Neurocomputing 113 (2013) 227–233

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

SIFT on manifold: An intrinsic description

Guokang Zhu [a,b], Qi Wang [a], Yuan Yuan [a,*], Pingkun Yan [a]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P.R. China
[b] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, P.R. China

ARTICLE INFO

Article history:
Received 21 July 2012
Received in revised form
2 November 2012
Accepted 13 January 2013
Communicated by M. Wang
Available online 4 March 2013

Keywords:
Computer vision
Manifold learning
Scale invariant feature transform
Locality preserving projections
Descriptor

ABSTRACT

Scale Invariant Feature Transform is a widely used image descriptor, which is distinctive and robust in real-world applications. However, the high dimensionality of this descriptor causes computational inefficiency when there are a large number of points to be processed. This problem has led to several attempts at developing more compact SIFT-like descriptors, which are suitable for faster matching while still retaining their outstanding performance. This paper focuses on the SIFT descriptor and explore a dimensionality reduction for its local representation. By using the manifold learning algorithm of Locality Preserving Projections, a more effective and efficient descriptor LPP-SIFT can be obtained. A large number of experiments have been carried out to demonstrate the effectiveness of LPP-SIFT. Besides, the practicability of LPP-SIFT is also shown in another set of experiments for image similarity measurement.

1. Introduction

Local interest points, extracted from images and described with an effective descriptor, are commonly employed in applications such as object recognition [1–4], robotic visual attention [5], stable target tracking [6,7], image retrieval [8,9] and image/video annotation [10,11]. Local interest points have the advantages of effectiveness, distinctiveness and robustness. In addition, when applied, no preprocessing onto the input images. Therefore, research toward local interest points has been standing as a hot topic over the past years.

As for the description of interest point, probably the most popular work is the one of Lowe's [12]. In this work, a 128-dimensional *Scale Invariant Feature Transform* (SIFT) descriptor is proposed. This descriptor is reported to have achieved tremendous success in a wide range of applications due to its superior computational effectiveness [13]. However, there are also a lot of complaints about the lack of efficiency caused by its high dimensionality. The 128-dimensional description is not an obstacle when only a few detected points are needed to be represented and matched, but it will become a demanding problem when there are millions of points to be processed with limited computational resources. This issue has led to a new task for researchers to develop more compact SIFT-like descriptors suitable for faster matching without loss of matching performance.

This paper focuses on the improvement of SIFT in practice, and explores a dimensionality reduction to its local representation using LPP. The proposed detector is named as LPP-SIFT, which can preserve the intrinsic geometry relationship among interest points with the dimensionality lower than the standard SIFT. Our primary motivation is to develop an intrinsic description more efficient and compact in the context of image similarity comparison while maintaining its excellent performance.

The rest of this paper is organized as follows. Section 2 describes the related works. Section 3 presents a review of the standard SIFT. Section 4 details the proposed LPP-based representation for local features. Section 5 shows the experiments for evaluating the descriptor and Section 6 concludes the paper.

2. Related works

Generally, efforts toward local interest points include two aspects: detector and descriptor. The fundamental requirement for the detector is that the local interest points should be detected with high repeatability, invariant to image transformation and robust to noise. In this aspect, *Laplacian-of-Gaussian* (LoG) is among the earliest detectors that are widely utilized [14,15]. Besides that, there are still other detectors widely employed in applications. For example, Lowe [12] proposes a detector to extract the extremum of *Difference-of-Gaussian* (DoG) which is proved to be the close approximation of the scale-normalized LoG. Mikolajczyk and Schmid [14] propose the Harris-Affine and Hessian-Affine detector, which uses a multi-scale version of the Harris and the Hessian Laplace detector to localize interest points respectively, and then adopt an affine shape-adapted smoothing method [16] for scale selection and affine adaptation. These two

* Corresponding author. Tel.: +86 29 88889302.
E-mail addresses: yuany@opt.ac.cn, yuan369@hotmail.com (Y. Yuan).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.01.020

detectors can obtain higher localization accuracy of local interest points than the DoG based approach, because the latter also responds to edges and detection in this case is unstable [14].

As for local descriptors, probably the most popular one is the SIFT descriptor proposed by Lowe [12]. Several evaluations (e.g., [13]) have demonstrated that this descriptor has superior performance compared to others. Given a local point, SIFT descriptor mainly covers two stages. First, local image gradients are computed around the keypoints and the major orientation of these gradients are obtained. Secondly, 16 histograms of the local patch are calculated and rotated relative to their corresponding major orientations. Each histogram contains eight bins, thus yields a 128 dimensional descriptor.

In light of dimensionality reduction of SIFT-like descriptors, *Principal Component Analysis* (PCA) is a popular and direct choice. Among these algorithms, probably the most representative two are PCA-SIFT [17] and *Gradient Location and Orientation Histogram* (GLOH) [13]. PCA-SIFT detects local interest points as the original SIFT does but describes it on a 39 × 39 patch by a PCA based dimensionality reduction. Ke and Sukthankar show in their work that the resulted feature vector is significantly shorter than the standard SIFT feature vector, while the performance of this descriptor is comparable to the original one. The GLOH descriptor proposed by Mikolajczyk and Schmid [13] constructs a histogram of 272 bins, and then uses PCA to reduce the size of the descriptor. Though proved to be more distinctive than SIFT with the same dimension, the computational cost is somehow expensive. In addition, performing PCA directly to reduce the dimensionality of the standard SIFT descriptors is also a popular way in computer vision community [18,19].

Despite the widespread use in various fields, the validity of PCA is limited by its priori assumption that the relationship among data is linear. However, in real-world applications, it is common where the relation among variables is nonlinear. In this case, nonlinear techniques (e.g., Isomap [20], *Locally Linear Embedding* (LLE) [21,22], and *Laplacian Eigenmap* (LE) [23]) might be more appropriate, which are proposed to discover the submanifold structure hidden in high dimensional ambient space. Though these methods have been successfully applied to some benchmark artificial datasets, the yielded mappings are only defined on the training data points and it is unclear how to extend the mapping for new test data points. Therefore, these nonlinear manifold learning techniques [20–25] are limited in applicability for information comparison tasks. In contrast, the *Locality Preserving Projections* (LPPs) [26] which definitely considers the structure of the manifold may be expediently and reliably applied to any new data point to locate it in the intrinsic low dimensional submanifold.

LPP [26] is a local structure preserving method, which can preserve the intrinsic geometric relationships of the data and share many important properties with nonlinear techniques such as LLE [21] or LE [23]. LPP builds a graph maintaining neighborhood relationship of the given dataset, and then uses the notion of the Laplacian of the graph to compute a projection matrix. This projection matrix can map the high dimensional data points to a subspace, and has the property that local neighborhood information is well preserved. This property makes the algorithm insensitive to outliers and noises to some extent. Since it is likely that a nearest neighbor seek in the locality preserving low dimensional submanifold will yield corresponding results to that in the high dimensional ambient space, the locality preserving quality of LPP is to be of effective and credible use in the information comparison applications.

3. Review of SIFT

The standard SIFT mainly covers three steps. First, keypoint candidates are determined in a series of DoG images by local extremum detection. Second, Taylor expansion of the scale-space function is employed to eliminate the unstable candidates of low distinctiveness and strong edge responses. In the end, local image gradients and orientations are computed around each survived keypoint.

In the first step, the keypoint candidates are identified efficiently by constructing a Gaussian pyramid and obtaining local extremum over a series of DoG images.

In the second step, the scale-space function can be approximated by using a second order Taylor expansion:

$$D(\mathbf{x}+\delta\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T\frac{\partial^2 D}{\partial \mathbf{x}^2}\delta\mathbf{x}, \tag{1}$$

where $\mathbf{x} = (x,y,\sigma)^T$ denotes a keypoint candidate whose coordinate is (x,y) and the scale factor is σ. The function value at the local extremum, $D(\widehat{\mathbf{x}}) = D(\mathbf{x}+\delta\widehat{\mathbf{x}})$, can be obtained by

$$D(\widehat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D}{\partial \mathbf{x}}\delta\widehat{\mathbf{x}}. \tag{2}$$

Consequently, a threshold $\gamma_1 = 0.03$ [12] is adopted to reject keypoint candidates $\{\forall\widehat{\mathbf{x}}, |D(\widehat{\mathbf{x}}) < \gamma_1\}$, because these candidates with low DoG value are also with low contrast and unstable.

Another important characteristic of DoG is that this operator will have strong responses to edges, and detections in this case are unstable. To remove such fake keypoint candidates which "have a large principal curvature across the edge but a small one in the perpendicular direction", Lowe suggests to use a 2 × 2 Hessian matrix H, whose eigenvalues can bo used to estimate the principal curvatures:

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \tag{3}$$

Let $\gamma_2 \geq 1$ be the ratio between the larger eigenvalue and the smaller one. Then algorithm just needs to reject those candidates satisfying:

$$\frac{Tr(H)^2}{Det(H)} \geq \frac{(\gamma_2+1)^2}{\gamma_2}. \tag{4}$$

The third and final step of SIFT identifies the dominant orientation around each survived keypoint and then builds a representation based on a patch of 16 × 16 pixels in its neighborhood. Patches are divided into 4 × 4 blocks. For each block, a histogram indicating eight gradient orientations is produced, and the feature vectors are therefore constructed with 128 dimensions.

The complexity of the SIFT descriptor can be varied by two parameters: the size of the $n \times n$ patch and the number r of orientations in the histograms. The resulting SIFT-like descriptor is of $n \times n \times r$ dimensions. With the increasing complexity, the descriptor will be more distinctive but it will also be overly sensitive to registration error, nonrigid transformation and occlusion. Lowe [12] shows that, setting $n=1$ will be very poor for its discriminative ability and the performance continue to rise until the complexity of the descriptors up to a 16 × 16 patch with eight quantized orientations. Descriptors of higher complexity will actually lead to adverse effects.

4. LPP-SIFT descriptor

Our approach for local descriptors utilizes the same inputs as the standard SIFT (i.e., the location, scale, dominant orientation and local patch of the keypoint). It contains three steps: (1) compute the projection matrix off-line with a set of training patches and descriptions; (2) calculate the SIFT descriptions of the examined keypoints; (3) project the descriptive vectors by the learned matrix to build the more compact descriptions. This reconstructed compact feature vectors are significantly smaller than the standard SIFT feature vectors.

To build the projection matrix, this work executed the standard SIFT on a database of diverse images and collected a large number of sample patches. Each was processed to create a 128-element vector.

Then the LPP technique [26] was chosen to pre-compute the projection matrix. Suppose the dataset of SIFT descriptions is $(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m)$ and let $\mathcal{N}_{\mathbf{x}_i}$ denote the k nearest neighbors of $\mathbf{x}_i$, $\mathcal{N}_{\mathbf{x}_j}$ denote the k nearest neighbors of $\mathbf{x}_j$. Then use $y_i = \mathbf{w}^T\mathbf{x}_i$ to denote the one-dimensional representation of $\mathbf{x}_i$ with the transformation vector $\mathbf{w}$, and define the similarity matrix $S(s_{ij} = s_{ji})$ as follows:

$$s_{ji} = \begin{cases} e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/t} & \text{if } \mathbf{x}_i \in \mathcal{N}_{\mathbf{x}_j} \text{ or } \mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}; \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The criterion for choosing a reasonable projection is to minimize the objective function as follows:

$$f = \frac{1}{2}\sum_{ij}(y_i - y_j)^2 s_{ij}. \tag{6}$$

This objective function undergoes a severe penalty if the neighboring points x_i and x_j are mapped far apart (i.e., $(y_i - y_j)^2$ is large). Therefore, minimizing f can ensure that if x_i and x_j are adjacent then y_i and y_j are close as well. Exercising some simple algebraic deduction, f can be rewritten to

$$\begin{aligned} f &= \frac{1}{2}\sum_{ij}(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 s_{ij} \\ &= \sum_{ij}\mathbf{w}^T\mathbf{x}_i d_{ii}\mathbf{x}_i^T\mathbf{w} - \sum_{ij}\mathbf{w}^T\mathbf{x}_i s_{ij}\mathbf{x}_j^T\mathbf{w} \\ &= \mathbf{w}^T X(D-S)X^T\mathbf{w} \\ &= \mathbf{w}^T XLX^T\mathbf{w}. \end{aligned} \tag{7}$$

where D is a diagonal matrix $(d_{ii} = \sum s_{ji})$, and $L = D - S$ is the Laplacian matrix [27]. As the bigger value of the d_{ii} corresponds to the more important y_i, there is a natural constraint:

$$Y^T DY = \mathbf{w}^T XDX^T\mathbf{w} = 1. \tag{8}$$

This minimization problem can be predigested to finding:

$$\begin{aligned} &\arg\min_{\mathbf{w}} \quad \mathbf{w}^T XLX^T\mathbf{w} \\ &\quad \text{s.t.} \quad \mathbf{w}^T XDX^T\mathbf{w} = 1, \end{aligned} \tag{9}$$

which can be translated as the generalized eigenvalue problem:

$$XLX^T\mathbf{w} = \lambda XDX^T\mathbf{w}. \tag{10}$$

Let the column vectors $\mathbf{w}_i (i = 0,1,\ldots,l-1)$ be the solution of the above generalized eigenvalue problem, ordered according to their eigenvalues, $\lambda_0 < \cdots < \lambda_{l-1}$. The final $n \times l$ projection matrix W_{LPP}, which projects the n-dimensional descriptive vector to the lower l-dimensional representation, is constructed as $W_{LPP} = (\mathbf{w}_0,\mathbf{w}_1,\ldots,\mathbf{w}_{l-1})$.

The images used in building the projection matrix were not used in any of the performance evaluating experiments. An appreciative value (48) was empirically determined for the dimensionality of the feature space in this work. The greater detail described in Section 4.3.1 discussed the effects of n on performance.

5. Experiments

The following parts first describe the experimental setup and then discuss the evaluation metrics used to compare the descriptors. In the end, experimental results are shown and analyzed in detail.

5.1. Experimental setup

Two types of experiments have been done to prove the effectiveness of the proposed descriptor. The localization and scale of the local regions are estimated by the Hessian-Affine or Harris-Affine detector [13].

The first type of experiments evaluates the descriptors' robustness. These experiments employ a dataset of 100 real-world images randomly sampled form a public data set provided by Achanta et al. [28], and extract more than 37,000 sample patches to train the projection matrix using LPP. In the evaluation process, a data set from the work of [15] is employed. In this process, keypoints in different images are first extracted and described with the obtained projection matrix. Then their matches are identified to see whether the descriptor is robust enough to find correspondences in various conditions.

The robustness is defined in terms of four different transformations (Fig. 1): (1) Scale change and rotation: the camera zoom is modulated and rotated at approximately 30–45 degrees around its optical axis. (2) Viewpoint change: the camera position is changed according to a series of significant fronto-parallel angles, which are approximately 50–60 degrees. (3) Blur: images are directly obtained by modulating the camera zoom and focus respectively. (4) Illumination change: photoing condition is changed with the camera aperture.

Each of the above four tests involves a sequence of six images with different geometric or photometric transformation. The geometric relationships between images are known or could be computed by a set of parameters. Therefore, this predefined mapping between images could be used to determine the ground truth matches.

The second type of experiments attempt to evaluate each descriptor's performance when integrating it into an image similarity comparison application. The database employed here is collected by Ke and Sukthankar [17], which contains 30 images with 10 common household items photographed from different viewpoints, scales and lightening conditions.

5.2. Evaluation metrics

The definition for a correct match between two keypoint descriptors must to be discussed firstly. There are three common criterions for a match of two keypoint descriptors: (1) threshold; (2) nearest neighbor based matching; (3) nearest neighbor distance ratio.

Suppose A and B are two local descriptors. According to the first criterion, A is a match of B if their Euclidean distance is below a given threshold. In the second criterion, A is defined as a match of B if A is the nearest neighbor to B and the distance between them is below a given threshold. With respect to the third criterion, two local descriptions (A and B) are matched if the ratio of distance (λ) between the first nearest neighbor (B) and the second nearest neighbor is below a given threshold. A comprehensive comparison [13] showed that the third criterion selects the best match based and penalizes those with many similar matches. This can improve the matching precision and exclude false matches. Therefore, this criterion is employed for defining correct matches in both types of experiments.

In the experiments for the robustness evaluation, the criterion called "*recall vs. 1-precision*" is employed. This criterion was first proposed by Ke and Sukthankar [17] and this experiment use its variant proposed by Mikolajczyk and Schmid in [13]. Recall is determined by the number of correct matches (c) with respect to the total number of corresponding regions (η) between two parallel images:

$$recall = \frac{c}{\eta}, \tag{11}$$

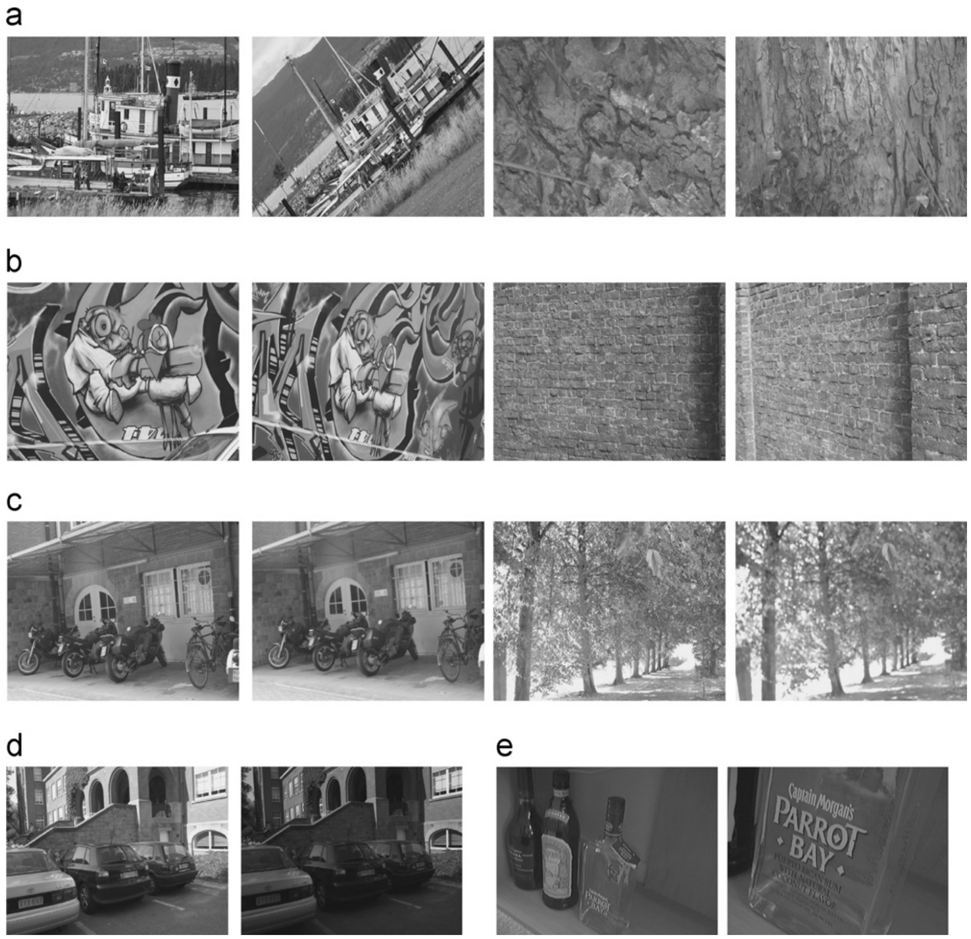


Fig. 1. Sample images used in the experiments. (a) scale change and rotation; (b) viewpoint change; (c) image blur; (d) illumination change; and (e) samples in image similarity comparison application.

and 1-precision is defined as the number of false matches (f) with respect to the total number of matches (τ):

$$1-precision=\frac{f}{\tau}. \tag{12}$$

To explain these two formulas, there need to introduce another concept: the overlap error [29]. The Hessian-Affine detector and the Harris-Affine detector could provide a normalized blob-like region determined by the affine adaptation process [16]. Denoting A, B as the local region and H as the homography between the two images of the same scene, the overlap error is defined as

$$\varepsilon_s=1-\frac{A\bigcap H^TBH}{A\bigcup H^TBH}. \tag{13}$$

A correct match is that the intersection of the corresponding local regions is more than 50% of the union of the two regions, i.e., $\varepsilon_s<0.5$. A false match is the opposite case. The total number of correspondence (i.e., the number of possible correct matches) for the given database is determined with the same criterion.

In the image similarity comparison application, each image was used as a query into the database. Given two images, all the local interest points need to be detected firstly, then their corresponding feature vectors are computed. The number of matched feature vectors between images counted by the "third criterion" described above is treated as a similarity. The results are measured on a three point scoring system as Ke and Sukthankar [17], i.e., 0, 1 and 2 respectively corresponds to the instance of zero, one or two correct matches appeared in the top three similarity comparison results.

5.3. Results

The following part presents the results compared between the LPP-SIFT and some other popular descriptors on transformation controlled experiments and an image similarity comparison application. The comparative descriptors include: the 128-dimensional SIFT, 36-dimensional PCA-SIFT, and 128-dimensional GLOH. For these three comparative descriptors, this paper employed the implementations of [13]. It is reported that the selected dimensionalities can achieve the corresponding best results. Besides, Harris-Affine and Hessian-Affine are selected as the local detectors.

5.3.1. Dimensionality selection

Fig. 2(a) and (b) show experimental results about the relationship between LPP-SIFTs performance and the dimensionality n of the feature vector in condition of image blur. It can be observed that a small value (e.g., $n=4$) of the dimensionality is poor at discriminative ability, but the performance continues to improve until the dimension reaches 48. After that, higher dimensions cannot improve the performance greatly, and on the contrary, the computational cost is intensive.

More experiments are then performed, which showed that the relationships mentioned above is also strongly stable at other three different image transformations. An extrapolation for these results is that the 48 dimensions have already made up 95% of the information contained in full-dimensional feature vector ($n=128$). Therefore, $n=48$ is chosen in the experiments.

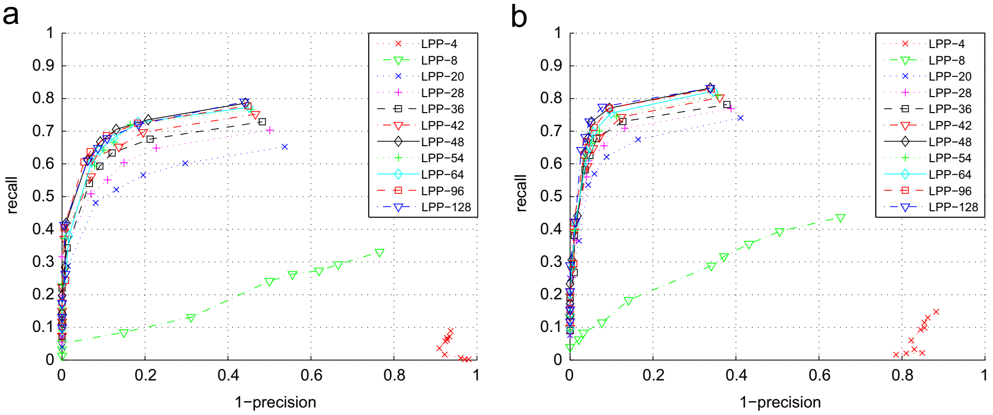


Fig. 2. Evaluation for different dimensions of LPP-SIFT descriptor. (a) and (b) show the results based on Harris-Affine and Hessian-Affine regions respectively.

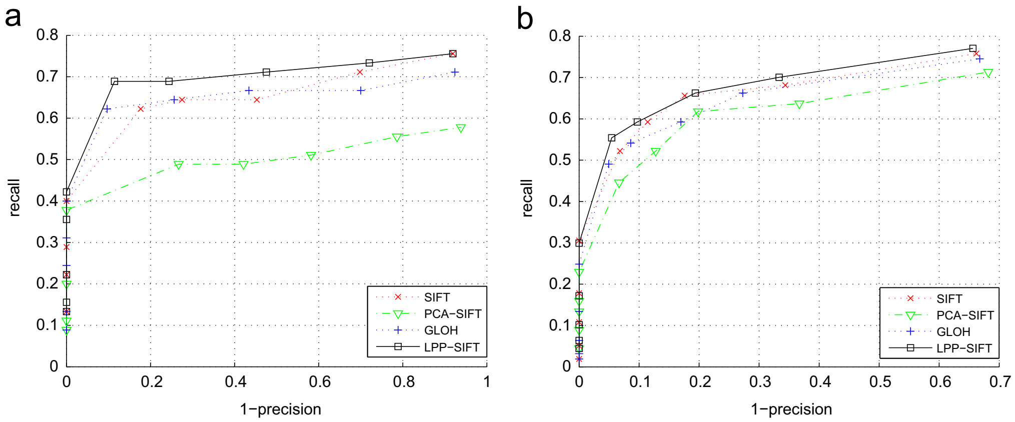


Fig. 3. Evaluation for scale changes of 2–2.5 combined with image rotations of 30°–45°. (a) and (b) show the results based on Harris-Affine and Hessian-Affine regions respectively.

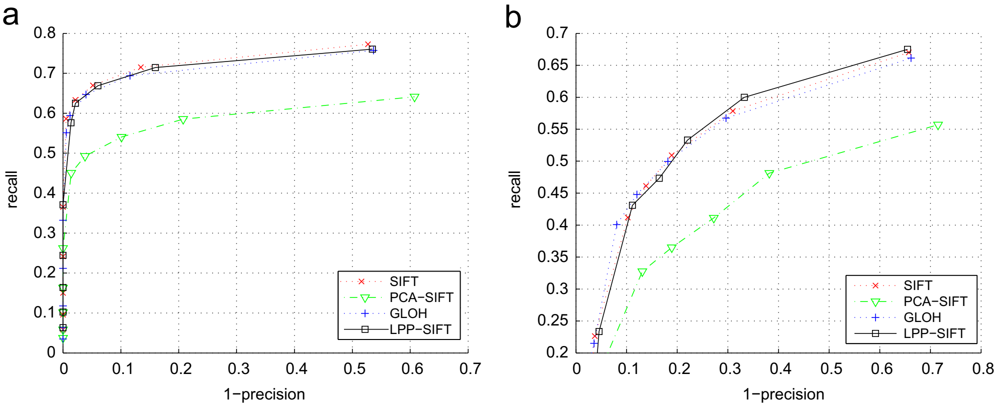


Fig. 4. Evaluation for viewpoint changes of 50–60 degrees. (a) and (b) show the results based on Harris-Affine and Hessian-Affine regions respectively.

5.3.2. Controlled transformation

Figs. 3–6 present the results comparing LPP-SIFT to other descriptors on the first set of experiments.

Fig. 3 shows the descriptors' performance in condition of scale and rotation changes with: (a) Harris-Affine detector and (b) Hessian-Affine detector. Scale changes and image rotations are implemented by modulating the camera zoom in the range of 2–2.5 and rotating it at approximately 30°–45° around its optical axis respectively. As shown in Fig. 3(a), the LPP-SIFT computed on Harris-Affine regions clearly dominates SIFT, PCA-SIFT, and GLOH. The rank of the descriptors computed on Hessian-Affine regions is the same, but the dominance of the LPP-SIFT is weakened as shown in Fig. 3(b).

Fig. 4 shows experimental results based on viewpoint transformation. As can be seen that the LPP-SIFT, standard SIFT and GOLH obtain nearly equivalent performance both on Harris-Affine and Hessian-Affine regions. LPP-SIFT computed on Hessian-Affine regions even perform slightly better than SIFT and GLOH at false matching rates above 0.3. PCA-SIFT computed on both Harris-Affine and Hessian-Affine regions are significantly inferior compared with others.

Robustness to image blur is evaluated in Fig. 5. As can be seen that all the descriptors can resist a significant amount of blur introduced

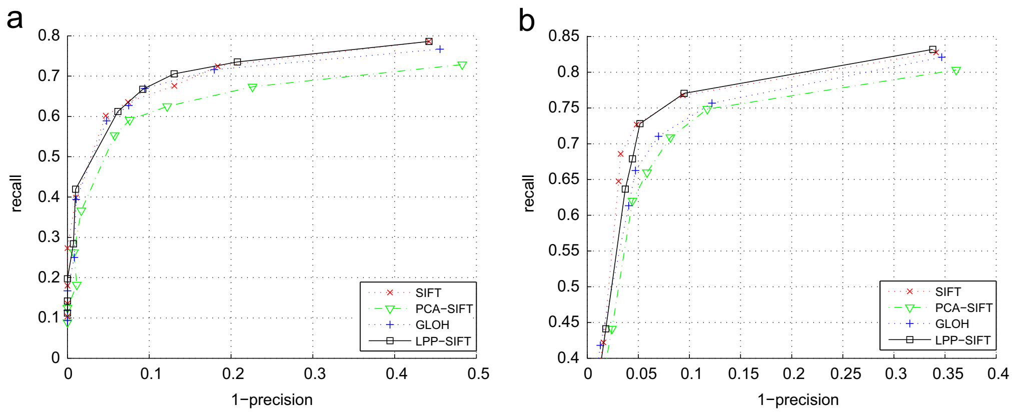


Fig. 5. Evaluation for blur introduced by changing the camera focus. (a) and (b) show the influence of image blur based on local regions detected with Harris-Affine and Hessian-Affine detector respectively.

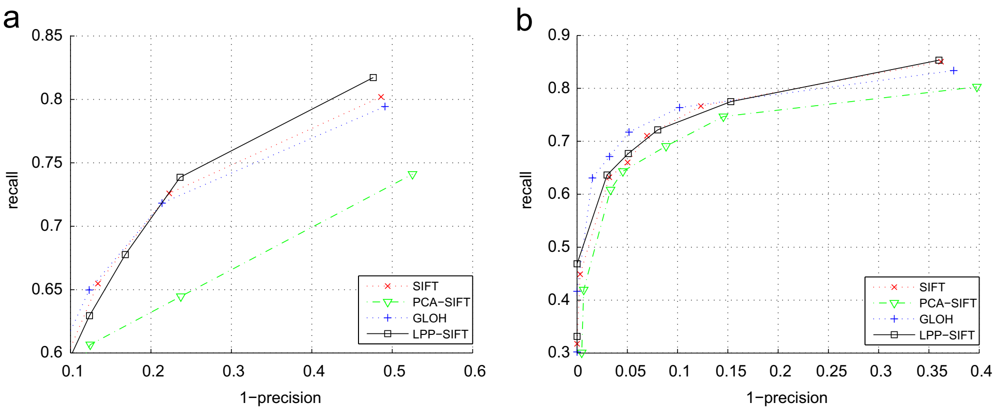


Fig. 6. Evaluation for illumination changes with the camera aperture. (a) and (b) show the results based on Harris-Affine and Hessian-Affine regions respectively.

by changing the camera focus. LPP-SIFT and SIFT computed on both Harris-Affine (Fig. 5(a)) and Hessian-Affine regions (Fig. 5(b)) always obtain better results than the other two obviously. Besides, LPP-SIFT is slightly better than SIFT at false matching rate above 0.1.

Fig. 6 presents results for photoing condition changes. PCA-SIFT computed for both Harris-Affine regions and Hessian-Affine regions obtain the significantly lower score than others. When the descriptors were computed on Harris-Affine regions, LPP-SIFT obtain significantly better score than SIFT and GLOH at false matching rate above 0.2. When the descriptors were computed on Hessian-Affine regions, it is difficult to distinguish the difference between LPP-SIFT and SIFT, whereas they both outperform PCA-SIFT and GLOH at false matching rate above 0.15.

All these comparative results in together are sufficient to demonstrate that, the compact LPP-SIFT is more resistant to geometric or photometric transformations. As all these descriptors shared the same locations and scales of keypoints, it is reasonable to believe that the enhanced robustness of the LPP-SIFT is mainly benefitted by the more instinct descriptions.

5.3.3. *Image similarity comparison application*

In practical applications, the most concerned properties of detectors are the efficiency and accuracy. When calculating Euclidean distance based matching degree between two local keypoints, it is clear that the n-dimensional descriptions will take n times subtraction and multiplication, and one square root operation. Therefore, reducing n can effectively improve the speed of the matching procedures. However, accuracy and speed are often mutually antagonistic. As for keypoint descriptors, an algorithm emphasizing on high speed is often associated with the sacrificing of accuracy, unless it can capture the essential characteristics of the keypoints with low-dimensional structure. In this experiment, the practicabilities of SIFT, PCA-SIFT, GLOH, and LPP-SIFT are compared based on the trade off between speed and accuracy.

Table 1
Image similarity comparison accuracy and time consuming using Harris-Affine detector.

Method	Accuracy (%)	Score of each image	Time (s)
SIFT	65	1.3 ± 0.7479	223.774
PCA-SIFT	61.67	1.2333 ± 0.7279	112.888
GLOH	66.65	1.3333 ± 0.6609	229.860
LPP-SIFT	68.34	1.3667 ± 0.6687	118.434

Table 2
Image similarity comparison accuracy and time consuming using Hessian-Affine detector.

Method	Accuracy (%)	Score of each image	Time (s)
SIFT	70	1.4 ± 0.6215	160.099
PCA-SIFT	60	1.2 ± 0.6644	90.460
GLOH	68.34	1.3667 ± 0.7184	157.995
LPP-SIFT	71.65	1.4333 ± 0.6261	95.789

Tables 1 and 2 present the results for an image similarity comparison application conducted in a small dataset [17]. It can be seen clearly that LPP-SIFT obtains the best matching accuracy with significant lower computational cost than standard SIFT and GLOH. When compare to the PCA-SIFT (36 dimensions), the LPP-SIFT (48 dimensions) clearly dominates the former on matching accuracy with approximately equivalent time cost, so it can be reasonable inferred that using the LPP-SIFT should be of practical benefits.

6. Discussion and conclusion

In this paper, a new compact representation for local image descriptor is reported. To prove the effectiveness and efficiency of the proposed descriptor, a large amount of experiments on robustness are conducted and evaluated in terms of image transformations and image similarity comparison. Experimental results demonstrate that the proposed LPP-SIFT substantially improves matching speed and accuracy in the context of image similarity comparison application. It meanwhile obtains better performance as SIFT in some cases.

When compared to PCA-SIFT and GLOH, LPP-SIFT is more distinctive and robust for both controlled and real-world conditions. The reason is that the local manifold structure is more important than the global structure in some applications such as image similarity comparison. Manifold learning techniques, e.g., LPP can preserve the local structure of the feature space and the intrinsic geometry relationship of the data, while PCA only see the global Euclidean structure which is not discriminative enough. Therefore it is comprehensible that the LPP-SIFT shows some significant improvements over PCA-SIFT and GLOH in both discrimination and robustness.

Acknowledgements

This work is supported by the National Basic Research Program of China (973 Youth Program) (Grant No. 2013CB336500), the National Natural Science Foundation of China (Grant Nos. 61172142, 61172143, and 61105012), and the Natural Science Foundation Research Project of Shaanxi Province (Grant No. 2012JM8024).

References

[1] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1582–1596.

[2] M. Song, Z. Liu, D. Tao, X. Li, M. Zhou, Image ratio features for facial expression recognition application, IEEE Trans. Syst. Man Cybern. Part B (T-SMC-B) 40 (2010) 779–788.

[3] C. Liu, L. Sharan, E.H. Adelson, R. Rosenholtz, Exploring features in a Bayesian framework for material recognition, in: Proceedings of IEEE CVPR, pp. 239–246.

[4] C.-C. Tsai, H.-Y. Lin, J. Taur, C.-W. Tao, Iris recognition using possibilistic fuzzy matching on local features, IEEE Trans. Syst. Man Cybern. Part B (T-SMC-B) 42 (2012) 150–162.

[5] M. Begum, F. Karray, G. Mann, R. Gosine, A probabilistic model of overt visual attention for cognitive robots, IEEE Trans. Syst. Man Cybern. Part B (T-SMC-B) 40 (2010) 1305–1318.

[6] S. Gauglitz, T. Höllerer, M. Turk, Evaluation of interest point detectors and feature descriptors for visual tracking, Int. J. Comput. Vision 94 (2011) 335–360.

[7] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, B. Girod, Unified real-time tracking and recognition with rotation-invariant fast features, in: Proceedings of IEEE CVPR, pp. 934–941.

[8] Y. Ke, R. Sukthankar, L. Huston, An efficient parts-based near-duplicate and sub-image retrieval system, in: Proceedings of ACM-MM, pp. 869–876.

[9] T. Tuytelaars, L.V. Gool, Bilateral filtering for gray and color images, in: Proceedings of Visual Information Systems, pp. 493–500.

[10] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Trans. Multimedia 11 (2009) 465–476.

[11] M. Wang, B. Ni, X.-S. Hua, T.-S. Chua, Assistive tagging: a survey of multimedia tagging with human–computer joint exploration, ACM Comput. Surv. 44 (2012).

[12] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2004) 91–110.

[13] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1615–1630.

[14] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vision 60 (2004) 63–86.

[15] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Gool, A comparison of affine region detectors, Int. J. Comput. Vision 65 (2005) 43–72.

[16] T. Lindeberg, J. Garding, Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure, Int. J. Comput. Vision 15 (1997) 415–434.

[17] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: Proceedings of IEEE CVPR, pp. 506–513.

[18] J.C. van Gemert, J. Geusebroek, C.J. Veenman, A.W.M. Smeulders, Kernel codebooks for scene categorization, in: Proceedings of ECCV, pp. 696–709.

[19] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1271–1283.

[20] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2322.

[21] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[22] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, The J. Mach. Learn. Res. 4 (2003) 119–155.

[23] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural. Comput. 15 (2003) 1373–1396.

[24] M. Brand, Charting a manifold, in: Proceedings of NIPS, pp. 961–968.

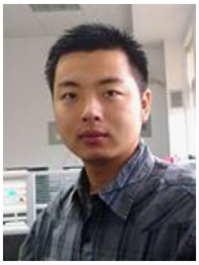
[25] H. Zha, Z. Zhang, Isometric embedding and continuum ISOMAP, in: Proceedings of ICML, pp. 864–871.

[26] X. He, P. Niyogi, Locality preserving projections, in: Proceedings of NIPS, pp. 1–8.

[27] F.R.K. Chung, Spectral Graph Theory, Regional Conference Series in Mathematics, American Mathematical Society, 1997.

[28] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: Proceedings of IEEE CVPR, pp. 1597–1604.

[29] K. Mikolajczyk, C. Schmid, An affine invariant interest point detector, in: Proceedings of ECCV, pp. 128–142.



Guokang Zhu is currently a PhD candidate with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Qi Wang received the B.E. degree in automation and PhD degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010 respectively. He is currently a postdoctoral researcher with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Pingkun Yan received the B.Eng. degree in electronics engineering and information science from the University of Science and Technology of China, Hefei, China and the PhD degree in electrical and computer engineering from the National University of Singapore, Singapore. He is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China. His research interests include computer vision, pattern recognition, machine learning, and their applications in medical imaging.

Yuan Yuan is a researcher (full professor) with Chinese Academy of Sciences, and her main research interests include Visual Information Processing and Image/ Video Content Analysis.