images are shared to the public and can be downloaded from: [21]. In our experiments, 100 representative query images are selected from the ground truth dataset for evaluation comparison. We use mean average precision (mAP) to evaluate the accuracy performance of all experiments.

We use an inverted-file index structure to index images. Each visual word is followed by an entry in the index that contains the list of images in which the visual word appears. For each indexed feature, we store its image ID, SIFT orientation, scale and the *x*- and *y*- coordinate, which will be used for generating geometric coding maps for retrieval. Similar to [8], the image similarity is formulated by the number of true matches.

Two approaches are considered for comparison. The first one is the Bag-of-Visual-Words approach with visual vocabulary tree [3], denoted as the "baseline" approach. The second one is re-ranking via geometric verification, which is based on the estimation of an affine transformation by a variant of RANSAC [19] as used in [11]. We call this method "RANSAC". In the experiment, all candidate images are involved in the RANSAC-based re-ranking.

We perform the experiments on a server with 2.4 GHz CPU and 8 GB memory. Fig. 3 illustrates the mAP performance of the comparison algorithms and our geometric coding (GC) approach. Table 1 shows the average time cost per query of all approaches. The time cost of SIFT feature extraction is not included.

Compared with the baseline, our approach is more time-consuming, since it is involved with geometric coding and verification. It takes the baseline 0.095 second to perform one image query on average, while for our approach the average query time cost is 0.155 second, 0.06 second more than the baseline. However, our approach increases the MAP from 0.37 to 0.54, a 46% improvement over the baseline. RANSAC is the most

time-consuming approach, due to the affine estimation from many random samplings. It costs 1.052 second on average per query, which is 6.7 times more than our approach. Also, it is notable that our approach achieves even better mAP performance than the "RANSAC" method.

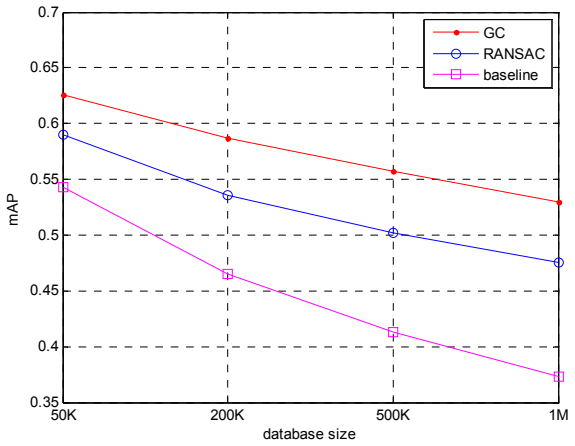

Figure 3. Comparison of mAP for different methods on the 1M database. (Best viewed in color PDF)

Table 1. The average time cost of the comparison methods and our geometric coding (GC) approach.

approach	baseline	RANSAC	GC
time cost (s)	0.095	1.052	0.155

4. CONCLUSION

In this paper, we propose a novel geometric coding scheme for SIFT match verification in large scale partial-duplicate image search. The geometric coding consists of geometric square coding and geometric fan coding. It efficiently encodes the relative spatial locations among features in an image and effectively discovers false feature matches between images. Our approach can effectively detect duplicate images with rotation, scale changes, occlusion, and background clutter.

In our future work, we will study better quantization strategy for visual codebook generation. Also, we will move on from duplicate image search to video copy detection, and explore its potential on image and video annotation [22].

5. ACKNOWLEDGMENTS

This work was supported in part to Dr. Li by NSFC under contract No. 60672161 and 863 Program under contract No. 2006AA01Z317, in part to Dr. Lu by Research Enhancement Program (REP) and start-up funding from the Texas State University, and in part to Dr. Tian by NSF IIS 1052851, Faculty Research Awards by Google, FXPAL and NEC Laboratories of America, respectively.

6. REFERENCES

[1] http://www. Tineye.com

[2] http://similar-images.googlelabs.com/

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

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

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

[6] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008.

[7] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive Visual Words and Visual Phrases for Image Applications. In *Proc. ACM Multimedia*, 2009.

[8] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. Spatial coding for large scale partial-duplicate web image search. In *Proc. ACM Multimedia*, 2010.

[9] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, and Q. Tian. Building Contextual Visual Vocabulary for Large-scale Image Applications. In *Proc. ACM Multimedia*, 2010.

[10] 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.

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

[12] 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.

[13] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *Proc. CIVR*, 2007.

[14] O. Chum, M. Perdoch, and J. Matas. Geometric minhashing: Finding a (thick) needle in a haystack. In *Proc. CVPR*, 2009.

[15] H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Proc. CVPR*, 2007.

[16] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling Features for Large Scale Partial-Duplicate Web Image Search. In Proc. CVPR, 2009.

[17] Y. Cao, C. Wang, Z. Li, L. Zhang, L. Zhang. Spatial-bag-of-features. In Proc. CVPR, 2010.

[18] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. of the ACM, 24: 381–395, 1981.

[19] O. Chum, J. Matas, and S. Obdrzalek. Enhancing RANSAC by generalized model optimization. In Proc. ACCV, 2004.

[20] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Proc. BMVC, 2002.

[21] http://home.ustc.edu.cn/~zhwg/download/DupGroundTruth Dataset.rar .

[22] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song. Unified Video Annotation via Multi-Graph Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, 2009.