2010 International Conference on Pattern Recognition

Adding Affine Invariant Geometric Constraint for Partial-Duplicate Image Retrieval

Zhipeng Wu[1], Qianqian Xu[1], Shuqiang Jiang[2], Qingming Huang[1], Peng Cui[2], Liang Li[2]

[1]*Graduate University of Chinese Academy of Sciences, Beijing, 100049, China*
[2]*Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China*
{zpwu, qqxu, sqjiang, qmhuang, pcui, lli}@jdl.ac.cn

***Abstract*—The spring up of large numbers of partial-duplicate images on the internet brings a new challenge to the image retrieval systems. Rather than taking the image as a whole, researchers bundle the local visual words by MSER detector into groups and add simple relative ordering geometric constraint to the bundles. Experiments show that bundled features become much more discriminative than single feature. However, the weak geometric constraint is only applicable when there is no significant rotation between duplicate images and it couldn't handle the circumstances of image flip or large rotation transformation. In this paper, we improve the bundled features with an affine invariant geometric constraint. It employs area ratio invariance property of affine transformation to build the affine invariant matrix for bundled visual words. Such affine invariant geometric constraint can cope well with flip, rotation or other transformations. Experimental results on the internet partial-duplicate image database verify the promotion it brings to the original bundled features approach. Since currently there is no available public corpus for partial-duplicate image retrieval, we also publish our dataset for future studies.**

Keywords-affine invariant geometric constraint; partial-duplicate image retrieval; internet partial-duplicate image database

I. Introduction

The notion of partial-duplicate images simply refers to the images which share the identical sub-area copies of an original. However, they are "partial" duplicate which implies that the duplicate areas are only parts of the whole images and located in different spatial regions with various kinds of transformations (e.g. rotation, scaling). One example that can be found in our daily experiences is about brand logo. In Figure 1-a, although they are not duplicate images, we still notice the perceivable connection among them given by the partial-duplicate areas of Nike logo.

Recently, with the rapid development of multimedia technology, manually generated partial-duplicate images have blossomed into a worldwide popularity on the internet. People like to use the software such as Photoshop to create interesting pictures. Moreover, Image Personalization Websites [1, 2, 3] which provide an easier way to generate partial-duplicate images begin to come into vogue. Figure 1-b simply illustrates some partial-duplicate images generated from these websites. The users just need to upload the original picture and select a template. Then the websites will generate the partial-duplicate pictures automatically.

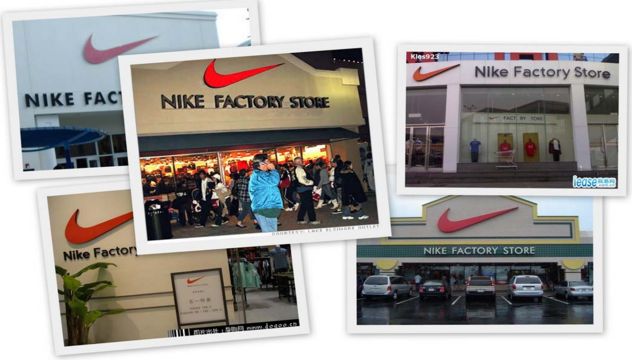
Figure 1-a. Partial-duplicate images in daily lives.

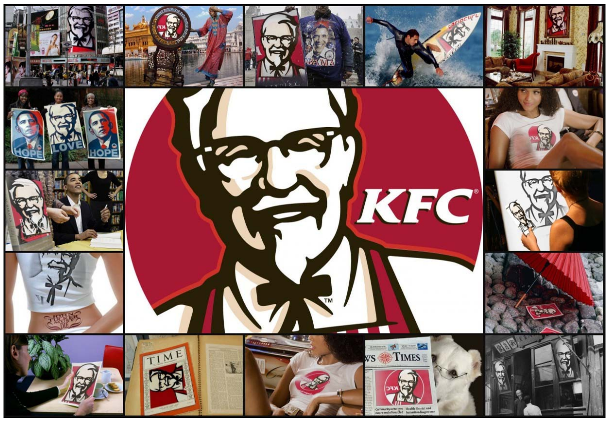
Figure 1-b. Websites generated images.
The one in the center is the source.

The emergence of partial-duplicate images brings a new challenge to the traditional image retrieval systems. Because the duplicate areas are only located at local regions, in such circumstances, global features (e.g. global color histogram

This work was supported in part by National Natural Science Foundation of China: 60833006, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.212

[4]) may lose the discriminative power. Alternatively, the proposal of local features such as SIFT [5] provides a much more promising orientation. In order to cope with large number of extracted features, local-sensitive hashing is adopted to index the feature descriptors [6]. To match the feature descriptors, [7] propose a one-to-one symmetric matching algorithm and [8] employ multi-level spatial matching.

State-of-the-art large scale image retrieval systems analogy the retrieval task with text indexing and retrieval schemes. They quantize SIFT features, treat the image as a collection of visual words [9] and build scalable vocabulary tree [10]. While quantization limits the discriminative power and the ignorance of geometric relationships among visual words remains a problem, geometric verification [11, 12] becomes an important post-processing step to refine retrieval precision. Due to the high computation complexity for full geometric verification on large scale image database, how to improve the efficiency and implement a framework for images, especially partial-duplicates comes into a hot topic.

To better fit the requirements of partial-duplicates, researchers bundle the visual words into groups instead of taking all of them as a whole [13]. By the detector of Maximally Stable Extremal Regions (MSER, [14]), each group of bundled features becomes much more discriminative than a single feature and the relative ordering relationship provides an efficient geometric constraint.

Although experiments on a large web image database show that bundled features promote the retrieval efficiency and precision on partial-duplicate image, the geometric relationship in it is still unconvinced. Intuitively, the relative ordering relationship of visual words is not rotation invariant, and the original approach of bundled features is only under the assumption of no significant rotation between duplicate images. In fact, we notice that under many circumstances, there are large rotations/flips occurring on web partial-duplicate images. In this paper, we improve the bundled features with an affine invariant geometric constraint. It employs the area ratio invariance property of affine transformation to build the affine invariant matrix for bundled visual words [15]. Such affine invariant geometric constraint can cope well with flip, rotation or other transformations, and we verify the promotion it brings to the original bundled features approach on the internet partial-duplicate image database.

II. Bundled Features and Relative Ordering Relationship

A. *Bundling point features by region features*

The idea of bundled features [13] is motivated by two popular image features: SIFT [5] and MSER [14]. While holding the powerful discriminative ability for image local regions, SIFT and MSER operates on different levels of local representation. SIFT detects interest point and describe the scale-invariant region centered on it. Instead, MSER detects affine-covariant stable region and takes the elliptical region as the unit to be described. The notion of bundled features is simply using region features (MSER) to bundle point features (SIFT) into groups which is a flexible representation that facilities partial matches. Figure 2 shows an example of bundled features.

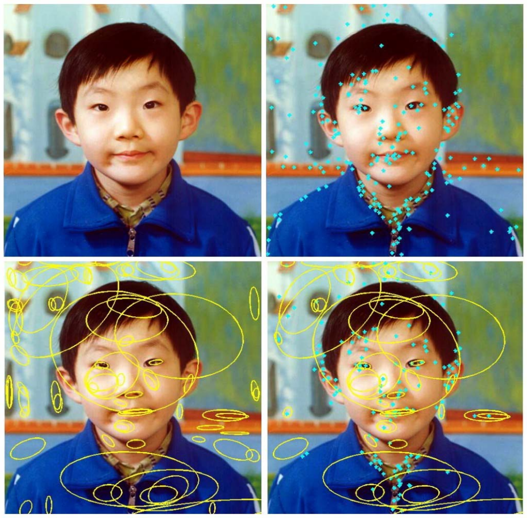

Figure 2. Bundling point features by region features.

Recent image retrieval approaches usually quantize SIFT features into visual words for better efficiency. However, in large scale image retrieval, one single feature has to be compared with millions or billions of features which may suffer from the loss of discriminative power caused by the quantization step. Motivated by the problem of mismatching SIFT features, the features are bundled into groups and employ group matching instead of single matching. Previous studies show the remarkable distinctness and repeatability of MSER. The bundled features $B = \{b_i\}$ can be defined as:

$$b_i = \{s_j \mid s_j \propto r_i, s_j \in S\} \tag{1}$$

where $S = \{s_j\}$ is SIFT features and $R = \{r_i\}$ is the MSER. $s_j \propto r_i$ means the point feature s_j falls inside region r_i.

The bundled features approach provides a more robust solution than single SIFT feature matching. Moreover, it allows partial group matching among image feature collections which is suitable for partial-duplicate image retrieval. As mentioned above, to obtain a better retrieval precision, we add geometric constraint to the features bundled together.

B. *Relative ordering relationship*

Assuming **p** and **q** are the two bundled features to be matched, the similarity score **M (q; p)** is closely related to the number of matched visual words and the geometric location consistency of the visual words in the two bundles. [13] defines the similarity score **M (q; p)** as:

$$\mathbf{M}\ (\mathbf{q};\ \mathbf{p}) = \mathbf{M}_{\mathbf{m}}\ (\mathbf{q};\ \mathbf{p}) + \boldsymbol{\lambda} \times \mathbf{M}_{\mathbf{g}}\ (\mathbf{q};\ \mathbf{p}) \tag{2}$$

where $\mathbf{M}_{\mathbf{m}}$ **(q; p)** denotes the membership term. It relies on the number of common visual words between two bundles. $\mathbf{M}_{\mathbf{g}}$ **(q; p)** denotes the geometric term. A simple way to implement it is by calculating the relative order relationship of the matched visual words on X- and Y- coordinates. Take

Figure 3-a as an example. By counting the visual words in bundle **p** and **q**, we number them in numerical order (#1, #2…). The relative order relationship (matching order) from **p** to **q** along X- coordinate is (#1, #3, #4) which results in geometric inconsistency 0. Similarly, in Figure 3-b, geometric inconsistency is -1. The $\mathbf{M_g}$ **(q; p)** is defined as the minimum value of inconsistency on X- and Y- coordinates. It is no larger than 0 and λ in (2) is the weighting parameter.

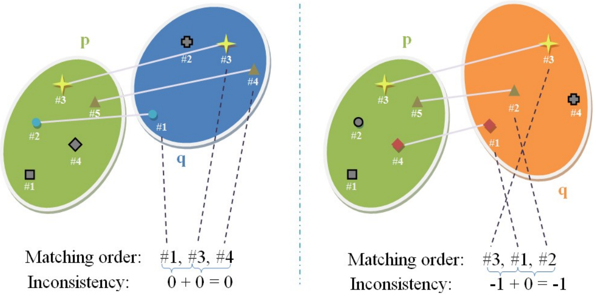
p
q
#1
#2
#3
#4
#5
Matching order: #1, #3, #4
Inconsistency: 0 + 0 = 0
p
q
#1
#2
#3
#4
#5
Matching order: #3, #1, #2
Inconsistency: -1 + 0 = -1

Figure 3. Geometric inconsistency.

III. AFFINE INVARIANT GEOMETRIC CONSTRAINT

The relative ordering relationship in [13] is sensitive to large rotations, especially under the transformations of horizontal /vertical flip. Figure 4-a and 4-b illustrate an example. Although we believe that the geometric structure should not be modified after the horizontal flip of bundle **q**, according to the figure, the relative ordering inconsistency dramatically changes from 0 to -2. Since there are only 3 common visual words between bundle **p** and **q**, the change of geometric inconsistency greatly decreases the matching precision.

To better handle this situation, researchers suggest ordering features along the dominant orientation of the bundling MSER detection [13]. However, the use of dominant orientation brings additional computational cost and only ordering the features along one direction is not robust enough as it ignores the 2-D spatial structure of bundled features. In order to mine the spatial relationship between the matched visual words, we improve the bundled features with an affine invariant geometric constraint based on the area ratio invariance property of affine transformation. To bundle **p** and **q**, supposing they share **n** common visual words (**p** $=\{s_{p1}, s_{p2},\ldots, s_{pn}\}$, **q** $=\{s_{q1}, s_{q2},\ldots, s_{qn}\}$, s_{pi} and s_{qi} are the ith matched visual words in **p** and **q** respectively), the affine invariant matrix $\mathbf{H}_{\textbf{Affine Invariant}}$ is actually an area ratio term based on the triangle generated by two visual words and the geometric center of all the features in bundle.

Having the geometric center $\vec{p}$, the triangle area matrix $\mathbf{H_p}$ (for bundle **p**) is calculated as:

$$\mathrm{H_p} = \begin{bmatrix} 0 & \mathrm{h}_{12} & \mathrm{h}_{13} & \cdots & \mathrm{h}_{1n} \\ \mathrm{h}_{21} & 0 & \mathrm{h}_{23} & \cdots & \mathrm{h}_{2n} \\ \vdots & & & & \vdots \\ \mathrm{h}_{n1} & \mathrm{h}_{n2} & \mathrm{h}_{n3} & \cdots & 0 \end{bmatrix}_{n\times n} \tag{3}$$

where element $\mathrm{h_{ij}}$ is the area size of triangle generated by vertices s_{pi}, s_{pj}, and $\vec{p}$. Intuitively, $\mathbf{H_p}$ is a square symmetric matrix with zero values along the main diagonal.

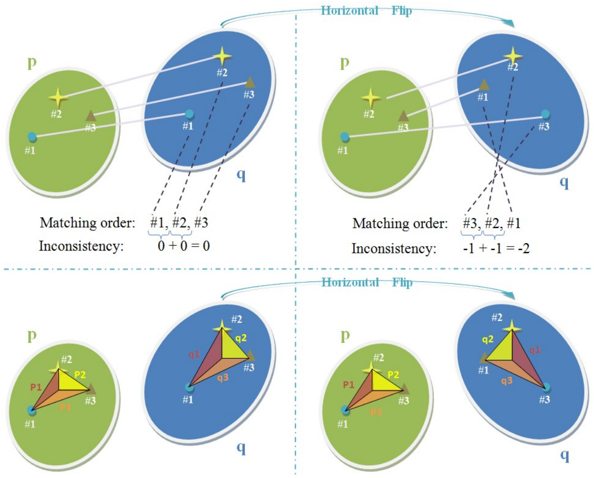

Figure 4. Examples of image horizontal flip.

The affine invariant matrix is based on the area ratio invariance property of affine transformation. Giving the largest element h_{uv} in $\mathbf{H_p}$, $\mathbf{H}_{\mathbf{p}\ \textbf{Affine Invariant}}$ is calculated by $\mathbf{H_p}$ dividing h_{uv} which preserves the area ratio invariance. Figure 4-c and 4-d illustrate an example. The area size of the triangles in bundle **p** is denoted as p1, p2, and p3; in **q** is q1, q2, and q3.The $\mathbf{H_p}$ and $\mathbf{H}_{\mathbf{p}\ \textbf{Affine Invariant}}$ is constructed as:

$$\mathbf{H_p}\begin{cases} h_{p11} = h_{p22} = h_{p33} = 0 \\ h_{p12} = h_{p21} = \mathrm{p1} \\ h_{p13} = h_{p31} = \mathrm{p2} \\ h_{p23} = h_{p32} = \mathrm{p3} \end{cases} \tag{4}$$

$$\mathbf{H}_{\mathbf{p}\ \textbf{Affine Invariant}} = \frac{1}{h_{uv}}\left[\mathbf{H_p}\right] \tag{5}$$

The geometric term $\mathbf{M_g}$ **(q; p)** in (2) is in proportion to the similarity of the two affine invariant matrixes. We then define $\mathbf{M_g}$ **(q; p)** as:

$$\mathbf{M_g}\ \mathbf{(q;\, p)} = \mathbf{n} \times \mathrm{corr}\,(\mathbf{H}_{\mathbf{p}\ \textbf{Affine Invariant}}, \mathbf{H}_{\mathbf{q}\ \textbf{Affine Invariant}}) \tag{6}$$

here **n** refers to the number of matched visual words and corr ( ) is the matrix correlation:

$$corr(A,\, B) = \frac{\sum_m \sum_n (A_{mn} - \overline{A})\,(B_{mn} - \overline{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \overline{A})^2)(\sum_m \sum_n (B_{mn} - \overline{B})^2)}} \tag{7}$$

where $\overline{A}$, $\overline{B}$ are the mean values of the elements in A and B.

Compared with the relative ordering relationship term in [13], the proposed approach takes advantage of the 2-D spatial distribution of the visual words and thus becomes more robust for geometric verification. Moreover, it is affine invariant and can cope well with large rotations and flips.

We will demonstrate the promotion it brings to the original bundled features in section 4.

IV. Experiment

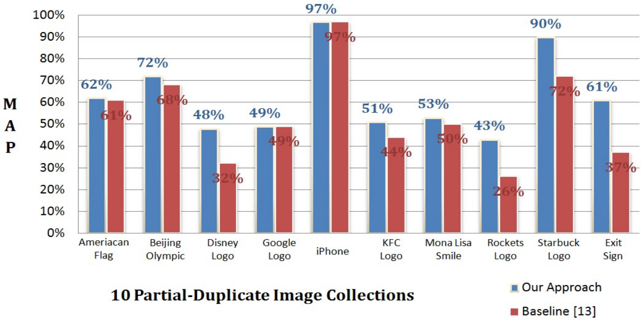

Figure 5. MAP of the queries from the 10 collections.

We introduce our published internet partial-duplicate image database (http://www.jdl.ac.cn/mova/Internet-Partial-Duplicate-Image-Database.rar) into the experiment. The internet partial-duplicate image database is consisted of 10 image collections which has 200 partial-duplicates in each. There are in all 2,000 images of the 10 collections: American Flag, Beijing Olympic Badge, Disney Logo, Google Logo, iPhone, KFC Logo, Mona Lisa Smile, Rockets Logo, Starbucks Logo, and Exit Sign. All of these images are transformed manually according to different templates provided by the Image Personalization Websites mentioned above. To construct a real online image retrieval environment, we add another 8,000 non-duplicate web images and there are altogether 10,000 images in the experiment corpus.

100 images from the 10 collections of partial-duplicates (10 collections×10 images in each) are randomly selected as the retrieval queries. We implement the original bundled features [13] as the baseline to be compared. Both of the proposed and baseline approach share the common visual vocabulary of 64,000 visual words, and the weighting parameter λ in (2) is set to 1 in our approach and 2 for [13]. Figure 5 illustrates the Mean Average Precision (MAP) of the queries from the 10 collections.

According to Figure 5, by adding an affine invariant geometric constraint, the MAP of all the queries obtained by our approach is 62.6%, which shows an obvious improvement than the baseline approach (MAP: 53.6%). Figure 6 shows a retrieval example.

V. Conclusion

The recently proposed bundled features show remarkable power for partial-duplicate images retrieval. However, the weak geometric constraint is only applicable when there is no significant rotation between the duplicates. In this paper, we introduce the affine invariant matrix to improve the original geometric verification step in bundled features. It takes advantage of the 2-D spatial distribution of the visual words and thus becomes more robust. Experiment shows the effectiveness of the proposed approach. Besides, we publish our internet partial-duplicate dataset for future studies.

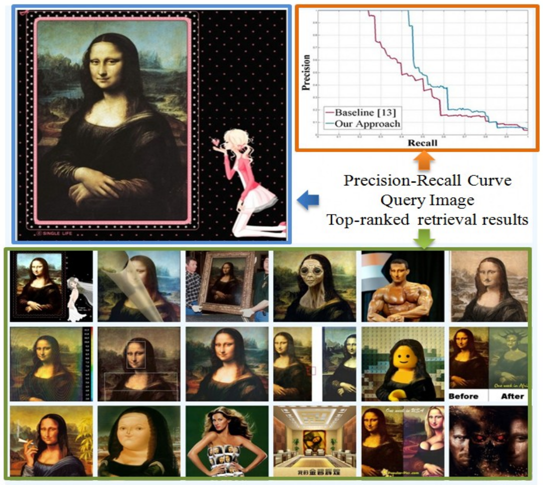

Figure 6. Retrieval example.

References

[1] Funnywow, http://www.funnywow.cn/.

[2] Keniu, http://www.keniu.com/.

[3] PhotoFunia, http://www.photofunia.com/.

[4] A. Qamra, Y. Meng, and E. Y. Chang, "Enhanced Perceptual Distance Functions and Indexing for Image Replica Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, 27(3): 379–391, 2005.

[5] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, 20: 91–110, 2003.

[6] Y. Ke, R. Sukthankar, and L. Huston, "Efficient Near Duplicate Detection and Sub-Image Retrieval," In Proceedings of ACM International Conference on Multimedia, pp. 869–876, Oct. 2004.

[7] W. L. Zhao, C. W. Ngo, H. K. Tan and X. Wu, "Near Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning," IEEE Trans. Multimedia, 9(5):1037-1048, 2007.

[8] D. Xu, T.J. Cham, S. Yan and S.-F. Chang, "Near Duplicate Image Identification with Spatially Aligned Pyramid Matching," In Proceedings of International Conference on Computer Vision, pp.1-7, Jun. 2008.

[9] J. Sivic, and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," In Proceedings of International Conference on Computer Vision, pp.1470-1477, Oct. 2003.

[10] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," In Proceedings of International Conference on Computer Vision, pp. 2161-2168, Oct. 2006.

[11] H. Jegou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search," In Proceedings of European Conference on Computer Vision, pp. 304–317, Oct. 2008.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," In Proceedings of International Conference on Computer Vision, pp. 1-8, Jun. 2007.

[13] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling Features for Large Scale Partial-Duplicate Web Image Search," In Proceedings of International Conference on Computer Vision, pp. 25-32, 2009.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdl, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," In Proceedings of British Machine Vision Conference, pp. 384-396, Sep. 2002.

[15] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2004.