Two Geometric Constraints for Bundled Features in Partial-Duplicate Image Retrieval

Zhipeng WU[†] Kiyoharu AIZAWA[†][‡]

[†] Dept. of Information and Communication Eng., The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
[‡] Interfaculty Initiative in Information Studies, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan

E-mail: { zhipengwu, aizawa }@hal.t.u-tokyo.ac.jp

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 two kinds of geometric constraints: Affine Invariant constraint (AI constraint) and Relative Saliency Ordering constraint (RSO constraint). Experimental results on the internet partial-duplicate image database verify the promotions the two geometric constraints bring to the original bundled features approach. Since currently there is no available public corpus for partial-duplicate image retrieval, the dataset is open for future studies.

Keyword Partial-Duplicate Image Retrieval, Bundled Features, Affine Invariant Constraint, Relative Saliency Ordering Constraint, Content-based Saliency Region

1. INTRODUCTION

The spring up of large numbers of online image data brings a new challenge to the content-based image retrieval (CBIR) systems. Recently, with the rapid improvement of multimedia technology, a plenty of partial-duplicate images are generated by software and image personalization websites [1-3]. 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 [1-3] 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.

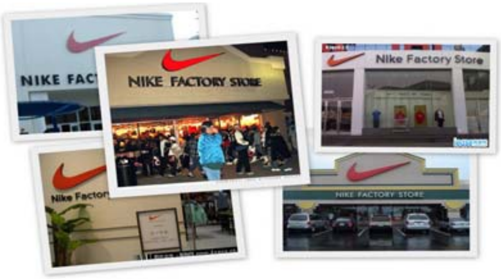


Figure 1-a Partial-duplicate images in daily lives

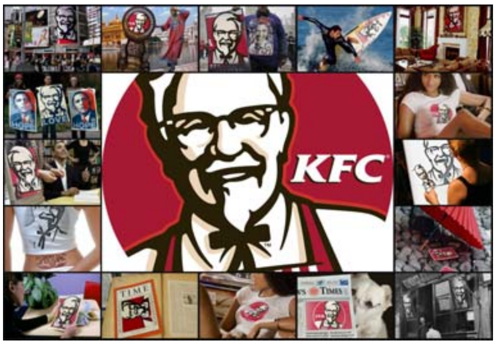


Figure 1-b Websites generated images.
The one in the center is the source

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 [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 review a further-developed geometric constraint for bundled features: Affine Invariant constraint (AI constraint) [15]. It employs the area ratio invariance property of affine transformation to build the affine invariant matrix for bundled visual words [16]. Such affine invariant geometric constraint can cope well with flip, rotation or other transformations.

In addition, we improve the bundled feature by providing a more sophisticated geometric constraint based on Relative Saliency Ordering (RSO constraint) [17]. With respect to the saliency values of the visual words, the bundle becomes more discriminative and robust for transformations such as rotation.

The rest of the paper is organized as follows: Section 2 reviews the bundled features and the relative ordering constraint. In Section 3 and 4, we introduce the AI and RSO geometric constraints as improvements for the original one used in bundled features. The experimental results can be found in Section 5, and we conclude this paper in Section6.

2. BUNDLED FEATURES

2.1. 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 performs partial matches. Figure 2 shows an example of bundled features.

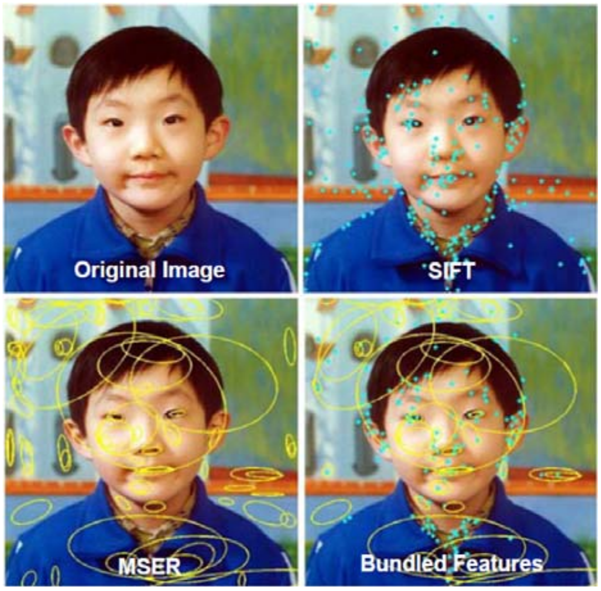


Figure 2 Bundling point features by region features

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 $B = < b_j >$ can be defined as:

$$b_j = < s_j \mid s_j \propto r_i, s_j \in S > \qquad (2.1)$$

where $S = < s_j >$ is the 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.

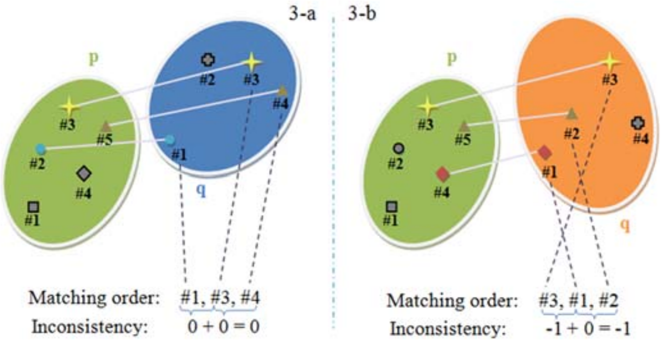
The bundled features approach provides a more robust solution than single SIFT feature matching. Moreover, it allows partial group matching among image feature collections which is suitable for partial-duplicate image retrieval. As mentioned above, to obtain a better retrieval precision, geometric constraint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

2.2. Relative Ordering Relationship Constraint

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:

$$\mathbf{M\ (q;\ p) = M_m\ (q;\ p) + \lambda \times M_g\ (q;\ p)} \qquad (2.2)$$

where **M$_m$ (q; p)** denotes the membership term. It relies on the number of common visual words between two bundles. **M$_g$ (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 a 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, the geometric inconsistency is -1. The **Mg (q; p)** is defined as the minimum value of inconsistency on X- and Y- coordinates. It is no larger than 0 and λ in (2.2) is the weighting parameter.



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

Figure 3 Geometric Consistency

3. AFFINE INVARIANT CONSTRINT

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.

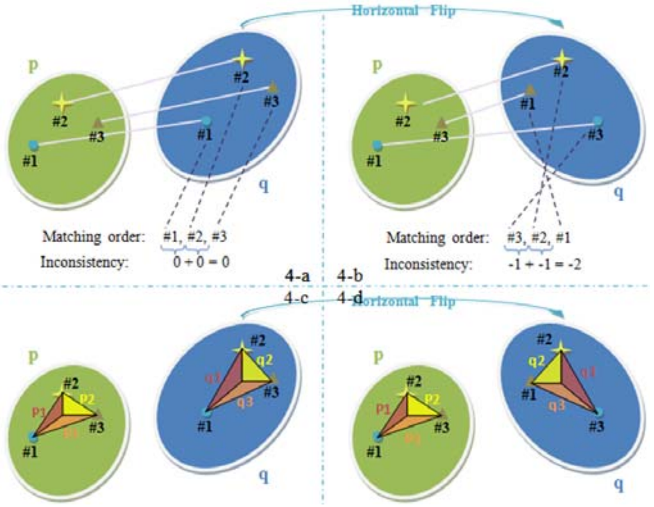


Figure 4 Examples of image horizontal flip

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 AI 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$_{p2}$, …, s$_{pn}$>, **q** = <s$_{q1}$, s$_{q2}$, …, s$_{qn}$>, s$_{pi}$ and s$_{qi}$ are the *ith* matched visual words in **p** and **q** respectively), the affine invariant matrix **H$_{\text{Affine Invariant}}$** 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 **H$_p$** (for bundle **p**) is calculated as:

$$\mathbf{H_p} = \begin{bmatrix} 0 & h_{12} & h_{13} & \cdots & h_{1n} \\ h_{21} & 0 & h_{23} & \cdots & h_{2n} \\ \vdots & & & & \vdots \\ h_{n1} & h_{n2} & h_{n3} & \cdots & 0 \end{bmatrix}_{n \times n} \qquad (3.1)$$

where element h$_{ij}$ is the area size of triangle generated by vertices s$_{pi}$, s$_{qi}$ and $\vec{p}$. Intuitively, **H$_p$** is a square symmetric matrix with zero values along the main diagonal.

The affine invariant matrix is based on the area ratio invariance property of affine transformation. Giving the largest element h_{uv} in **H$_p$**, **H$_{\text{p Affine Invariant}}$** is calculated by **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 bundle **q** is q1, q2, and q3. The $\mathbf{H_p}$ and $\mathbf{H_{p\,Affine\,Invariant}}$ is constructed as:

$$\mathbf{H_p}\begin{cases} h_{p11} = h_{p22} = h_{p33} = 0 \\ h_{p12} = h_{p21} = \mathrm{p1} \\ h_{p13} = h_{p31} = \mathrm{p2} \\ h_{p23} = h_{p32} = \mathrm{p3} \end{cases} \qquad (3.1)$$

$$\mathbf{H_{p\,Affine\,Invariant}} = \frac{1}{h_{uv}}\left[\mathbf{H_p}\right] \qquad (3.3)$$

The geometric term $\mathbf{M_g}$ **(q; p)** in Equation (2.2) is in proportion to the similarity of the two affine invariant matrixes. We then define $\mathbf{M_g}$ **(q; p)** as:

$$\mathbf{M_g\,(q;\,p)} = \mathbf{n} \times corr\,(\mathbf{H_{p\,Affine\,Invariant}},\, \mathbf{H_{q\,Affine\,Invariant}}) \qquad (3.4)$$

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

$$corr(A,B) = \frac{\sum_m\sum_n (A_{mn}-\overline{A})\,(B_{mn}-\overline{B})}{\sqrt{(\sum_m\sum_n (A_{mn}-\overline{A})^2)(\sum_m\sum_n (B_{mn}-\overline{B})^2)}} \qquad (3.5)$$

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], AI constraint 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.

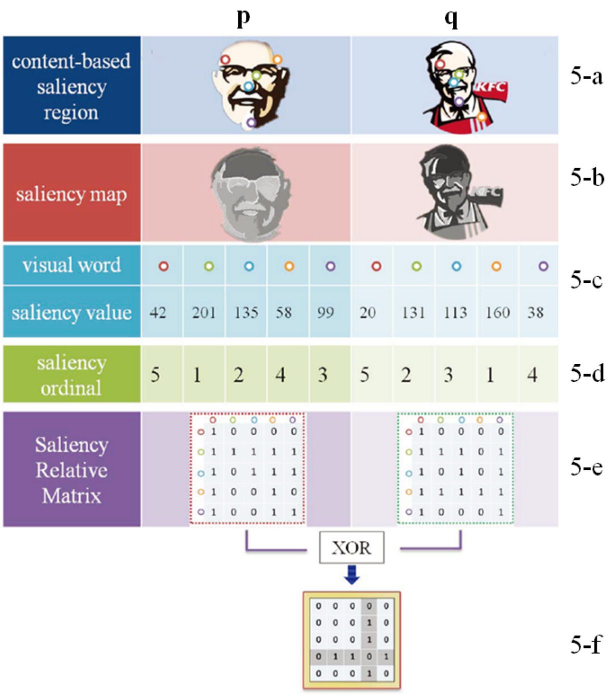
4. RSO CONSTRAINT

Image saliency analysis and visual attention has been extensively studied recently [18-20]. Motivated by both saliency analysis and bundled feature approach, we add saliency information into the process of bundle generation [17]. Moreover, by organizing the visual words according to their relative saliency order, a novel geometric constraint is proposed.

Take Figure 5 as an example. There are five matched visual words detected in bundle **p** and **q** (5-a), and the saliency values for them are shown in 5-b. One basic method is to directly rank the saliency values. According to 5-d, the saliency ordinal vectors for bundle **p** and **q** are: <5, 1, 2, 4, 3> and <5, 2, 3, 1, 4>. As we mentioned above, using relative order can make the approach robust to visual word miss-matching and detection failure. If we directly compare the two ordinal vectors, four of the total five elements in the vectors are not equal (5=5, 1≠2, 2≠3, 4≠1, 3≠4). This is not satisfying because there is only one pair of miss-matched visual words between bundle **p** and **q** (the visual word colored in orange). Therefore, with respect to relative ordering rather than the naïve order rank, we design the Saliency Relative Matrix (SRM) for each of the detected bundles:

$$SRM = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix} \qquad r_{ij} = \begin{cases} 0 & \alpha_i > \alpha_j \\ 1 & otherwise \end{cases} \qquad (4.1)$$

where the element r_{ij} in SRM is defined by the saliency values α_i of visual word v_i and α_j of visual word v_j in the bundle. The SRM is an anti-symmetric matrix which preserves the relative saliency order among visual words. Figure 5-e illustrates the SRMs calculated for bundle **p** and **q**.



	p					q				
saliency value	42	201	135	58	99	20	131	113	160	38
saliency ordinal	5	1	2	4	3	5	2	3	1	4

1	0	0	0	0
1	1	1	1	1
1	0	1	1	1
1	0	0	1	0
1	0	0	1	1

1	0	0	0	0
1	1	1	0	1
1	0	1	0	1
1	1	1	1	1
1	0	0	0	1

0	0	0	0	0
0	0	0	1	0
0	0	0	1	0
0	1	1	0	1
0	0	0	1	0

Figure 5 ROS Constraint for bundled features

By comparing the similarity between SRMs extracted from the two bundles, the Relative Saliency Ordering constraint (ROS constraint) is implemented as the matrix XOR operation, as shown in Figure 5-f.

The geometric term $\mathbf{M_g}$ **(q; p)** in Equation (2.2) is now defined as:

$$\mathbf{M_g\,(q;\,p)} = \mathbf{n} \times P(SRM_p \oplus SRM_q) \qquad (4.2)$$

where **n** is the number of matched visual words and *P*( ) refers to the proportion of '1' elements in the whole matrix. Compared with the relative ordering relationship term in [13], ROS constraint adds saliency information into the approach and thus becomes more discriminative. In addition, this constraint works well under conditions such as large rotation and horizontal/ vertical flip.

5. EXPERIMENT

5.1. Dataset

We create a public partial-duplicate image dataset for the experiments in this paper. The dataset is called "Internet Partial-Duplicate Image" [21]. 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 [1-3].

5.2. Evaluation of AI Constraint

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 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 Equation 2.2 is set to 1 in our approach and 2 for [13]. Figure 6 illustrates the Mean Average Precision (MAP) of the queries from the 10 collections.

According to Figure 6, by adding an affine invariant geometric constraint, the MAP of all the queries obtained by our approach is 62.6%, which shows significant improvement than the baseline approach (MAP: 53.6%). Figure 7 shows a retrieval example.

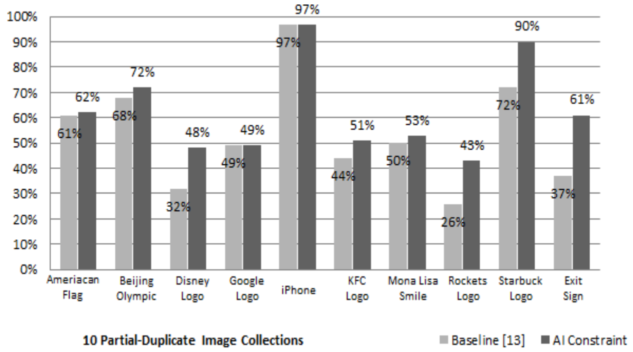


Figure 6 Evaluation of AI constraint

5.3. Evaluation of ROS Constraint

In this experiment, we add another 30,000 distracter web images into "Internet Partial-Duplicate Image" database and there are totally 32,000 images in the experiment dataset. We use a traditional BOV approach [10] as the baseline approach and a dictionary with 5,000 visual words is clustered with hierarchical *k*-means. Besides, in [22], a multi-description is designed for partial-duplicate image retrieval.

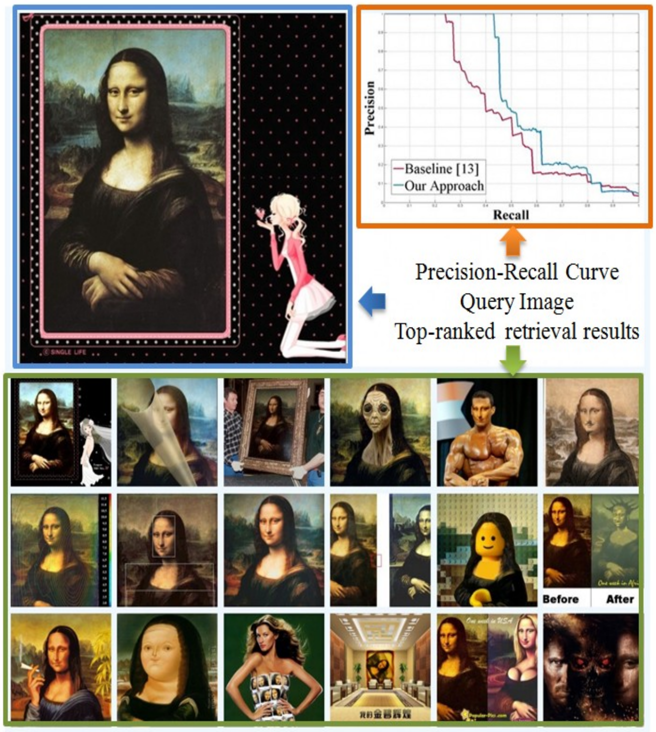


Figure 7 Partial-duplicate image retrieval example

We select 50 representative images from the 10 image collections as the queries. Following Section 5.2, Mean Average Precision (MAP) is adopted as the evaluation metric. Figure 8 illustrates the experimental results.

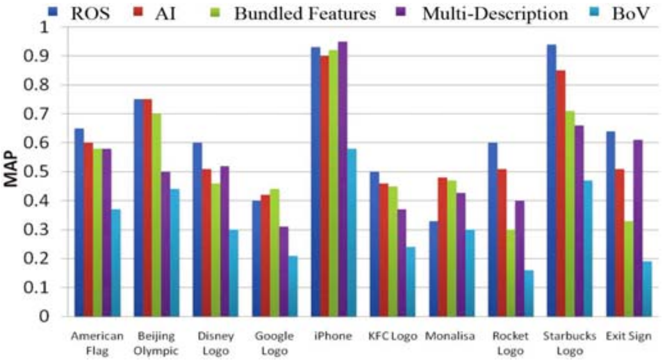


Figure 8 Evaluation of ROS constraint

According to Figure 8, comparing with other state-of-the-art methods, ROS constraint based bundled feature performs better. It's MAP is 63.4% while the MAPs received by "BOV", "bundled feature", "AI based bundled feature" and "multi-description" are 32.6%, 53.6%, 59.9% and 53.27% respectively.

6. CONCLUSION

Recently proposed bundled features show remarkable discriminative power in partial-duplicate image retrieval. However, the weak geometric constraint in the original approach is only applicable when there are no significant rotations and flips. In this paper, we improve the bundled features with two kinds of geometric constraints: Affine Invariant constraint (AI constraint) and Relative Saliency Ordering constraint (RSO constraint). Experimental results on the internet partial-duplicate image database verify the promotions the two geometric constraints bring to the original bundled features approach.

Reference

[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,” Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 379-391, 2005.

[5] D.G. Lowe, “Distinctive image features from scale invariant keypoints,” Proc. International journal of computer vision, vol. 60, no. 2, pp. 91-110, 2004.

[6] Y. Ke, R. Sukthankar, and L. Huston, “Efficient near-duplicate detection and sub-image retrieval,” Proc. ACM International Conference on Multimedia, pp., 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,” Proc. IEEE Transactions on Multimedia, vol. 9, no. 5, pp. 1037-1048, 2007.

[8] D. Xu, T.J. Cham, S. Yan, and S.F. Chang, “Near duplicate image identification with patially Aligned Pyramid Matching,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-7, 2008.

[9] J. Sivic, and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” Proc. IEEE International Conference on Computer Vision, pp. 1470-1477, 2003.

[10] D. Nister, and H. Stewenius, “Scalable recognition with a vocabulary tree,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161-2168, 2006.

[11] H. Jegou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” Proc. European Conference on Computer Vision, pp. 304-317, 2008.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, 2007.

[13] Z. Wu, Q. Ke, M. Isard, and J. Sun, “Bundling featurees for large scale partial-duplicate web image search,” Proc. IEEE Conference on Computer Vision and Pattern recognition, pp. 25-32, 2009.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” Proc. Image and Vision Computing, vol. 22, no. 10, pp. 761-767, 2004.

[15] Z. Wu, Q. Xu, S. Jiang, Q. Huang, P. Cui, and L. Li, “Adding Affine Invariant Geometric Constraint for Partial-Duplicate Image Retrieval,” Proc. International Conference on Pattern Recognition, pp. 842-845, 2010.

[16] R. Hartley, and A. Zisserman, Multiple view geometry in computer vision: Cambridge university press, 2003.

[17] L. Li, Z. Wu, Z-J. Zha, S. Jiang, and Q. Huang, “matching content-based saliency regions for partial-duplicate image retrieval,” Proc. IEEE International Conference on Multimedia and Expo, pp., 2011.

[18] L. Itti, and C. Koch, “A comparison of feature combination strategies for saliency-based visual attention systems,” Proc. SPIE human vision and electronic imaging IV (HVEI¡¯99), vol. 3644, pp. 373-382, 1999.

[19] H. Liu, S. Jiang, Q. Huang, and C. Xu, “A generic virtual content insertion system based on visual attention analysis,” Proc. ACM Intetrnational Conference on Multimedia, pp. 379-388, 2008.

[20] U. Neisser, “Cognitive psychology,”

Appleton-Century-Crofts New York, 1967.

[21] "Internet Partial-Duplicate Image,"

http://www.jdl.ac.cn/mova/Internet-Partial-Duplicate-Image-Database .rar.

[22] J.C. Van Gemert, C.G.M. Snoek, C.J. Veenman, A.W.M. Smeulders, and J.M. Geusebroek, “Comparing compact codebooks for visual categorization,” Proc. Computer Vision and Image Understanding, vol. 114, no. 4, pp. 450-462, 2010.