Partial Near-Duplicate Image Identification with Global Geometric Consistency of Subset-of-features

Peng Li
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

Han-Bing Yan
National Institute of Network and Information Security
CNCERT/CC
Beijing, China

Gang Cui
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China

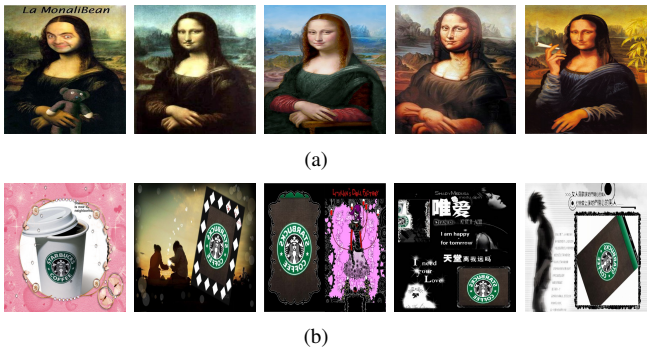
Yue-Jin Du
National Institute of Network and Information Security
CNCERT/CC
Beijing, China

***Abstract*—The task of determining whether two images are near-duplicate or not becomes increasingly important in many applications, such as copyright infringement detection, sub-image retrieval and spam image filtering. Traditional methods for near-duplicate image identification (NDII) usually extract image local features firstly, and then quantize them as bag of words (BOW); the frequency histogram is finally taken as the representation of image for NDII. However, the mismatches between local features, the lower distinctiveness, polysemy and synonymy of visual words all degrade the accuracy of NDII, especially for partial NDII. Although some geometric verification procedures have been taken, these methods are still affected by the mismatches and ambiguity of visual words. In this paper, we propose a novel scheme for verifying the global geometric consistency of subset-of-features for improving the accuracy of BOW model. If there is a subset of matched pairs of local features obtained by BOW model, in which the ratios of scales and differences of orientations are consistent, we take these two images as near-duplicate images. The cardinality of the subset can also be used for measure the similarity of the two images. Experimental results show that the proposed method can improve the accuracy of NDII prominently, and it is also effective and robust for the retrieval of some typical partial near-duplicate images.**

***Keywords*-near-duplicate image identification; sub-image retrieval; local feature; bag of visual words; geometric consistency**

I. INTRODUCTION

Near-duplicate image identification (NDII) refers to determine whether two images are near-duplicate to each other or not, but with differences in scale, rotation, translation, viewpoint, chopping and format changing, etc. With the proliferation of digital images and their widespread distribution over the Internet, NDII has become a critical issue for a variety of real-world applications, for example, copyright infringement detection, sub-image retrieval, spam image filtering, illegal use of images/videos detection [1,2]. The near-duplicate regions can be the whole image or only a small portion of the image. Fig.1 show two different type of near-duplicate images, the images in Fig.1a are similar to each other on the whole scale, whereas the images in Fig.1b are only near-duplicate in a small portion of the whole image, the STARBUCKS LOGO. Typically, the first one can be taken as the special case of the second one, which is also a challenge for NDII. In this paper, we will focus on the more general problem, partial NDII (PNDII).



(a)

(b)

Figure 1: Samples of near-duplicate images: (a) near-duplicate images of Mona Lisa; (b) partial near-duplicate images of STARBUCKS logo.

Global statistic representations, such as color histograms, shape and texture features, are known to be sensitive to real-world image variations, and often cannot give adequate descriptions of an image's local structures and discriminating features. Therefore, it is not suitable for NDII, especially for PNDII. Local features are often more reliably detected and matched across different examples of an object or scene with varying viewpoints, lighting conditions, and various geometric transformations. However, it is very complex and time-consuming to use the descriptors directly due to the large amount of local features of each image and the high dimensionality of descriptors. "Bag-of-Words"(BOW) model is an effective approach for improving the matching performance of local features. With this model, each local feature is assigned to one of the "visual word". Each image is then represented by a histogram of word frequency. However, the most critical problem of such method is that the spatial information among local features is totally neglected. Also visual words are less distinctive than local feature descriptors. Despite its simplicity and efficiency, the

ambiguous visual words will introduce large number of false matches when each descriptor is matched independent to others. The polysemy and synonymy of visual words degrade the accuracy of BOW model [3].

Several methods have been proposed to improve the problems mentioned above by capturing the geometric or spatial arrangement of visual words, such as using the relative location of matched keypoints for filtering mismatches [4], bundling the local features within the same MSER region for enhancement matching [5], verifying the global features of the region formed by the matched local features [6]. But these methods are also affected by the repeatability of local features. The polysemy and synonymy of visual words, which mean that the local features not matched with each other may be quantized to the same visual word, and the matched local features may be quantized to different visual words, make PNDII more challenging.

The spatial information is usually reintroduced as a post-processing step to re-rank only the retrieved images obtained by BOW model for reserving the efficiency. In this paper, we propose an image representation approach using visual word histogram with weak geometry information. We first obtain the potential matched images with the frequency histogram of visual words. And then in the second stage, the weak geometric information is used for verifying whether the match is correct or not, which is also the focus of this paper.

In this paper, we adopt Lowe's difference of Gaussian (DoG) detector and SIFT descriptor for obtaining local features [7]. This detector is scale and rotation invariant, and can tolerant certain amount of affine transformation. It is observed that the ratios of the scales of correct matched local features of two images are approximately equal to the variance of scale. And the differences of the orientations of the matched points are approximately equal to the rotation angle of the near-duplicate region. It shows that the ratio of scales and difference of orientations of correct matches are consistent. Typically, the local descriptors quantized to the same visual words are all matched to each other. Therefore, if we can obtain a subset from all the matches formed by the local features quantized to the same words, and the cardinality of the subset is no less then a threshold, the two images are near-duplicate, otherwise not. Our approach incorporates the global geometric consistency of the subset of matched local features. The experiments show that it is robust to the quantization error of the BOW model.

The remainder of this paper is organized as follows. Section 2 reviews some closely related works of near-duplicate image identification. Section 3 firstly describes the representation of image, and then introduces the principle of our approach by example. Section 4 presents the formal description of the proposed global geometric consistency of subset-of-features method. And section 5 shows the experimental results and related analysis. Finally, the conclusion and perspective are summarized in section 6.

II. Related Work

NDII is a challenging task. Despite that many solutions to the NDII problem have been proposed, by and large, contemporary solutions mainly focus on how to improve the accuracy and efficiency of BOW model using spatial and geometric relationship of local features. In this section, we review some closely related works.

Min-Hash [8] uses an approximate set intersection between visual word histograms to discovering near-duplicate images. However, it is ineffective for identifying image with small near-duplicate regions. Chum et al. [9] proposed Geometric minHash (GmH) improves upon standard min-Hash using the spatial extent of image features. Lee et al. [10] proposed Partition minHash (PmH). PmH first divides the image into different sizes of overlapping patches, and then matches the visual word in each patch with standard minHash independently. It shows that PmH outperforms min-Hash in terms of precision and recall. Zhou et al. [4] proposed to encode the spatial relationships among local features in an image to discover false matches of local features for enhancing the accuracy of visual word matching. Wu et al. [5] proposed to bundle SIFT local features within the MSER regions for weak geometric verification. These bundled features provide a flexible representation that allows simple and robust geometric constraints. However, the weak geometric constraint is only applicable when there is no significant rotation between duplicate images. Zhang et al. [2] proposed a part-based image similarity measure derived from stochastic matching of attribute relational graphs that represent the compositional parts and part relations of image scenes. But this method is with high computation cost. Xu et al. [11] proposed to divide the images into different parts with different sizes of patches. And then they find the most matching score with EMD and linear programming. This is the most formal matching approach of PNDII. But it is still high computation cost. Lv et al. [12] propose to segment the images and then use EMD* and compact data structure for efficient matching. Wang et al. [6] proposed to further confirm the matching of local features; the color histograms of areas formed by matched keypoints in two images are compared for verifying the matching result. Zhao et al. [13] proposed to further improve the accuracy by Scale Rotation invariant Pattern Entropy (SR-PE). SR-PE is a pattern evaluation technique capable of measuring the spatial regularity of matching patterns formed by local keypoints. But this method may filter out some near-duplicate image with small similar regions in the first stage. Additionally, Hu et al. [14] introduced the coherent phrase model which incorporates the coherency of local regions to reduce the quantization error of the BOW model. Local regions are characterized by visual phrase of multiple descriptors instead of visual of single descriptor. Therefore, it is more distinctive. Zhang et al. [15] proposed to encode more spatial information for

BOW representation through the geometry-preserving visual phrases. These methods can improve the distinctiveness of visual word effectively. However, the accuracy is affected by the repeatability of local features.

III. Problem Statement

In this section, we first introduce the representation of images with BOW model and weak global geometric information. Then we show the principle of our approach by example.

A. Representation of Images

Let D be a finite set of image database. We first extract the SIFT local features for each image I_i in D, and typically have the following information for i-th local feature p_i of I_i: a descriptor d_i, an orientation value o_i, a scale value σ_i and the coordinate (x_i,y_i) of the keypoint. $\{(x_i,y_i),\sigma_i,o_i\}$ is the geometric information of p_i. Then, we generate the visual words by clustering the sample features with k-means. Let $VD=\{e_i|1\le i\le m\}$ be the visual dictionary, where e_i is visual word, and m is the size of the dictionary. Next, each local feature is assigned to a visual word by finding the nearest descriptor of word in VD. And also we reserve the scale and orientation information of the corresponding local feature. Finally, the representation of p_i is shown as follow: $p_i\{d_i,(x_i,y_i),\sigma_i,o_i\}\rightarrow p_i\{e_i,\sigma_i,o_i\}$. Because we don't try to solve the transformation matrix of the corresponding matched local features. Therefore, we ignore the coordinate information, and use only the global scale and orientation information for geometric consistency verification. Fig.2 shows the representation of I_i with the structure of the histogram of the frequency of visual words and weak geometric information.

$$I_i = \begin{bmatrix} e_1 & \rightarrow & 0 & \rightarrow & \cdots \\ e_2 & \rightarrow & 2 & \rightarrow & \{\sigma_j,o_j\},\{\sigma_k,o_k\} \\ \cdots & \rightarrow & \cdots & \rightarrow & \cdots \\ e_i & \rightarrow & 3 & \rightarrow & \{\sigma_i,o_i\},\{\sigma_l,o_l\},\{\sigma_h,o_h\} \\ \cdots & \rightarrow & \cdots & \rightarrow & \cdots \\ e_m & \rightarrow & 0 & \rightarrow & \cdots \end{bmatrix}$$

Figure 2: Image representation with BOW model and weak geometric information.

Considering both the efficiency and effectiveness, the whole approach consists of two phases, finding the potential near-duplicate images and verifying the geometric consistency. The two stages of our approach use different information of our representation method. Given a query image Q, we first use the intersection distance τ to measure the similarity between two visual word frequency histograms H_{I_i} and H_Q of images I_i and Q, where $\tau(H_{I_i},H_Q)=\sum_{v=1}^{m} min(H_v(I_i),H_v(Q))$, $H_v(\cdot)$ represents the v-th bin of the histogram. And then, we can obtain the subset $I=\{I_i|\tau(H_{I_i},H_Q)\ge T_1 \ \&\ I_i\in D\}$, the parameter T_1 is set to gate whether a image is the potential near-duplicate image of Q or not. I_i is the potential near-duplicate image of Q.

Next, the weak geometric information is introduced in the post-processing step for geometric consistency verification. For enhancing matching efficiency, geometric consistency verification is applied only to the potential near-duplicate images in I obtained in the first phase.

B. Principle of Our Approach

Typically, the local features corresponding to the same visual words are matched with each other. Because of the mismatches of local features, and the polysemy and synonymy of visual words, different local features may be quantized to the same word, but they are not matched in fact. One local feature may be also matched with many keypoints in the other images. Therefore, geometric verification is necessary for enhancing the accuracy of BOW model. The weak geometric information in our representation is used for confirming that whether I_i is true near-duplicate image of Q or not. Our method doesn't try to find out the mismatches between the two images. And we only try to find out certain geometric consistency from the scale and orientation information between the matched features corresponding to the same visual words. The problem can be that given I_i and Q with our representations, the purpose is to confirm that whether I_i is truly near-duplicate to Q or not.

As we know, a SIFT keypoint is a circular image region with an orientation [7,16]. Fig.3 shows the geometric frame of the three parameters [17]: the keypoint center coordinate, its scale (the radius of the region), and its orientation (an angle expressed in radians). It is observed that the distributions of scales and orientations of the matched features are consistent when image is rotated or scaled. It means that the ratio of scales is approximately equal to the variance of scale, and the difference of orientations is approximately equal to the variance of rotation.

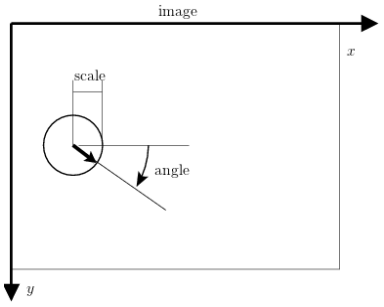


Figure 3: The frame of SIFT keypoint's geometric information.

As aforementioned, there must be similar regions between near-duplicate images. Therefore, we can use this geometric consistency to confirm that whether the match is correct or not. Fig.4 shows the keypoints found in two near-duplicate images of a boy. For illustration purposes, only 4 keypoints are shown in each image. The image on the right is a rotated, scaled version of the one on the left. The keypoints located in this pair of images are shown as blue circles, with lines denoting dominant orientations and radius denoting scale. Note that the changes of size and orientation of keypoints reflect how the image was scaled and rotated. (p_1,q_1),(p_2,q_2),(p_3,q_3),(p_4,q_4) are the pairs of matched SIFT feature points in these two images. It is observed that $\frac{\sigma_{p1}}{\sigma_{q1}} \approx \frac{\sigma_{p2}}{\sigma_{q2}} \approx \frac{\sigma_{p3}}{\sigma_{q3}} \approx \frac{\sigma_{p4}}{\sigma_{q4}}$, and o_{p_1}-$o_{q_1} \approx o_{p_2}$-$o_{q_2} \approx o_{p_3}$-$o_{q_3} \approx o_{p_4}$-o_{q_4}. Therefore, we can utilize this consistency among the matched keypoints. That is: if there is a subset of the matched features formed by all the local features corresponding to the same visual words, which meets the geometric consistency mentioned above, and the cardinality of the subset is greater than the threshold T_2, the two images are near-duplicate. Intuitively, if two images are near-duplicate, it is highly possible that there is a subset which meets the characteristics mentioned above, otherwise not. And the cardinality of the subset can also be used for measuring the similarity of the two near-duplicate images.

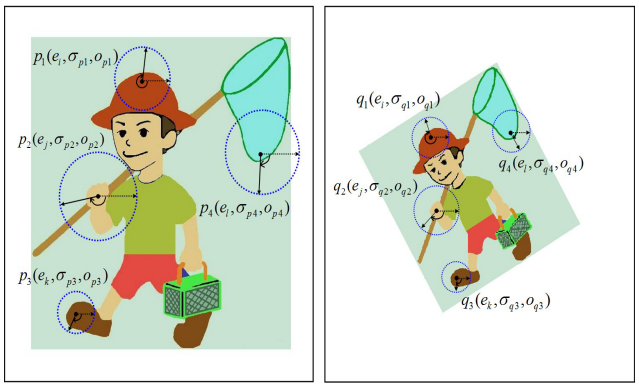


Figure 4: Geometric consistency of the scale ratio and orientation difference of the matched features.

IV. GGC-SOF: Global Geometric Consistency of Subset-of-features

Given two images I_i and Q, and the visual words of $I_i \cap Q$ are $M \subseteq VD$. PM are all the possible matched pairs of local features corresponding to all the same visual word $e_k \in M$. If there is a subset $S \subseteq PM$, the minimum radiuses of the clusters of the ratios of the scales and the differences of the orientations of the matched local feature pairs are less than thresholds C_σ and C_o respectively, and $|S| \geq T_2$, then, I_i is near-duplicate to Q. The formal description of GCC-SOF is given below.

$$F = |S|$$
$$s.t.$$
$$S = \{\{(e_k, \sigma_{Q_1}, o_{Q_1}), (e_k, \sigma_{Ii_1}, o_{Ii_1})\}, ...\} \subseteq PM$$
$$C(\frac{\sigma_{Ii_1}}{\sigma_{Q_1}}, ...) \leq C_\sigma$$
$$C(o_{Ii_1} - o_{Q_1}, ...) \leq C_o$$

where $C(\frac{\sigma_{Ii_1}}{\sigma_{Q_1}},...)$ stands for the minimum radius of the cluster of $(\frac{\sigma_{Ii_1}}{\sigma_{Q_1}},...)$, $C(\sigma_{Ii_1}$-$\sigma_{Q_1},...)$ has the similar meaning. C_σ and C_o are the cluster thresholds of the ratios of scales and the differences of orientations, respectively. If F=$|S| \geq T_2$, I_i is near-duplicate to Q. By adjusting C_σ, C_o and T_2, we can enhance the accuracy of matching and measure the similarity of the two images flexibly.

V. Experimental Evaluation And Analysis

To verify the effectiveness of the proposed approach, a testing benchmark with 2,000 images provided by Wu [6] is used. In this dataset, the images are consisted of 10 collections, of which 200 image pairs are identified as ground-truth near-duplicates. Additionally, we obtain the other 2,000 non-duplicate images from the Internet to construct the final testing dataset. In our experiments, the values of C_σ and C_o are 0.1 and 0.05, respectively. If the scale of the subset is bigger than 10, the two images are near-duplicate.

Our method is the post-processing stage of BOW model for PNDII. The experiments are conducted and compared with BOW model and our GGC-SOF. The size of visual vocabulary is 100,000. For the experiments, 100 images from the 10 collections of near-duplicate images (10 collections$\times$10 images in each) are randomly selected as the query images, and all the other images are the samples. Table 1 illustrates the average precision and average recall of the queries from the 10 collections.

Table I: The performance of our method compared with BOW model

Image Collections	Average Precision		Average Recall	
	BOW	GGC-SOF	BOW	GGC-SOF
American Flag	26.6%	85.8%	76.5%	48.5%
Beijing Olympic	29.1%	92.5%	71.7%	52.2%
Disney Logo	16.2%	84.8%	72.1%	33.3%
Google Logo	36.3%	87.5%	36.7%	24.2%
iPhone	57.4%	94.2%	84.9%	53.3%
KFC Logo	57.8%	88.1%	51.2%	35.3%
Mona Lisa Smile	27.7%	97.1%	78.9%	39.4%
Rockets Logo	13.9%	57.1%	49.7%	25.1%
Starbuck Logo	21.2%	97.8%	88.4%	78.2%
Exit Sign	23.9%	81.9%	71.6%	36.3%

According to Table 1, by our geometric consistency verification, the precision of all the retrieval results obtained by our approach is approximate 86%, which shows an obvious improvement than the BOW's average precision: 31%. It also shows that there are lots of mismatches caused by

the polysemy and synonymy of visual words. And lots of local features not matched with each other are assigned to the same visual words. Our method improves the accuracy effectively. However, for the collection of Rockets logo, the precision is low because of the less SIFT local features around the near-duplicate region. It also shows that the recall of our method is low. It is observed from the dataset that some near-duplicate images in the same collections change a lot. Additionally, we typically need to examine the trade-off between recall and precision. In this experiment, we set the parameters, such as C_σ and C_o, small for obtaining a high accuracy. Therefore, the recall of our method can also be increased by increasing the parameters properly.

Through our experiments, we also find that our method is effective for the challenging problem of partial near-duplicate image identification. We have tested some retrieval of query images with small near-duplicate regions to the results. Fig.6 shows the top ten retrieval results of some query images ranked by the maximum size of the global geometric consistency subset. It can be seen that our method is effective for partial near-duplicate images with small similar regions. Therefore, the proposed method can be taken as an as effective post-processing step to re-rank the retrieved images for enhancing the accuracy of BOW model.

Figure 5: Top ten retrieval results of some query images.

VI. Conclusion

In this paper, we have presented an effective method, global geometric consistency of subset-of-features, for partial near-duplicate image identification using the consistency of the ratios of scales and the differences of orientations of the matched local feature points. The preliminary experimental results show the effectiveness of our approach, especially for the identification of partial near-duplicate images. Our method is robust to the ambiguity of visual word, and it is more flexible by adjusting the parameters. However, for some query images with less SIFT keypoints in the near-duplicate regions, the accuracy of results is not high. In addition, it is time consuming for finding the global geometric consistency subset. We need to find the proper indexing structure for accelerating the searching process. These are also the further works to be improved.

Acknowledgment

This work is supported by the National Science Foundation of China (Project: 61171193). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] X. Yang, Q. Zhu, K. Cheng. MyFinder: near-duplicate detection for large image collections, Proc. ACM MM'9, 2009, pp.1013-1014.
[2] D. Zhang, S. Chang. Detecting image near-duplicate by stochastic attribute relational graph matching with learning, Proc. ACM MM'04, 2004.
[3] M. Duan, X. Wu. Visual polysemy and synonymy: toward near-duplicate image retrieval, Front. Electr. Electron. Eng. China, 2010, pp.419-429.
[4] W. Zhou, Y. Lu, H. Li, Y. Song, Q. Tian. Spatial coding for large scale partial-duplicate web image search, Proc. ACM MM'10, 2010, pp.511-520.
[5] Z. Wu, Q. Ke, M. Isard, J. Sun. Bundling features for large scale partial-duplicate web image search, Proc. CVPR, 2009, pp.25-32.
[6] Y. Wang, Z. Hou, K. Leman, N.T. Pham, T. Chua, R. Chang. Combination of local and global features for near-duplicate detection, Proc. MMM, 2011, pp.328-338.
[7] D.G. Lowe. Distinctive image features from scale-invariant keypoints, International J. of Computer Vision, 2004:2(60), pp.91-110.
[8] O. Chum, J. Philbin, A. Zisserman. Near duplicate image detection: min-Hash and tf-idf weighting, Proc. BMVC, 2008.
[9] O. Chum, M. Perdoch, J. Matas. Geometric min-Hashing: finding a (thick) needle in a haystack, Proc. CVPR, 2009, pp.17-24.
[10] D.C. Lee, Q. Ke, M. Isard. Partition min-hash for partial duplicate image discovery, Proc. ECCV, pp.648-662.
[11] D. Xu, T. Cham, S. Yan, S. Chang. Near duplicate image identification with spatially aligned pyramid matching, Proc. CVPR, 2008, pp.1-7.
[12] Q. Lv, M. Charikar, K. Li. Image similarity search with compact data structures, Proc. ACM CIKM'04, 2004, pp.208-217.
[13] W. Zhao, C. Ngo. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection, IEEE Trans. on Image Processing, 2009, pp.412-423.
[14] Y. Hu, X. Cheng, L. Chia, X. Xie, D. Rajan, A. Tan. Coherent phrase model for efficient image near-duplicate retrieval, IEEE Trans. on Multimedia, 2009, pp.1434-1445.
[15] Y. Zhang, Z. Jia, T. Chen. Image retrieval with geometry-preserving visual phrases, Proc. CVPR, 2011, pp.809-816.
[16] A. Vedaldi. An open implementation of the SIFT detector and descriptor, UCLA CSD Technical Report, 2007.
[17] http://www.vlfeat.org/api/sift_8h.html