9-2015

Near-Duplicate Image Retrieval Based on Contextual Descriptor

Jinliang Yao
Hangzhou Dianzi University

Bing Yang
Hangzhou Dianzi University

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Near-Duplicate Image Retrieval Based on Contextual Descriptor

Jinliang Yao, Bing Yang, and Qiuming Zhu*

***Abstract*—The state of the art of technology for near-duplicate image retrieval is mostly based on the Bag-of-Visual-Words model. However, visual words are easy to cause mismatches because of quantization errors of the local features the words represent. In order to improve the precision of visual words matching, contextual descriptors are designed to strengthen their discriminative power and measure the contextual similarity of visual words. This paper presents a new contextual descriptor that measures the contextual similarity of visual words to immediately discard the mismatches and reduce the count of candidate images. The new contextual descriptor encodes the relationships of dominant orientation and spatial position between the referential visual words and their context. Experimental results on benchmark Copydays dataset demonstrate its efficiency and effectiveness for near-duplicate image retrieval.**

***Index Terms*—Near-duplicate image retrieval, visual word, contextual descriptor, spatial constraint.**

I. Introduction

GIVEN a query image, our objective is to find its near-duplicate versions in a large scale image database. In this paper, the near-duplicate versions of the image are referred to as those images that are usually generated from the original image by certain ways of altering and editing, such as cropping, scaling, rotation, color changing, compression, text addition, framing, and other non-affine geometric transformations. One instance of near-duplicate images is shown in Fig. 1. Two images in Fig.1 come from one original image by adding text, scaling cropping, etc. We consider one image as near-duplicate image of the other.

In near-duplicate image retrieval systems, the state of the art scheme is based on the Bag-of-Visual-Words model [1]. In this scheme, local features are quantized to visual words. Inverted file indexing is then applied to register images via these visual words. However, visual words have much less discriminative power than text words due to the impact of quantization and image editing operations. The approaches of only indexing the images with the set of visual words suffer from lack of precision.

In order to improve the retrieval precision of the visual words approaches, geometric verification for rejecting mismatches of visual words has become very popular as a visual words post-verification step. Zheng [2] proposed a visual phraselet method based on the pairs of visual words to refine spatial constraints. Zhou [3] designed a spatial coding technique to discard mismatches of visual words. Wu [4] built bundled features that are detected by grouping local features within MSER regions. The similarity of the bundled features is measured by their spatial orders. However, these above methods need to obtain the matched pairs of visual words between a query image and a candidate image first, and then calculate the spatial similarity of the matched visual words between the two images for rejecting mismatches of visual words. Due to the expensive computational cost and large number of candidate images in large scale datasets, these rejecting mismatches processes are usually applied to only some top-ranked candidate images. This practice causes poor precision for near-duplicate image retrieval.

(a) (b)

Figure 1. An example of near-duplicate images.

In tackling the problems of visual words post-verification processes, one basic idea that has been explored is to design a local spatial descriptor which can be used to immediately filter the mismatches of visual words according to the similarity of local spatial descriptors. Liu [5] tried this idea, and proposed spatial contextual binary signatures for visual words. Liu's method firstly divides the surrounding local features into different parts and computes the weighted sum of these surrounding features. Then an orthogonal projection matrix is used to reduce the dimension of the feature vector. Finally, the reduced feature vector is quantized by using a threshold. This method does not pay enough attention to the impact of missing local features. It is vulnerable to some image editing operations, such as scaling. Different from Liu's method, Zheng [9] embedded the binary color feature of keypoint into the inverted index files to check for visual word matching.

In this paper, we focus on the impact of image editing operations and propose a contextual descriptor which enumerates the spatial information of local features in the context. This new descriptor is an improved version of our prior

Manuscript received Aug., 2014; revised Sep.2014; accepted February 16, 2013. Date of current version March 07, 2013. This work was supported by the National Natural Science Foundation of China under Grant 61202280.

J. Yao and B. Yang Authors, are with Computer Science School, Hangzhou Dianzi University, Hangzhou, 310018, China (e-mail: yaojinl@hdu.edu.cn).

Q. Zhu Author, is with Computer Science department, University of Nebraska at Omaha, NE, 68182, USA (e-mail: qzhu@mail.unomaha.edu)

work [10]. The new descriptor improves the compactness of the old version and is demonstrated in near-duplicate image retrieval. The proposed descriptor can tolerate missing a part of local features, increase the discriminative power of visual words and be embedded into an inverted file indexing structure. Different from Liu's method [5], our proposed descriptor encodes the spatial relations of the context by order relation which is robust to most of image editing operations. In addition, the dominant orientation of local feature is adopted to represent local feature because of its robustness. Experiments show that our proposed contextual descriptor approach achieved considerable improvements over the baseline approach and other visual words post-verification approaches.

II. CONSTRUCTING CONTEXTUAL DESCRIPTORS

All of local features in images are selected as the referential local features to construct their contextual descriptors. The proposed contextual descriptors can be constructed by the following three steps.

① Select interest points (IPs) from the neighbors of the referential local feature as the context which is a set of local features.

② Extract the contextual features between the referential local feature and its context.

③ Generate contextual descriptors by encoding the contextual features.

The detail processing is introduced as follows. We use SIFT as the descriptor of local features. A SIFT descriptor (S_i) is characterized by a feature vector (F_i), a dominant orientation (θ_i), a feature scale (σ_i), and a spatial position (Px_i, Py_i). That is, a SIFT descriptor can be denoted as [$F_i, \theta_i, \sigma_i, Px_i, Py_i$].

A. Selecting the context

Many image editing operations, such as scaling, compression, greatly affect the results of local feature detection for near-duplicate image retrieval. For example, Small scale SIFT descriptors disappear after a resolution reduction operation, when the image in Fig.2 (a) is transformed into Fig.2 (b) with 1/3 of resolution reduction. In practical implementation, we could only select some of the neighbors as the context due to the consideration of conserving storage space. Meanwhile, the context of the same referential local feature in different resolution image should include as many of the same neighbors as possible. If small scale local features are selected as the context of the large scale referential local features, the contextual descriptor becomes unstable because small scale local features easily disappear in low resolution images.

Selecting larger scale neighbors as the context can reduce the impact of image scaling transformation. Therefore, we select a fixed number (N) of local features as the context in terms of the weighted sum of the scale and distances differences between a referential local feature and its neighbors. The weighted sum (W_i) is computed with (1).

$$W_i = \frac{C*\sqrt{(Px_i - Px_o)^2 + (Py_i - Py_o)^2}}{\sqrt{Img\_W^2 + Img\_H^2}} + \frac{(1-C)*(\sigma_i - \sigma_o)}{\max(\sigma)} \quad (1)$$

$$Neighbors(o) = \underset{N}{Min}(W_i) \qquad i \in V \quad (2)$$

where C and (1-C) denote the weight for the difference of distance and scale, respectively. The subscripts 'o' and 'i' in the expression denote the referential local feature and the other local features in image, respectively. $\sqrt{Img\_W^2 + Img\_H^2}$ and $\max(\sigma)$ are used to normalize the distance and the scale difference. The selected neighbors are obtained by (2), where V denotes all local features in the image. We call the selected neighbors $\{S_i \mid \underset{N}{Min}(W_i), i \in V\}$ as the context of the referential local feature (o). N is used to set the size of context.

Fig. 2 shows an example of selecting the context in different resolution images. The red, yellow, and white lines denote the scale and dominant orientation of the referential local feature, its context, and the non-context neighbors, respectively. From Fig.2(a), we can find that some small scale local features are not chosen as the context. Therefore this way ensures the contexts in different resolutions have a higher possibility to have the same local features.

(a) (b)

Figure 2. The context of the same local feature in different resolution image; (a) is the original image; (b) is edited with resolution 1/3 of the original image.

B. Extracting the features of the context

The compactness and robustness of the relational features between the referential local feature and the local feature in its context are important to near-duplicate image retrieval. We found the dominant orientations of local features are more stable and compact than visual words obtained by quantization of local feature. Therefore we utilize dominant orientations to represent local features. In order to keep the robustness to scaling operation, the proposed descriptor only explores the directional relationship. As mentioned before, we compute the directional relationship ($\alpha(n)$) and the dominant orientation relationship ($\beta(n)$) between the referential local feature (l) and its context (n) by using (3) and (4) respectively.

$$\alpha(n) = |\arctan2(Py_n - Py_l, Px_n - Px_l) - \theta_l| \quad (3)$$

$$\beta(n) = |\theta_n - \theta_l| \quad (4)$$

Where arctan2(y, x) is an angle in radians between the positive x-axis and the line connecting the origin of the plane and the point given by the coordinates (x, y). In (3) and (4), subtracting θ_l is to keep the robustness to image rotation.

The computational processes for getting the values $\alpha(n)$ and $\beta(n)$ were shown in Fig.3. After this process, the context is represented as a set of contextual features $\{[\alpha(n), \beta(n), Px(n), Py(n)], \ n \in N\}$.

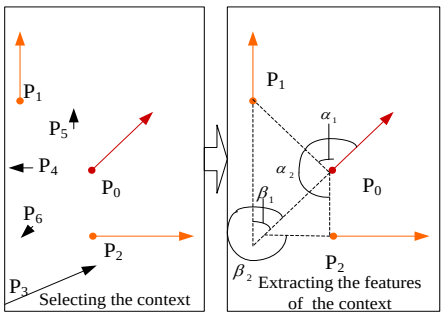

Figure 3. Selecting the context of local feature P_0 and extracting the contextual features of the context $\{P_1, P_2\}$.

C. Generating contextual descriptors

In order to obtain the compactness of the proposed contextual descriptor, a coding procedure for the contextual features $\{[\alpha(n), \beta(n)]\}$ is executed. The $\alpha(n)$ and $\beta(n)$ are quantized into a value q(n) in the range of 0 to 255 which can be represented as a byte in terms of (5).

$$q(n) = \left\lfloor \frac{\alpha(n)}{A} \right\rfloor * 2^4 + \left\lfloor \frac{\beta(n)}{B} \right\rfloor \quad (5)$$

where the multiplier 2^4 is adopted as an operator of shifting 4 bits. Therefore, the front 4 bits of q(n) are used to save the quantization result of $\alpha(n)$; and the last 4 bits of q(n) are used to save the quantization result of $\beta(n)$. In (5), A and B are two quantization factors.

After a local feature in the context is represented by a byte, the context is organized as a sorted array [q(1),q(2),…,q(n),…,q(N)] by the distance between them and the referential local feature. The subscript of q(n) in the array represents the order relation. The quantization result of the nearest local feature in the context is saved in the first position of the contextual descriptor array; and the furthest one is saved in the last position. In near-duplicate images, true matches of the context preserve the order relation of local features. Fig. 4 shows two contextual descriptors of the same referential local feature in two near-duplicate images.

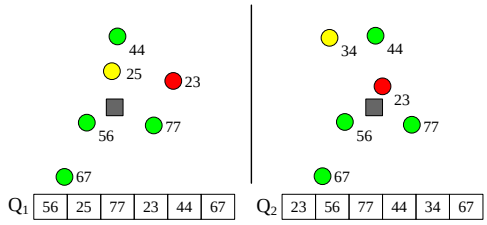

Figure 4. Two contexts and their contextual descriptors. The gray squares are the referential local features. The circles are their context. The red, yellow, and green denote the mismatches, missing neighbors, and true matches, respectively. The values are the quantization results of the context.

D. Matching contextual descriptors

In the retrieval stage, the maximal length of the ordered matching lists between two contextual descriptors is used to measure their similarity. The ordered matching list refers to the matched items which positions in the two descriptor arrays keep the same order. The calculating process takes place as follows.

① Obtain the matched items between two contextual descriptors.

② Get the positions of the matched items in their own contextual descriptor arrays, respectively.

③ Enumerate possible position lists where the items keep the same order.

We take Fig. 4 as an example to show how to get the ordered matching lists. According to the contextual descriptors in Fig.4, the matched items {56,77,23,44,67} are obtained. The positions of the matched items in the two descriptor arrays are [0,2,3,4,5] and [1,2,0,3,5], respectively. The position list of Q_1 is an ascending order, we need to choose some positions from Q_2 and keep them an ascending order. [1,2,3,5] and [0,3,5] are possible position lists. The position list with the maximal length is [1,2,3,5] which corresponds the ordered matching list [56,77,44,67]. The ordered matching lists preserve the distance order relation of the context. A threshold (Ts) for the maximal length of the ordered matching lists is used to verify if it is a true match.

III. Experiments

To evaluate the effectiveness of our proposed contextual descriptor for near-duplicate image retrieval, we conducted experiments on the Copydays dataset [6] which is exclusively composed of personal holiday's photos. Each image has suffered three kinds of editing operations: JPEG compression, cropping and "strong." The motivation is to evaluate the behavior of the indexing algorithms for most common image copies. This dataset has 157 original images. Each original image has 19 corresponding near-duplicate images. Because the size of the Copydays dataset was relatively small for algorithm testing purpose, the methods were evaluated in a large scale image dataset by adding distracter images. In our experiments, Flickr 1M image dataset [7] which was retrieved from Flickr was used as distracter images. To evaluate the performance with respect to the size of dataset, some smaller datasets (100K, 200K, etc) were built by sampling the Flickr 1M dataset. Mean Average Precision (mAP) [8] was used to measure image retrieval accuracy.

Our experiments focused on the effectiveness of the contextual descriptors, rather than on how to get visual words. Product quantization method [9] was used to transform local feature into visual words which had high transformation efficiency. In the experiments, the size of codebook is set 2^{21}. Therefore for all the experiments, visual words were obtained from SIFT by product quantization method.

An inverted-file index structure was used for our proposed near-duplicate image retrieval method. Each visual word has an entry in the index that contains the list of image ID and the contextual descriptor. In retrieval stage, the similarity of the contextual descriptor is used to verify if the matched visual word is a true match. We sorted the candidate images by the count of the matching visual words which were verified by their contextual descriptors.

A. Impact of parameters

The proposed contextual descriptor was evaluated against

different context sizes (N) and similarity thresholds (Ts). These two parameters are related to each other. So we use a table to show the results in different parameters on the 100K distracter image dataset. The performance of mAP with different context sizes and similarity thresholds is shown in Table. 1.

TABLE 1
THE MAP RESULTS WITH DIFFERENT CONTEXT SIZE AND SIMILARITY THRESHOLD

Ts \ N	6	8	12	16
3	0.868	0.809	0.639	0.505
4	0.870	**0.871**	0.787	0.656
5	0.817	0.837	0.795	0.686
6	0.666	0.802	0.784	0.685

When N and Ts were set 8 and 4, respectively, our method obtains the highest mAP. With the increase of N, the mismatch probability of items in the contextual descriptors also increases, so mAP is decreasing. However, with the decrease of N, the items in the contextual descriptors are easily missed because of image editing operations, so mAP is decreasing too. Ts is co-related with N. When N is bigger; Ts need be set to a bigger value.

B. Evaluation

We experimented with three methods: the baseline method, the visual words post-verification method, and the embedding method for comparison with our contextual descriptors. This baseline method sorts the candidate images by the count of the matching visual words without visual words post-verification and contextual descriptor verification. The chosen visual words post-verification is a spatial coding method [3] which is denoted as "Rerank." In our implementation, the parameters r and the threshold for checking the value of S were set to 2 and 0.7, respectively. The embedding method is another contextual descriptors method [5] which is denoted as "Embedding." In our implementation, the Hamming distance threshold was set to 4. Fig. 5 shows the results of different methods.

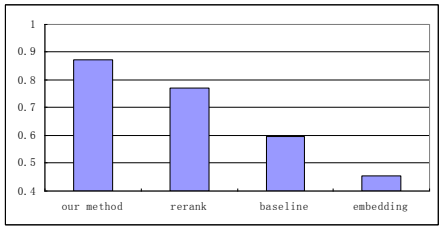

Figure 5. Comparison of mAP for different methods on the 100k database.

From Fig.5, it can be observed that our approach outperformed the other three methods. The mAP of the baseline method was 0.598. Our approach increased it to 0.871. Since the embedding method's contextual descriptor [5] was sensitive to the missing of its neighbors, its performance was lower than our approach.

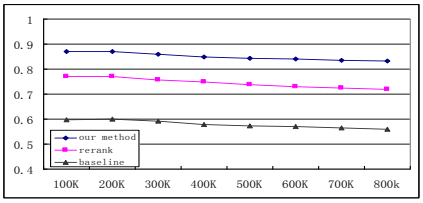

Figure 6. The change of mAP on different size databases.

From Fig. 6, it can be seen that the mAP of these methods decrease with the increase of database size. The rate of change of our method was lowest which changed from 0.871 to 0.832 among these methods.

TABLE 2
THE STORAGE PER VISUAL WORD IN INVERTED FILE AND AVERAGE QUERY TIME COST FOR DIFFERENT METHODS ON 100K DATASET.

	Query time (S)	Storage (bytes)
Our method	13	12
Rerank	18	8
Embedding	28	12
baseline	4	4

Tab.2 shows our method needs more storage to save the contextual descriptor of visual word. But its average query time is lower than the "Rerank" method. The query time does not include the time cost of obtaining the SIFTs from images.

IV. CONCLUSION

In this paper, we described a new contextual descriptor which improves the discrimination power of visual words. The proposed contextual descriptor efficiently encodes the neighbors' local descriptor and relative spatial relation and effectively discovers false matches of visual words between images. As for near-duplicate image retrieval, our contextual descriptor achieves better performance than some visual words post-verification methods and consumes much less query time.

The proposed contextual descriptor strictly encodes the spatial relations. It is robust to image editing operators, such as rotation, scaling, and cropping. However, the descriptor is not robust to perspective transformation of image. As demonstrated in the experiments, our approach is very effective and efficient for large scale near-duplicate image retrieval, but it does not work as well on general object retrieval.

REFERENCES

[1] P. James, O. Chum, M. Isard, J. Sivic, and A. Zisserman. "Object retrieval with large vocabularies and fast spatial matching," In Proc. IEEE CVPR, 2007, pp. 1-8.

[2] L. Zheng and S. Wang, "Visual Phraselet: Refining Spatial Constraints for Large Scale Image Search," IEEE SIGNAL PROCESSING LETTERS, vol. 20, no. 4, pp.391-394, APRIL 2013.

[3] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian. "Spatial coding for large scale partial-duplicate web image search," In Proc. of ACM MM, New York, NY, USA, 2010, pp.511-520.

[4] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," IEEE CVPR, 2009, pp25-32.

[5] Z. Liu, H. Li, W. Zhou, and Q. Tian. "Embedding spatial context information into inverted file for large-scale image retrieval," In Proc. ACM MM, Nara, Japan, 2012, pp. 199-208.

[6] H. Jegou, M. Douze and C. Schmid, "Hamming Embedding and Weak geometry consistency for large scale image search," ECCV, 2008, pp.304-317.

[7] M. J. Huiskes, B. Thomee, M. S. Lew. "New Trends and Ideas in Visual Concept Detection: the MIR flickr retrieval evaluation initiative," In Proc. ACM MIR, Philadelphia, USA, 2010, pp.527-536.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman. "Object retrieval with large vocabularies and fast spatial matching," CVPR, 2007, pp.1-8.

[9] H. Jegou; M. Douze, C. Schmid, "Product Quantization for Nearest Neighbor Search," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.33, no.1, pp.117-128, Jan. 2011.

[10] J. Yao, B. Yang, Q. Zhu, "Rejecting Mismatches of Visual Words by Context Descriptors," presented at the 13th ICARCV, Singapore, 2014.

[11] L. Zheng, S. Wang, Q. Tian, "Coupled Binary Embedding for Large-Scale Image Retrieval," IEEE Trans Image Process., Vol.23,no.8,pp.3368-80,2014.