MULTI-DESCRIPTION OF LOCAL INTEREST POINT FOR PARTIAL-DUPLICATE IMAGE RETRIEVAL

Liang Li[1,2], *Shuqiang Jiang*[1], *Qingming Huang*[1,2]

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

ABSTRACT

In partial-duplicate image retrieval, images are commonly represented using Bag-of-visual-Words (BoW) built from image local features, such as SIFT. Therefore, the discriminative power of the local features is closely related with the BoW image representation and its performance in different applications. In this paper, we first propose a rotation-invariant Local Self-Similarity Descriptor (LSSD), which captures the internal geometric layouts in the local textural self-similar regions around interest points. Then we combine LSSD with SIFT to develop a multi-description of images for retrieving partial-duplicate. Finally, we formulate the Semi-Relative Entropy as the distance metric. Retrieval performance of this multi-description evaluated in the Oxford building dataset and an image corpus crawled from Google shows that the average precision achieves 11.1% and 2.8% improvement, respectively, comparing with state-of-the-art bundling feature.

Index Terms— Partial-duplicate, LSSD, Semi-Relative Entropy

1. INTRODUCTION

As image manipulation software becomes more powerful, more pirated images are published on the web. A very insidious form of image manipulation is the creation of fake photographs by cutting and pasting pieces extracted from different original sources. For example, one could crop figures of two celebrities from two different photographs and create a fake composite image showing them embracing, even if they may never have met in reality [1]. The problem of matching a small portion of one image to another related image is termed *partial-duplicate image retrieval* [3]. It is an important approach of finding copyrighted images [1] (potentially modified), detecting forged images [2] and image retrieval [3, 4,19].

Bag-of-visual-Words (BoW) is commonly utilized to represent images and has been adopted for similarity measure in many applications such as object recognition [6], texture recognition [7], image retrieval [3] and image re-ranking [5]. In BoW image representation, the local features are first extracted and then replaced by their nearest visual words, which are generated by clustering a large number of local features. Hence, extracting the local features for interest points is a critical step. However, as illustrated in [3, 5, 6, 12], in BoW representation the commonly used SIFT descriptor is not descriptive enough and is prone to noise. This motivates some improvement works to strengthen the descriptive power of the BoW representation. For instance, in [3], bundling feature is proposed to capture the spatial contexts between local features by bundling the SIFT [9] and MSER [10] into local groups. In [6], the authors select the most discriminative visual word combinations with Adaboost for object recognition. In [5], Descriptive Visual Words and Descriptive Visual Phrases are proposed as the visual correspondences to text words and phrases. In [11], Shechtman et al. propose an approach for measuring similarity based on matching local self-similarity of color. In [12], the authors present a novel method for recognizing object categories when using multiple cues. However, most of the approaches are to model the spatial relation of visual words, which lost much useful information as a result of quantification. Such as bundling feature, it does not work well when images are rotated or flipped. Intuitively, the feature with a good description for both local

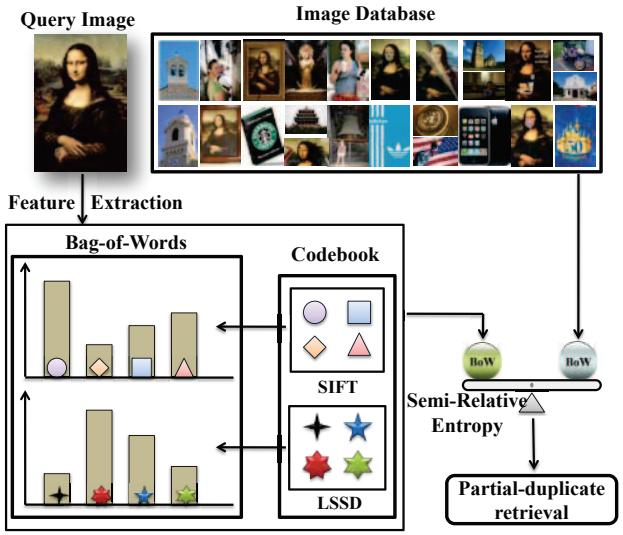

Fig. 1. The proposed framework of the system

interest points and spatial relation could be used as an effective representation for partial-duplicate image retrieval.

Based on the above analysis, we propose a novel rotation-invariant Local Self-Similarity Descriptor (LSSD). As a descriptor of spatial information, LSSD captures the internal geometric layouts in the local textural self-similar regions around interest points. The details of LSSD will be presented in section 2. As a local descriptor, SIFT has the ability of robust and effective representation for an interest point. For better discriminative power, we combine LSSD and SIFT to form a multi-description for a local interest point. Fig. 1 illustrates the framework of our algorithm. At first, LSSD and SIFT features are extracted. Then we represent images with two bags of visual words of SIFT and LSSD. Finally, according to the unique characteristic of partial-duplicate image retrieval, we propose the Semi-Relative Entropy (SRE) as the distance metric between two images based on the multi-description. Experiments on partial-duplicate image retrieval show that it is more reasonable than other distance metric such as Euclidean distance or χ^2 distance.

In summary, there are three contributions of our work: 1) the proposed LSSD presents good descriptive ability for geometric layout and is robust for small local distortions; 2) the multi-description for local interest points is discriminative and suitable for partial-duplicate image retrieval; 3) our proposed Semi-Relative Entropy is proved to be reasonable and shows promising performance for partial-duplicate image retrieval.

The remainder of the paper is organized as follows. Section 2 introduces the proposed Local Self-Similarity Descriptor in detail. Section 3 derives the proposed similarity measure formulation. The results of experiments will be discussed in section 4. Finally, section 5 concludes the paper.

2. THE LOCAL TEXTURAL SELF-SIMILARITY DESCRIPTION

In this paper we present a Local Self-Similarity Descriptor (LSSD) which captures internal geometric layouts of local textural self-similarities around interest points and is robust to small local distortions. Figure 2 illustrates the process of extracting LSSD.

2.1. Distinctive interest point

The LSSD is built upon interest points. Interest points are commonly employed in a number of real-world applications such as object recognition [5, 6] and image retrieval [3, 4], because they can be computed efficiently. In addition, they are resistant to partial occlusion and relatively insensitive to changes in viewpoint. We detect image interest points using the DoG (Difference of Gaussian), proposed by D. G. Lowe *et al.* [9], which has been shown to be robust and effective for detecting interest points. From the detected interest points, we can obtain three cues for each interest point *i.e.*1) scale; 2) location factor; 3) orientation factor.

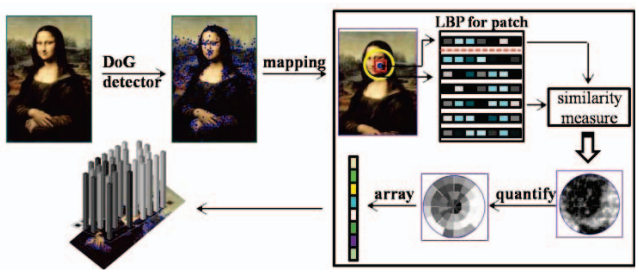

Fig. 2. Extracting Local Self-Similarity Descriptor

2.2. The extraction of LSSD

The LSSD is extracted in a circular region centered at an interest point (*i.e.* typically of radius 20 pixels). At first, for each point in the circular region, we compute LBP histogram [7] for its surrounding image patch (typically 7×7). Then we construct the self-similarity map, where the value of each point is the similarity distance between this point and central point. We use the χ^2 distance to measure the distance. Since the LBP only contains the information of texture, as supplementary, we add a regular term to measure the bias of luminance. Finally, the final distance measure formulation is:

$$SS_k(p,q)=\sum_i \frac{\left(H_p(i)-H_q(i)\right)^2}{\left(H_p(i)+H_q(i)\right)}+\tau\times\sin\left(\frac{\pi}{2}\cdot\frac{\left|G_p-G_q\right|}{255}\right) \qquad (1)$$

where, G_p denotes the gray value of point p and H_p stands for the LBP histogram of the patch centered at p; τ is set to 0.2 empirically.

After we obtain the self-similarity map, we transform it into a polar coordinate, which is partitioned into 64 bins symmetrically (4 radial intervals, 16 angle intervals). As a tradeoff between robustness and discrimination, the second maximal value in each bin is used as the value of each bin in the descriptor. Finally, we cascade these 64 bins clockwise and radially to form the LSSD. The value of the vector is normalized to the range 0-1. In Sec 2.1, we obtain the orientation factor of an interest point, and cascade these bins with the orientation factor as the initial direction to keep rotation invariance of LSSD. Algorithm 1 presents the detailed computation of LSSD.

We note that not all descriptors are significant. In the

Algorithm1: extract the Local Self-Similarity Descriptor (LSSD)

Input: image I, interest points k_i, $i=1,\ldots\ldots,K$
Output: the set of LSSD $\{S_i\}$
1: for each interest point k_i **do**
2: Get a circular region centered at k_i with radius R
3: Get LBP histogram H' of a square centered at k_i
4: **for** each point $p_{i,j}$ in the circular region **do**
5: Compute LBP histogram $H_{i,j}$ of a square centered at $p_{i,j}$
6: Get the similarity by comparing H' with $H_{i,j}$
7: **end for**
8: Get a circular self-similarity map
9: Transform the map into a polar coordinate
10: Quantize the polar coordinate into S_i with bins
11:end for

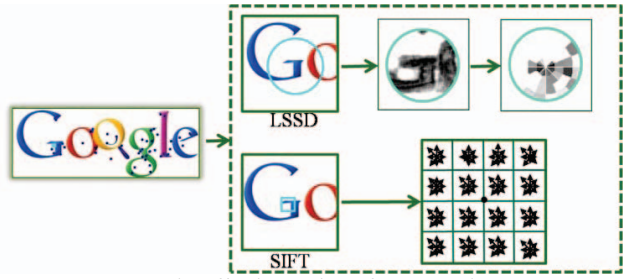

Fig. 3. The effective region of LSSD and SIFT

paper, we filter out two uninformative descriptors: 1) all the bins of a descriptor that are below a threshold (typically 0.05), which demonstrates that this descriptor does not capture any useful information. 2) all the bins of a descriptor that are above a threshold (typically 0.95), which demonstrates that this interest point is an outlier.

2.3. The multi-description of LSSD and SIFT

In our work we quantize LSSD and SIFT into the BoW representation independently and bind both histogram representations into a multi-description representing an image. Fig. 3 illustrates the effective region of LSSD and SIFT. LSSD has the capability of capturing the internal geometric layout for a larger region centered at an interest point and SIFT is a robust and effective representation of content for an interest point. Such multi-description has strong descriptive power including geometry and content in the local region so that it is valid to occlusion.

3. SIMILARITY MEASURE WITH SEMI-RELATIVE ENTROPY

We quantize LSSD and SIFT into visual words separately. For each image, we can represent it with our proposed multi-description. We evaluate the similarity between images by formulating a Semi-Relative Entropy (SRE).

In probability theory and information theory, the relative entropy [13] (also KL-divergence, information divergence, or information gain) is a non-symmetric measure of the difference between two probability distributions P and Q:

$$D_{KL}(P \| Q) = \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{2}$$

Besides non-symmetric, the relative entropy has another disadvantage: some items are neutralized because function $log(\cdot)$ is not always larger than zero. So we propose the Semi-Relative Entropy $D_{SRE}(P \| Q)$, which is given by:

$$D_{SRE}(P \| Q) = \sum f\left(\left| p(i) \cdot \log \frac{p(i)}{q(i)} \right| + \left| q(i) \cdot \log \frac{q(i)}{p(i)} \right| \right) \tag{3}$$

where, $f(\cdot)$ is a monotonically increasing function and is defined as the sigmoid function:

$$f(s) = 1/(1+\exp(-s)) \tag{4}$$

By means of the function, we can balance the effect of each item.

Finally, we formulate the SRE to evaluate the similarity between images:

$$SRE(Q \| S) = \alpha \cdot D_{SRE}(Q_p \| S_p) + \beta \cdot D_{SRE}(Q_q \| S_q) + r \cdot D_{SRE}(Q_c \| S_c) \tag{5}$$

$$s.t. \quad \alpha + \beta = 1, r = 2 \cdot \sqrt{\alpha \cdot \beta}$$

where, $\alpha = 0.4$, based on the performance on a separate validation dataset. Q_p, Q_q are the probability density distribution of different BoW for query image Q and S_p, S_q are the probability density distribution of different BoW for Source images. Q_c, S_c are the cross-modality probability density distribution , which is given by:

$$Q_c(i,j) = p_i \times q_j \quad p_i \in Q_p, q_j \in Q_q \tag{6}$$

In Equation (5), $D_{SRE}(Q_p \| S_p)$ and $D_{SRE}(Q_q \| S_q)$ denote the SRE between homogeneous visual words, while $D_{SRE}(Q_c \| S_c)$ denotes the SRE between joint heterogeneous visual words.

4. EXPERIMENTAL RESULTS

4.1. Experiment datasets

We use two datasets, *the Oxford building dataset* [15], and *an image corpus crawled from Google image* [16], for partial-duplicate image retrieval. For the Oxford building dataset, we select the images with "good" and "ok" label in the ground truth and there are 566 images with 11 different landmarks in all; for Google image corpus, we crawl 1375 images with 16 categories from Google image; Moreover, we choose 20 categories with 1799 images from *the Caltech 256 dataset* [17] as a near-partial-duplicate image retrieval dataset. Finally, we choose 6k negative images.

Following [3], we use Mean Average Precision (MAP) as our evaluation metric. 3 query images are selected for each category and then for each query image we compute the precision-recall curve, from which we obtain the average precision. At last, we take the mean value over all queries.

4.2. Comparison of discrimination for descriptor

We compare the discrimination of four descriptions in partial-duplicate image retrieval: the multi-description of SIFT and LSSD, another multi-description of SIFT and COLOR, SIFT, and LSSD. Images are represented with BoW of these four descriptions and SRE is used as the distance metric.

In the experiment, the COLOR description is a 128-Dimension gray histogram for a square patch (typically 21×21pixels) centered at the interest point detected by DoG detector [9]. In LSSD, the parameter α in Equation (5) takes 0.4 and the texture is represented by $LBP_{8,1}^{riu2}$ [7], which can be fast computed. Although LSSD is robust to small scale variance, we use the original query image and its down-sampled image to retrieve images for higher precision. The visual vocabulary of SIFT is 1.2 k and that of LSSD and COLOR is both 300.

Fig. 4-(a) shows the comparison of the above four descriptions. As expected, two multi-descriptions show the better performance, especially that the combination of LSSD

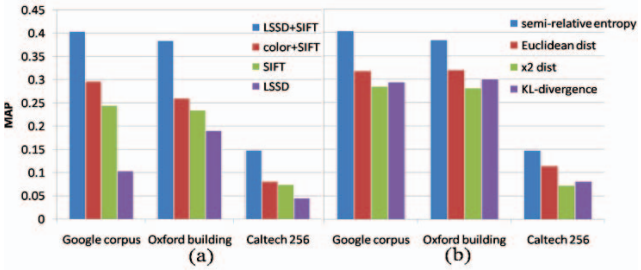

Fig. 4. Comparison of different descriptions and different distance metrics

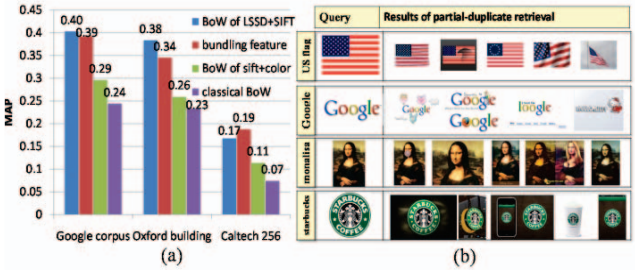

Fig. 5. Performance and examples of partial-duplicate image retrieval

and SIFT has a 13.4% improvement over the SIFT. Another result is that LSSD by itself has less descriptive ability than SIFT, which results from that LSSD only captures internal geometric layouts so as to have less uniqueness.

4.3. Comparison of distance metrics

We measure the similarity of the multi-description of LSSD and SIFT with four distance metrics: Euclidean distance, χ^2 distance, KL-divergence and our proposed SRE. Fig. 4-(b) shows that the SRE has the best performance. SRE decreases the weight of the common items between the query image and the candidate image while it increases the weight of the different items.

4.4. Performance for partial-duplicate image retrieval

We evaluate the performance of four main methods in these databases: classical BoW, BoW of the multi-description for COLOR and SIFT, BoW of our proposed multi-description for LSSD and SIFT, and "bundling feature" [3].

Fig. 5-(a) shows the results of above approaches with four observations: 1) the multi-description of LSSD and SIFT is effective. MAP of our method reaches 0.38 and 0.4 in *the oxford building dataset* and *the image corpus crawled from Google,* which shows a 15.02% and 15.91% improvement than classic BoW; 2) our approach outperforms "bundling feature" with 11.1% and 2.8% in these two image datasets; 3) the BoW of SIFT+ COLOR is not as good as SIFT+LSSD, because color information is variable after images are manipulated or pirated; 4) our approach has low MAP in *the Caltech 256* because the LSSD does not work where the textures of similar parts differ greatly. Fig. 5-(b) shows some retrieval instances from *the Google corpus*.

5. CONCLUSION

In this paper, a novel local self-similarity descriptor (LSSD) is proposed and it can capture internal geometric layouts in the local region around interest points. Then, a multi-description of local interest point is developed based on LSSD and SIFT for partial-duplicate image retrieval and promising results are achieved. In the future, we plan to advance the LSSD and apply it into object recognition.

6. ACKNOWLEDGEMENTS

This work was supported in part by National Natural Science Foundation of China: 60833006 and 60702035, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042.

7. REFERENCES

[1] J. Fridrich, D. Soukal, and J. Lukas. Detection of copy-move forgery in digital images. *Digital Forensic Research Workshop*, 2003.

[2] S. Berrani, L. Amsaleg, and P. Gros. Robust content-based image searches for copyright protection. *Proc. ACM Workshop*, 2003.

[3] Z. Wu, Q. F. Ke, and J. Sun. Bundling features for large- scale partial-duplicate web image search. *Proc. CVPR,* 2009.

[4] Y. Ke, R. Suthankar, and L. Huston. Efficient Near-Duplicate Detection and Sub-Image Retrieval. *Proc. ACM Multimedia,* 2004.

[5] S Zhang, Q Tian, G Hua, Q Huang, S Li. Descriptive Visual Words and Visual Phrases for Image Applications. *Proc. ACM Multimedia*, pp. 229-238, 2009.

[6] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. *Proc. CVPR*, 2008.

[7] Timo Ojala, Matti Pietikäinen, Topi Mäenpää. Multi-resolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *PAMI, 2002.*

[8] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc.CVPR*, 2006.

[9] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2(60):91–110, 2004.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *BMVC*, pp.384–393, 2002.

[11] Eli Shechtman, Michal Irani. Matching Local Self-Similarities across Images and Videos. *Proc.CVPR,* 2007.

[12] F. Khan, J. Weijer, M.Vanrell. Top-Down Color Attention for Object Recognition. *ICCV*, 2009.

[13] Kullback,s. The Kullback-Leibler distance. The American Statistician 41 (4): 340–341. 1987

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

[15] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/index.html

[16] http://images.google.com/

[17] http://www.vision.caltech.edu/Image_Datasets/Caltech256/

[18] M. Marszalek and C. Schmid. Spatial weighting for bag-of-features. Proc. CVPR, pp. 2118-2125, 2006.

[19] Z. Wu, Q. Xu, L. Li, P. Cui, S. Jiang and Q. Huang. Adding Affine Invariant Geometric Constraint for Partial-duplicate Image Retrieval, Proc. ICPR, 2010.