Performance Enhancement of Patch-based Descriptors for Image Copy Detection

Junaid Baber, Maheen Bakhtyar, Waheed Noor, Abdul Basit, Ihsan Ullah
Department of Computer Science and Information Technology
University of Balochistan

***Abstract*—Images have become main sources for the information, learning, and entertainment, but due to the advancement and progress in multimedia technologies, millions of images are shared on Internet daily which can be easily duplicated and redistributed. Distribution of these duplicated and transformed images cause a lot of problems and challenges such as piracy, redundancy, and content-based image indexing and retrieval. To address these problems, copy detection system based on local features are widely used. Initially, keypoints are detected and represented by some robust descriptors. The descriptors are computed over the affine patches around the keypoints, these patches should be repeatable under photometric and geometric transformations. However, there exist two main challenges with patch based descriptors, (1) the affine patch over the keypoint can produce similar descriptors under entirely different scene or the context which causes "ambiguity", and (2) the descriptors are not enough "distinctive" under image noise. Due to these limitations, the copy detection systems suffer in performance. We present a framework that makes descriptor more distinguishable and robust by influencing them with the texture and gradients in vicinity. The experimental evaluation on keypoints matching and image copy detection under severe transformations shows the effectiveness of the proposed framework.**

***Keywords*—Content-based image copy detection, SIFT, CSLBP, robust descriptors, patch based descriptors**

I. Introduction

With the availability of Internet and powerful open access image editing tools, it has become so easy and convenient to edit the images and distribute to others. Millions of images are shared and uploaded daily on image sharing sites such as Facebook, Flicker, and ImageShake. Image databases are increasing exponentially in size which causes so many problems for efficient image indexing and retrieval such as image piracy and redundancy. Much of work has been reported to address these problems. However, partial image copy detection and near duplicate detection are still very challenging [1], [2].

Generally, Image retrieval applications can be categorized into three types: near duplicate image detection, image copy detection, and similar image detection. In near duplicate detection, the task is to detect all image copies along with many other challenges such as the same scene captured by different viewpoint or captured on the different time. Image copy detection and partial duplicate detection are interchangeably used [1] and this is the subset of near duplicate image detection — original image is altered with changes in color, scale, partial occlusion, rotation, etc. Finally, similar image detection tends to find those images which are similar based on their visual contents, texture, or attributes. Similar image detection applications are widely used by many image search engines such as Google image. In these applications global features such as color histograms, variance, and image entropy are mostly used. In this paper, we mainly focus on content-based image copy detection which comprises of image copy detection and near duplicate image detection, these applications are potentially needed for piracy, copyright violation, and efficient image searching.

There are two famous techniques to prevent copyright violation; Watermarking and Content Based Copy Detection (CBCD), respectively. Watermarking technique involves embedding the information in the digital signal in such a way that it is difficult to remove [3] and carried along with the signal whenever distributed or shared. The information in watermark can be visible or invisible, whereas invisible information is widely used. There are many schemes proposed for digital watermarking such as spectrum watermarks [4], quantization watermarks [5], and blind detection watermarks [6]. Watermarking is the process of the identification of codes that store the information of the owner. The main challenges for watermarks are the translation, scale change, rotation, and cropping. For watermarking systems, signatures are extracted before distribution which makes this technique not feasible for on-line applications where digital signals are already distributed. Whereas, CBCD is the complementary approach to watermarking. In CBCD systems, robust signatures or fingerprints from query images are extracted, and then compared to the signatures or fingerprints of gallery images (database) to determine whether the query image has the copy in given gallery or not. One of the main advantages of CBCD system over watermarking is that signature extraction is not required before distribution, signature or fingerprint can be called the feature vector of the digital content. The feature vector should be distinguishable, robust and efficient [7]. The main challenge for CBCD systems is that the copy of digital content can be modified or transformed to deceive the copy detection systems. The list of top 10 challenge transformations are shown in Table I.

Local keypoint descriptors are widely used to represent the images in CBCD. Firstly, keypoints are detected from the images, these keypoints should have high repeatability under various affine and geometrical transformations. Some of the most successful algorithms for keypoint detection include SIFT [8], SURF [9], and Harris and Hessian affine

TABLE I
LIST OF CHALLENGES FOR COPY DETECTION

#	Description
C1	Camcording
C2	Picture in picture
C3	Insertions of pattern
C4	JPEG compression
C5	Change of illumination
C6	Cropping
C7	Blurring
C8	Image flipping
C9	Text insertion
C10	Decrease in image quality – This includes combination of all nine challenges along with image shifting, contrast, and image morphing

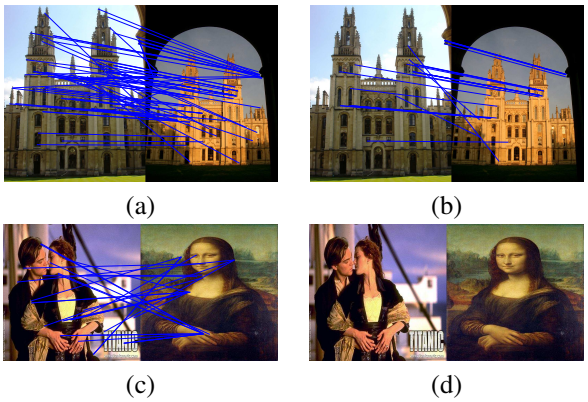


Fig. 1. SIFT descriptor matching, (a) and (c) SIFT descriptor, (b) and (d) neighbor base SIFT descriptor with two neighbors. Where (a) shows the indiscriminateness of SIFT descriptors and (c) shows the ambiguity.

keypoints [10]. Secondly, local patches around the keypoint are estimated [10]. Thirdly, robust and discriminative descriptors which are invariant to many transformations such as scale, rotation, affine distortion, 3D viewpoint change, image noise, and illumination change are computed [8]. Finally, image is represented by the set of local keypoint descriptors which are later used for image retrieval, object recognition, copy detection, augmented reality and many others.

The two main limitation for this practice are: (1) keypoint descriptor is the representation of local image patch which is centered to keypoint, this patch can produce the similar descriptor under entirely different scene or context (ambiguity), and (2) the similar patch can produce different descriptors under image noise (indiscrimination). The example of ambiguity and indiscrimination are shown Figure 1.

To overcome these limitation, geometric relationship between local features is widely used by researchers in the application of visual object categorization and copy detection [11]–[14]. It has been argued that the performance of local features is significantly improved from *bag of visual words* to *bag of pair of visual words* [12]. But there are some limitations such as (1) the combination of possible visual word pairs grow quadratically w.r.t the vocabulary size. To overcome this problem, different features are used on these pair of visual words [15]. (2) Different features can be treated as its second limitation as they require additional information like class labels and does not necessarily lead to better performance [15]. These two limitations are discussed in Morioka et al [14], where they propose a reverse technique. Instead paring the visual words, they pair the raw descriptors before learning the visual words. It has been experimentally shown that paring before visual words learning is more effective then visual word paring themselves [13], [14].

In this paper, we extend our previous work [16]. We propose a framework to improve the performance of patch based descriptors for CBCD. In our framework, the descriptors are enriched by geometrically spatial relationship around the keypoint region which makes descriptor more discriminative. The rest of the paper is organized as follows. In Section II, we briefly discuss some related work for content-based image copy detection. In Section III, two famous descriptors are explained, and different approaches to make them robust are discussed. In Section IV, detail experimental evaluation is given, and paper is concluded in Section V.

II. RELATED WORK

This section is divided into two parts. In first part, we briefly discuss about image copy detection, and in second part, we briefly explain two famous patch based descriptors.

A. Content-based image copy detection

Content-based copy detection has been boosted for last six year due to local features. Local features have been proven to be more resistant and robust for severe image transformations compared to global features. The descriptor SIFT [8] is also one of the major reasons for local features popularity. Many CBCD and image retrieval systems have been proposed based on SIFT and other local features such as GLOH [17], CSLBP [18], SURF [9], and BIG-OH [19].

Chang et al. [20] proposed RIME (Replicated IMage dEtector) to detect pirated copies of images on Internet using wavelets and color space. The system has good accuracy for basic types of transformations. Kim [21] use Discrete Cosine Transform (DCT) for CBCD, as DCT is more robust to many distortions and changes in images. They converted the images into YUV format and only Y component is used in proposed method, as they argue that colors do not play important role in copy detection but colors are vital part in image retrieval (images similar based on color, texture, or objects). They successfully detected the copies of the test images with and without modifications, however they fail to detect the copies with 90° or 270° rotation [22]. Basit et al. [23], [24] proposed a method joint localization to track the target with unmanned vehicles. The method fuses the robot kinematics and target dynamics in single space model to produce better results. The global features are efficient for simple types of transformations, however, in case of severe transformations the performance of global features is very poor, for example, in case of cropping, occlusion, and aspect ratio change.

Xu et al. [25] proposed CBCD system based on SIFT and spatial features. They detect homogeneous and larger

circular patches using SIFT detector, and then multi-resolution histograms are computed as feature vectors. However, the performance of this system is poor in occlusion types of transformations. Zhou et al. [2] proposed a framework for partial image copy detection for large scale applications by using bag-of-visual-words model. They quantized the SIFT in descriptor space and orientation space. They encode the spatial layout of keypoints by XMAP and YMAP strategy, which helps to remove the outliers. However, their framework is not effective for object retrieval type of applications.

B. Patch based descriptors

Most of the successful feature point descriptors described in the literature can be classified into two types: those based on gradient histograms [8], [9], [17], [26] and those based on local pixel intensity differences without explicit gradient calculations [18], [27]–[29] We will call the latter group "intensity-based" descriptors. Two famous descriptors are used in the proposed framework for experiments: Scale-invariant feature transform (SIFT) [8], and Center-Symmetric Local Binary Pattern (CSLBP) [18].

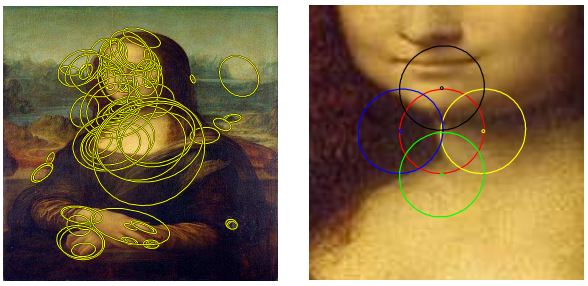
1) Center-Symmetric Local Binary Pattern: CSLBP is an extension of Local Binary Patterns (LBP). In LBP, each pixel value p is compared with its N neighbors with radial distance R. If the pixel value of p is less then its neighbors then output is set to one else output is set to zero. For every p there will be N comparisons and the output for each pixel p will be of N bits and that can be presented by decimal number. In practice, the values of N and R are 8 and 1, respectively. For a given image or patch, the histogram of LBP is computed where the length of histogram is 2^N. Whereas, CSLBP is quantized representation of LBP. In $CSLBP$, instead of comparing each neighbor with p, only center-symmetric neighbors are compared:

$$\begin{aligned} CSLBP_{N,R,T}(p) &= \textstyle\sum_{i=1}^{\frac{N}{2}} s(|n_i| - |n_{i+\frac{N}{2}}|)2^{i-1}, \\ s(j) &= \begin{cases} 1 & j > T \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

The length of histogram in CSLBP is $2^{\frac{N}{2}}$ which is quite shorter than the histogram of LBP. The suggested values for N, R and T are $8, 1, 0.01$, respectively.

To compute the $CSLBP$ descriptor, the given patch P is divided into spatial grid of $G_x \times G_y$ and the histogram of $CSLBP$ is computed for each cell. Finally, all histograms are concatenated to one vector. The length of CSLBP descriptor is $G_x \times G_y \times 2^{\frac{N}{2}}$. That is quite often the double of SIFT descriptor. For our experiments the values for $CSLBP_{N,R,T}$ are $CSLBP_{8,1,0.01}$, and the highest efficiency is obtained by keeping $G_x = 4$ and $G_y = 4$ that makes $CSLBP$ the length of 256.

2) Scale-Invariant Feature Transform: The SIFT descriptor is the representation of gradient orientation histograms. To compute the SIFT descriptor,the given patch P is divided into grid of $G_x \times G_y$. In each cell the gradient magnitude, $g(x, y)$, and orientation, $\theta(x, y)$, are computed for each pixel. The gradient orientation are quantized into 8 directions and histogram of quantized orientation is computed. Each sample added to histogram is weighted by its gradient magnitude and Gaussian weight. For Gaussian weight, circular window with a σ that is 1.5 times that of the scale of keypoint is taken [8]. The Gaussian weight is used to give more preference to those pixels that are near to center. Finally, gradient orientation histograms of all cells are concatenated to one vector, SIFT. The maximum efficiency of SIFT is also obtained by keeping the $G_x = 4$ and $G_y = 4$. Therefore, the SIFT descriptor is of 128 length $(8 \times 4 \times 4)$.



(a) (b)

Fig. 2. Neighbor based descriptor computation. (a) shows the keypoints with their elliptical patches, and (b) shows the neighbors of one randomly selected keypoint.

III. Methodology for descriptors computation

In this section, we explain our methodology for descriptors computation which improves the performance for CBCD.

A. Feature Extraction and Matching

Local keypoints are extracted and represented as $q = (x, y, \theta, \sigma, \mathcal{P}_q, \mathbf{d}_q^c)$, where x and y are the coordinates, θ is dominant orientation, σ is scale, $\mathcal{P}$ is 2D affine region of size 41×41 centered q, and $\mathbf{d}$ is a descriptor vector around keypoint and $c \in \{\text{CSLBP}, \text{SIFT}, \ldots\}$.

Two images are said to be similar, or match, if they have many similar descriptors. Two descriptors $\mathbf{d}_1$ and $\mathbf{d}_2$ are said to be similar if they are close to each other based on some distance measure. In our framework, we have used Euclidean distance as distance measure. The Euclidean distance is defined as

$$\mathcal{E}(\mathbf{d}_1, \mathbf{d}_2) = \sqrt{\sum_{i=1}^{m} (\mathbf{d}_1(i) - \mathbf{d}_2(i))^2}, \tag{2}$$

More specifically, given two images Q and R with local keypoints sets E and F, respectively, we perform nearest neighbor (NN) matching subject to a reliability constraint. We consider the keypoint pair (e_i, f_j), where $e_i \in E, f_j \in F$, to be similar if their descriptors $\mathbf{d}_i$ and $\mathbf{d}_j$ satisfy the following two conditions for distance measure $\mathcal{E}(\cdot, \cdot)$:

- Nearest neighbors

$$\mathcal{E}(\mathbf{d}_i, \mathbf{d}_j) = \min_{\mathbf{d}_k \in F} \mathcal{E}(\mathbf{d}_i, \mathbf{d}_k) \tag{3}$$

- Reliable match

$$T_m \cdot \mathcal{E}(\mathbf{d}_i, \mathbf{d}_j) < \min_{\mathbf{d}_l \in F, l \neq j} \mathcal{E}(\mathbf{d}_i, \mathbf{d}_l) \quad (4)$$

where $T_m > 1$ is a threshold ensuring a stable match under noise conditions.

This method is widely used in computer vision applications [8], [17], [18]. The descriptors with this configuration will be represented by sign D. To make matching fast, we use lookup table in which precomputed distances are stored. In our implementation of descriptors, we used unsigned 8-bits for each element of descriptor vector. Let $\mathcal{N} = \{0, 1, 2, \ldots, 255\}$ be the unsigned 8-bit arrays, then $\mathcal{T}$ is a function defined as follow:

$$\mathcal{T} : \mathcal{N} \times \mathcal{N} \rightarrow \mathcal{D} \quad (5)$$

where $\mathcal{D}$ contains the squared difference between two unsigned 8-bits values. Making use of $\mathcal{T}$ as lookup table, Euclidean distance can be redefined as follow:

$$\mathcal{E}(\mathbf{d}_1, \mathbf{d}_2) = \sqrt{\sum_{i=1}^{m} \mathcal{T}(\mathbf{d}_1(i), \mathbf{d}_2(i))} \quad (6)$$

Using $\mathcal{T}$ we can save two mathematical operations, subtraction and squaring (multiplication). We linearly increase the database size of descriptors and find the first nearest neighbor for single descriptor.

B. Geometrical influenced descriptors computation

For given image, local keypoints are detected. For detected keypoints, affine patches are estimated and normalized and finally respective descriptors are computed, as suggested by Mikolajczyk and Schmid [17]. For keypoints, Harris-affine detector is used which calculates the elliptical like patches for corner like structures [10], [30], [31]. On average, Harris-affine detector returns 1400 points by keeping default parameters on OVG dataset.

For keypoint patch normalization and then descriptor computation, we have used the same steps suggested by Mikolajczyk and Schmid [17].

The performance of descriptors can't be increased by increasing their dimensions with different configuration values of spatial grid (G_x and G_y) nor by increasing the region size around the keypoint (shown in Figure 5). We use different configuration to enrich the descriptor vectors by the gradients or texture in the vicinity. We aim to improve the performance of descriptors before quantization like computation of visual words or codebooks.

1) Configuration I: Pairing K spatially close feature descriptors: In this approach, we use K nearest neighbor approach. Each keypoint q is paired with other keypoint descriptor r which is nearest neighbor of q in descriptor space. We name this approach nearest neighbor descriptor (NND) paring from now onwards. This approach is sensitive to image noise which are discussed in detail in experimental section.

2) Configuration II: Increasing the patch size: To add the spatial information around the keypoint patch, we increase the patch size. The patch size of each keypoint is selected based on their scale to make keypoint scale invariant. Since, the patch $\mathcal{P}_q$ for given keypoint q is invariant to scale and affine distortions, therefore, in general scenario the performance of descriptors can be decreased. However, in case of CBCD, the performance is neither increased nor decreased. Experiments show that the performance in case of JPEG compression is increased.

3) Configuration III: Pairing neighbor region descriptors: In this configuration, we enrich the keypoint descriptors by the geometrical texture and gradient in vicinity. Instead of increasing the scale or adding an other keypoint descriptor to given keypoint descriptor $\mathbf{d}_q$, we take some local image patches near $\mathcal{P}_q$ and add the gradient or intensity information to $\mathbf{d}_q$. This is achieved by taking an other point(s) (x', y') at the pixel distance R_n from keypoint q, the patch over (x', y') is computed by keeping the same parameters of keypoint q except (x, y), finally newly computed descriptors are concatenated to $\mathbf{d}_q$. The example of neighbor local image patches near keypoints is shown in Figure 2. During experiments, we take upto $\mathcal{N}$ neighbor points, where $\mathcal{N} = 4$ and name it neighbor based descriptors (NBD). It is not necessarily needed to compute the full length descriptors from neighbor points. We take less number of spatial grid around the patch to compute the descriptors.

The value of R_n is carefully selected. Since, we are using elliptical regions, so we take *vertex* or *co-vertex* points for the distance of R_n so that neighbor patches have region overlap with $\mathcal{P}$ as shown in Figure 2.

IV. Experiments and Results

To validate the performance of our proposed framework we experiment on two applications. In first one, keypoint matching accuracy provided by homography are computed under different types of image transformations. The transformations mainly include $C4, C5$ and $C7$, the list of transformations can be seen in Table I. In second application, the performance is shown for image retrieval application. For image retrieval, query image is matched with all the images and ranked list based on distances is obtained. The images during this experiments have severe transformations which are the mixture of all listed transformations listed in Table I.

A. Benchmark Datasets

Two different benchmark datasets are used to validate the features. Dataset provided by oxford vision group (OVG)[1] is used for first application experiment. This is standard dataset and used in many papers [17], [18], [32], [33], which comprises of challenging geometric and photometric transformations. Three types of transformations are used in this experiment: illumination change, image blur (also scale change), and JPEG compression. For any given distortion (transformation)

[1] http://www.robots.ox.ac.uk/~vgg/research/affine/

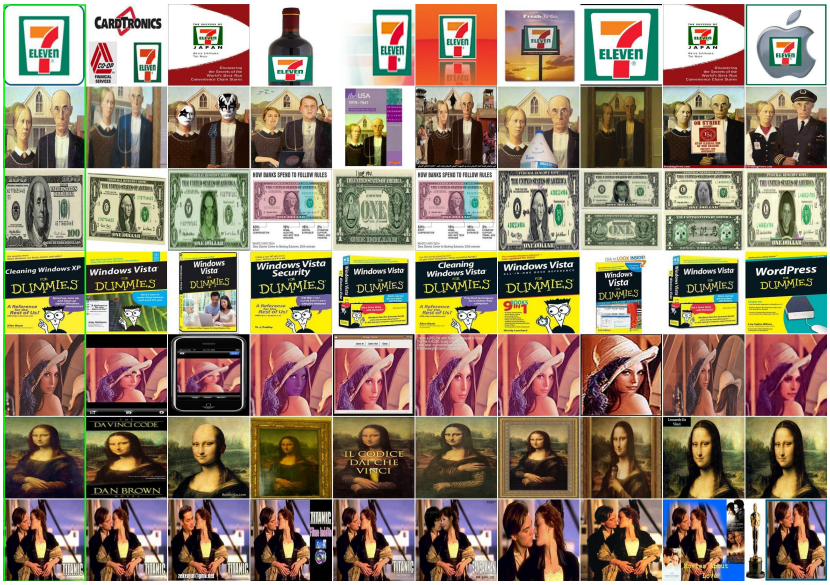


Fig. 3. Dataset used in image-to-image matching for partial image copy detection.

G, base image I_0 of some particular scene is provided with five more gradual distorted images, $S = \{I_0, I_1, \ldots, I_5\}$. Since, distortion get severe gradually therefore I_z has less severe transformation compared to I_{z+1} where $z \in \{1, 2, \ldots, 4\}$, all images in dataset are related by homography. Detail on acquisition of images and transformations in the dataset can be found in the original work [10], [17].

In second experiment, image retrieval datasets known as PICDD [1], [2] and Oxford [34] are used. In PICDD dataset, the images are collected and manually annotated of different scenes and objects. We randomly select 10 objects/scenes with their 10 transformed copies, the transformations are the union of all transformations listed in Table I. The example of random sample from PICDD dataset is shown in Figure 3. The second dataset, Oxford dataset [34], contains 11 landmarks with 55 query images among total of 5K images obtained from Flicker [2].

B. Evaluation matrices

The *precision*, *recall* and *F-score* are used as evaluation matrices

$$\begin{aligned} recall &= \frac{\#\ correct \text{ matches}}{\text{Total } correspondences} \\ precision &= \frac{\#correct \text{ matches}}{\text{Total } matches} \\ F\text{-score} &= 2 \times \frac{precision \times recall}{precision + recall} \end{aligned} \tag{7}$$

For point-to-point correspondences, the number of *correct matches* and *total correspondences* are determined by the *overlap error* [10], [17]. The overlap error is the estimation of region overlap in case of transformations, and can easily be estimated with the help of homography, more details can be found in original work [10], [17]. The values of precision and recall are obtained by the changing the T which is explained in Section III-A. In case of *F-score*, *precision* is computed on gradually increasing *recall* and only the maximum values of *F-score* are reported. For Experiment II, precision is computed on ranked list. For each image there are 10 copies in database.

[2]http://www.flickr.com/

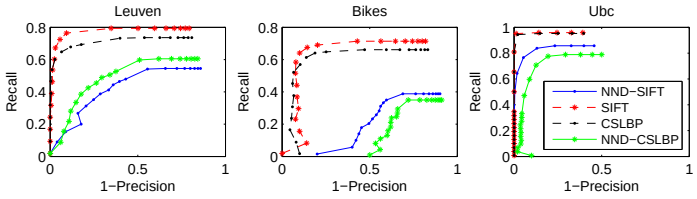


Fig. 4. Performance of NND for Experiment I

C. Experiment I: Point to point correspondences

In this experiment, local keypoints are detected from images $\{I_0, I_1, \ldots, I_5\}$. All descriptors SIFT, CSLBP, NBD-SIFT, NBD-CSLBP, NND-SIFT, NND-CSLBP, DSIFT, and DCSLBP are computed from all keypoints in all images. In each transformation, image I_0 is matched with rest of the 5 gradual deformed copies. Due to the page and space limitation, we only report the correct and false matches between one pair imagesI_0 and I_3). As explained in previous section that images are related by homography and number correct matches are obtained by overlap error, and due to homography total number of correspondences are already know. For matching we used the methodology explained in Section III-A. We visualize the correct and false matches by recall vs 1-precision, and F-score.

D. Experiment II: Image retrieval

For Experiment II and dataset PICDD, there are 100 images including 10 query images. For image I_Q, similarity based matching with all the images in databases are computed. Based

TABLE II
LIST OF SPATIAL GRID CONFIGURATION USED FOR DESCRIPTOR COMPUTATION. FOR CONF 1 TO CONF 4, NEIGHBORS ARE PROGRESSIVELY INCREASED FROM 1 TO 4.

Type	Description
Conf 1	Spatial grid around keypoint and neighbor is 1 x 2.
Conf 2	Spatial grid around keypoint and neighbor is 2 x 2.
Conf 3	Spatial grid around keypoint is 3 x 3 and grid around neighbors is 2 x 2.
Conf 4	Spatial grid around keypoint and neighbor is 3 x 3.
SIFT / CSLBP	Spatial grid size $\{(2 \times 2), (3 \times 3), \ldots, (6 \times 6)\}$
DCSLBP	Double support region around keypoint for CSLBP
DSIFT	Double support region around keypoint for SIFT

on similarity score rank list is maintained. The similarity score between I_Q and image I_d is computed as follow

$$\mathcal{M}(I_Q, I_d) = \frac{|I_Q \cup I_d|}{|I_Q|} \times 100 \quad (8)$$

where $|I_Q \cup I_d|$ represents the similar features between I_Q and I_d, and $|I_Q|$ is total number of features in image I_Q. The retrieval accuracy is shown in Figure 6. It can be seen that NBD SIFT retrieve, on average, top three images as true positives whereas all other descriptors retrieve only top one as true positive.

We also evaluate the performance of NBD descriptors on larger dataset (Oxford dataset). We follow same steps used in BIGOH [19]. We evaluate the retrieval performance of proposed descriptors by mean average precision (mAP). This is obtained by computing the matching scores of query images with all the images in the database exhaustively. Rank list is obtained for every query image and calculate the precision of retrieval at that cutoff (the number of copies divided by the number of gallery images with match score above the threshold), and finally mean of average precision is calculated. The mAP is shown in Table III.

E. Results and Discussions

The NND does not add any improvement in descriptors performance as they are very sensitive to image noise. It has been observed that keypoint changes their position in features space over small distortion. The chances to get 1-NN after some image noise is very small. The performance of NND is shown in Figure 4, NND-SIFT and NND-CSLBP are obtained by paring the keypoint descriptors with their first nearest neighbor descriptor. For Configuration II and III, we use different configuration values of spatial grid over patch before descriptors computation. The list of spatial grid configurations is shown in Table II. The different configuration of spatial grid show that descriptors performance can not be improved while increasing the spatial grid size. Whereas, it can be seen that equivalent performance is achieved with

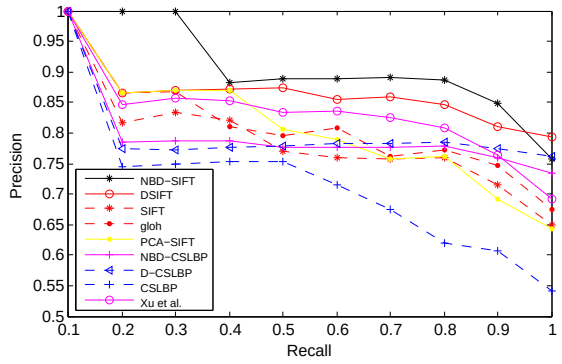


Fig. 6. Performance of SIFT and Neighbor SIFT for image retrieval.

reduced dimensions in the case of NBD descriptors. The descriptors performance for Configuration II and III is shown in Figure 5. For Configuration II, we can see that descriptor performance is not improved. There is only slight different in F-score except JPEG compression, D-CSLBP is better then CSLBP. For Configuration III (neighbor based descriptors), the performance of SIFT and CSLBP is increased. We also get equal performance despite of lower dimensions.

Descriptors computed on many neighbors give better performance, but if the position of keypoint is drastically changed then these descriptors give poor performance. Based on experiments, we recommend to use up to three neighbors. It can be seen that only increasing the scale does not bring any improvement. Instead it improves when descriptor computed from it original patch is concatenated with its neighbors, as affine region around the keypoint have vital contribution for descriptor robustness.

The framework also shows good performance for image retrieval experiment where severely transformed copies of query images are retrieved based on their descriptors matching. Neighbor based descriptors increase the discrimination power and also decrease the ambiguity. The discriminative power of SIFT after enriching it by its neighbor can be seen in Figure 6, where the values of G_x and G_y are (3×3), and for neighbor patches it is 2×2 (only two neighbors are used). The NBD descriptors contains more local information around the keypoint patch therefore when matching score of SIFT and NBD SIFT is computed on true negative images are very different. Ideally, the average matching score with true negative images should be zero but it is not obtained in practical. The average matching score of NBD SIFT is ***0.5*** whereas for SIFT it is **3.7** on true negative images. This clearly shows that NBD based descriptors have less ambiguity. An example of image-to-image matching based on local keypoints is shown in Figure 1 where it is can be seen that NBD SIFT does not have any matching between different images and have correct matching in case of similar images with symmetric structure.

V. CONCLUSION

This paper introduces the neighbor based descriptors. Descriptors are influenced by the texture and gradient in vicinity.

TABLE III
RETRIEVAL ACCURACY (MAP) OF ALL DESCRIPTORS ON LARGER DATASET.

Datasets	SIFT	CSLBP	GLOH	NBD SIFT	NBD CSLBP	D-SIFT	PCA-SIFT
PICDD	0.644	0.510	0.690	0.730	0.590	0.601	0.663
Oxford	0.711	0.602	0.711	0.767	0.610	0.612	0.690

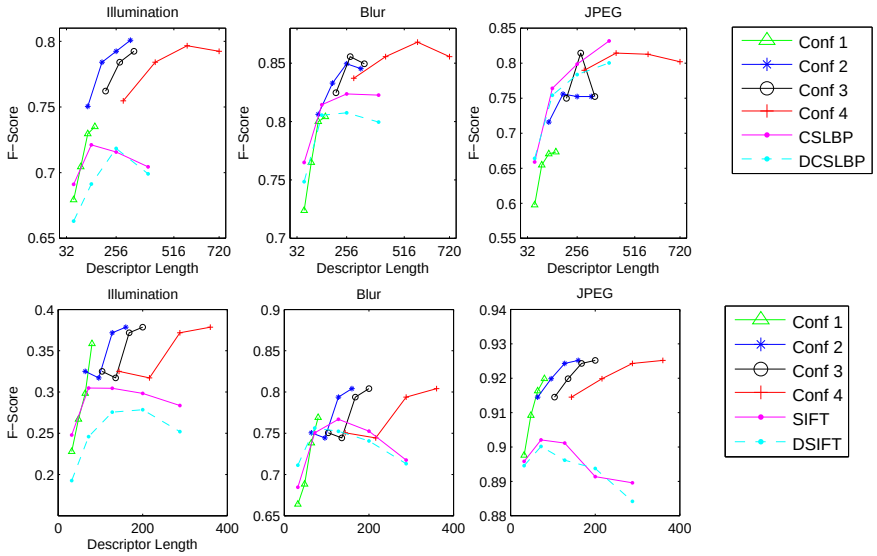


Fig. 5. Results of Keypoint matching for neighbor based descriptors. Top row shows the performance of CSLBP, and bottom row shows of SIFT. Configuration of descriptors computation are shown in Table II. During descriptor computation one to for neighbors are used.

Results on two famous descriptors, SIFT and CSLBP, are shown in Figure 5 and Figure 6. We obtained better performance despite of lower length after influencing the descriptor by their neighbors on particular transformations. Currently, we are trying to explore the retrieval efficiency for very large datasets in real time. Computation of neighbor descriptors increases the computation time. We can also compute hybrid descriptors, adding the power of different descriptors into one. For example, keypoint descriptor can be computed by SIFT and neighbors can be computed by CSLBP or visa-versa.

ACKNOWLEDGMENTS

We are thankful to National Institute of Informatics, Tokyo, for supporting the research.

REFERENCES

[1] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Computer Vision and Pattern Recognition*, 2009.

[2] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the international conference on Multimedia*, 2010.

[3] C.-T. Hsu and J.-L. Wu, "Hidden digital watermarks in images," *Image Processing, IEEE Transactions on*, 1999.

[4] I. Cox, J. Kilian, F. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *Image Processing, IEEE Transactions on*, 1997.

[5] B. Chen and G. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," in *Information Theory, 2000. Proceedings. IEEE International Symposium on*, 2000, p. 46.

[6] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *Image Processing, IEEE Transactions on*, 1999.

[7] S. Lian, N. Nikolaidis, and H. T. Sencar, "Content-based video copy detection a survey," in *Intelligent Multimedia Analysis for Security Applications*, 2010.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, 2008.

[10] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *IJCV*, 2004.

[11] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, 2003.

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

[13] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *ECCV*, 2010.

[14] N Morioka and S Satoh, "Compact correlation coding for visual object categorization," in *ICCV*, 2011.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, 2005.

[16] J. Baber, S. Satoh, C. Keatmanee, and N. Afzulpurkar, "Improving the performance of sift and cslbp for image copy detection," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*. IEEE, 2013, pp. 803–807.

[17] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on PAMI*, 2005.

[18] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, 2009.

[19] J. Baber, M. N. Dailey, S. Satoh, N. Afzulpurkar, and M. Bakhtyar, "BIG-OH: BInarization of gradient orientation histograms," *Image and Vision Computing*, vol. 32, no. 11, pp. 940–953, 2014.

[20] E. Y. Chang, J. Z. Wang, C. Li, and G. Wiederhold, "RIME: A Replicated Image Detector for the World-Wide Web," in *Storage and Retrieval for Image and Video Databases*, 1998.

[21] C. Kim, "Content-based image copy detection," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 169–184, 2003.

[22] Y. Wan, Q. Yuan, S. Ji, L. He, and Y. Wang, "A survey of the image copy detection," 2008, pp. 738 –743.
[23] A. Basit, M. N. Dailey, J. Moonrinta, and P. Laksanacharoen, "Joint localization and target tracking with a monocular camera," *Robotics and Autonomous Systems*, vol. 74, Part A, pp. 1 – 14, 2015.
[24] A. Basit, W. S. Qureshi, M. N. Dailey, and T. Krajník, "Joint localization of pursuit quadcopters and target using monocular cues," *Journal of Intelligent & Robotic Systems*, vol. 78, no. 3, pp. 613–630, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10846-014-0081-2
[25] Z. Xu, H. Ling, F. Zou, Z. Lu, and P. Li, "A novel image copy detection scheme based on the local multi-resolution histogram descriptor," *Multimedia Tools and Applications*, vol. 52, pp. 445–463, 2011.
[26] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Proc. CVPR*, 2004, pp. 511–517.
[27] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *ECCV*, 2010.
[28] M. Heikkilä and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.
[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 11/2011 2011.
[30] B. Adam, "Reliable feature matching across widely separated views," in *CVPR*, 2000.
[31] T. Lindeberg and J. Gårding, "Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure," *Image and Vision Computing*, 1997.
[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *ECCV*, 2010.
[33] M. Ambai and Y. Yoshida, "CARD: Compact And Real-time Descriptors," in *International Conference on Computer Vision*, 2011.
[34] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.