Pattern Recognition Letters 34 (2013) 1211–1220

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

An improvement to the SIFT descriptor for image representation and matching

Kaiyang Liao, Guizhong Liu *, Youshi Hui

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 18 July 2011
Available online 3 April 2013

Communicated by S. Sarkar

Keywords:
Local descriptor
SIFT
Point matching

ABSTRACT

Constructing proper descriptors for interest points in images is a critical aspect for local features related tasks in computer vision and pattern recognition. Although the SIFT descriptor has been proven to perform better than the other existing local descriptors, it does not gain sufficient distinctiveness and robustness in image match especially in the case of affine and mirror transformations, in which many mismatches could occur. This paper presents an improvement to the SIFT descriptor for image matching and retrieval. The framework of the proposed descriptor consists of the following steps: normalizing elliptical neighboring region, transforming to affine scale-space, improving the SIFT descriptor with polar histogram orientation bin, as well as integrating the mirror reflection invariant. A comparative evaluation of different descriptors is carried out showing that the present approach provides better results than the existing methods.

1. Introduction

Local interest point matching is matching the corresponding points between two or more images. It has proven to be very successful in many pattern recognition and computer vision tasks such as wide baseline matching, object recognition and tracking, texture recognition, image retrieval and reconstruction, robot localization, video data mining, building panoramas, stereo correspondence, recovering camera motion, and recognition of object categories (Tuytelaars and Van Gool, 2004; Shin and Tjahjadi, 2010; Zhang and Wang, 2011; Wu and Rehg, 2011; Arican and Frossard, 2012). The two essential aspects of local interest point matching are detection and description of interest points (Li and Ma, 2009). The detection of interest points determines the reliable feature points that are used to match, and at the same time determines the proper neighboring regions that are used in computing the descriptors. The description of an interest point involves creating a distinctive and robust descriptor for it.

For a good feature descriptor, two criteria should be considered in describing the extracted feature points (Li and Ma, 2009). The first is the distinctiveness, which means that the extracted feature should have enough information to distinguish the interest points, so the descriptor of an interest point must be as discriminative as possible. The other criterion is the robustness to resist photometric and geometric deformations. These feature descriptors should be robust with respect to photometric changes such as illumination direction, color, highlight, and intensity change. The extracted features should also be invariant to different geometric variations such as rotation, translation, scaling, mirror reflection and even viewpoint change.

Many different descriptors for interest points have been developed and have proven to be very successful in applications. Excellent reviews on the existing descriptors can be found in Li and Allinson (2008), Brown et al. (2011), Florindo et al. (2012). These descriptors can be divided into five classes:

The first class is the distribution-based descriptors. These techniques used histograms to represent different characteristics of the shape or appearance. Belongie et al. (2002) proposed a shape context descriptor, which at a reference shape point captures the distribution of the remaining points relative to it. Carneiro and Jepson (2002) proposed a phase-based local feature which is based on the phase and amplitude responses of complex-valued steerable filters. Lowe (2004) proposed a scale invariant feature transform (SIFT), which combines a scale invariant region detector and a descriptor based on the gradient distribution in the corresponding regions. Several attempts to improve the SIFT descriptor have been reported in the literature. The PCA-SIFT (Ke and Sukthankar, 2004) descriptor is an extension of the SIFT descriptor, which reduces the dimension of the SIFT descriptor vector from 128 to 36 using PCA. The GLOH (Mikolajczyk and Schmid, 2005) is also an extension of the SIFT descriptor designed to increase its robustness and distinctiveness. Morel and Yu (2009) proposed an affine SIFT, which simulates all the distortions caused by variations in the direction of the camera's optical axis, and then the SIFT is imposed on the simulated images. Guo et al. (2010) presented a mirror reflection invariant descriptor (MIFT) which is inspired from SIFT.

The second class is the differential-based descriptors. These descriptors employ a set of image derivatives computed up to a given order in a point neighborhood. Florack et al. (1991) proposed a

* Corresponding author. Tel./fax: +86 29 82667836.
E-mail addresses: liugz.xjtu@gmail.com, liugz@xjtu.edu.cn (G. Liu).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.03.021

descriptor based on the differential invariants, which combines components of the local derivatives to obtain rotation invariance. Schmid and Mohr (1997) described the interest points using local differential gray-level invariants, and the descriptors are invariant to scale, intensity, and rotation transformations.

The third class is the filter-based descriptors. The steerable filter descriptor (Freeman and Adelson, 1991) employed quadrature pairs of derivatives of Gaussian and Hilbert transforms to synthesize any filter of a given frequency with arbitrary phase. The Gabor filter descriptor (Lee, 1996) used a set of Gabor filters tuned to various frequencies and orientations to represent the image patterns. Baumberg (2000) proposed a complex filter which uses the Gaussian derivatives. Moreno et al. (2009) improved the SIFT descriptor with the Gabor smoothing derivative filters. Gómez and Romero (2011)) introduced a curvelet based descriptor which is calculated from the statistical pattern of the curvelet coefficients.

The fourth class is the color-based descriptors. It makes use of the color invariance robust against varying imaging conditions. Gevers and Smeulders (1999) proposed some new color models for the purpose of recognition of multicolored objects. Diplaros et al. (2006) described a method to merge the color and shape invariant information in the context of object recognition and image retrieval. Abdel-Hakim and Farag (2006) introduced the CSIFT as a colored local invariant feature descriptor. Stokman and Gevers (2007) proposed a generic selection model to select and weight the color invariant models for discriminatory and robust image feature detection. Verma et al. (2011) presented new color SIFT descriptors, which extended the SIFT descriptor to different color spaces.

The fifth class is other descriptors. Apart from the above basic descriptor types, there are also other extended descriptors. Ojala et al. (2002) proposed a local binary pattern (LBP) by building statistics on the local micropattern variations. Chen et al. (2010) developed a Weber local descriptor (WLD) based on the perception of human beings, which is robust to noise. Chen and Sun (2010) presented a new image descriptor to represent the normalized region, which primarily comprises the Zernike moment (ZM) phase information.

The SIFT descriptor is one of the most successful and popular local image descriptor among all the above mentioned descriptors. It has been proven to perform better than the other local invariant feature descriptors (Mikolajczyk and Schmid, 2005) until recent time. However, the SIFT descriptor is neither mirror reflection invariant nor completely invariant to the viewpoint change, and its scale and rotation invariance is not so exact for digital images. In this paper, we propose to improve on the SIFT descriptor by considering all the above mentioned disadvantages. Firstly, a normalized elliptical neighboring region is used to enhance the invariance to viewpoint change. Secondly, the affine scale-space is applied to increase the scale invariance. Thirdly, the polar histogram orientation bin is used to improve the rotation invariance. Finally, rearranging the descriptor is used to ensure the mirror reflection invariance.

The remainder of this paper is organized as follows. Section 2 presents our proposed algorithm. Section 3 introduces the evaluation criteria and the data sets for the experiments. In Section 4 the experimental results and analysis are provided. Finally, the paper is concluded in Section 5.

2. An improvement to SIFT local region descriptor of images

2.1. Normalizing elliptical neighboring region

Recently, several researches in the literature have focused on improving local features to be invariant to the viewpoint change. Mikolajczyk and Schmid (2004) proposed Harris-Affine and Hessian-Affine to obtain invariance to viewpoint change by the affine adaptation process based on the second moment matrix. They make effort to obtain the affine invariance in the stage of feature detection. Morel and Yu (2009) proposed an affine SIFT, which simulates all the distortions caused by variations in direction of the camera's optical axis, and then the SIFT is imposed on the simulated images. Since it has to simulate all the distortions, the affine SIFT will extract many extra feature points which reduce the efficiency of image matching and retrieval. We are going to propose a descriptor to obtain the invariance to viewpoint change by using the normalized elliptical neighboring region in the stage of feature description.

For each interest point, the SIFT descriptor computes the dominant histogram of gradient orientations in a small circular neighboring region around the point. The size of the circular region is determined by the scale of the interest point, but the shape of the neighboring region is not actually invariant to affine transformations for the different structures of neighboring regions (Li and Ma, 2009). For example, the image structure in the red circular region in Fig. 1(a) can be mapped to an elliptical region in Fig. 1(b) after an affine transformation. If we placed the same circle around the feature point in the transformed image which contains an elliptical region, there will be additional image structures in the circular region that will distort any invariant measures calculated, see the red circle in Fig. 1(b). We can see that the same image structure exists in the yellow elliptical region. Since the areas within the two red circles in Figs. 1(a) and (b) do not have the same image structures, using a circular region in the situation of large affine transformation will not produce a robust and distinctive feature descriptor. So we first get an elliptical neighboring region around the interest point, and then normalize the elliptical region to a circular one. We compute the histogram of dominant gradient orientations in the normalized circular neighboring region around the point.

The second moment matrix is often used to detect feature or to describe local image structures. We can use the second moment matrix to estimate the elliptical neighborhood of an interest point (Baumberg, 2000; Lindeberg and Gårding, 1997). After an affine transformation, the second moment matrix μ at a given interest point x is defined by:

$$\mu\left(x,\sum_I,\sum_D\right)=\begin{pmatrix}\mu_{11} & \mu_{12}\\ \mu_{21} & \mu_{22}\end{pmatrix}=\det\left(\sum_D\right)g\left(x,\sum_I\right)*\left((\nabla L)\left(x,\sum_D\right)(\nabla L)\left(x,\sum_D\right)^T\right) \tag{1}$$

where $\sum_D$ is a symmetric positive definite 2×2 matrix corresponding to the local scale and $\sum_I$ is a symmetric positive definite 2×2

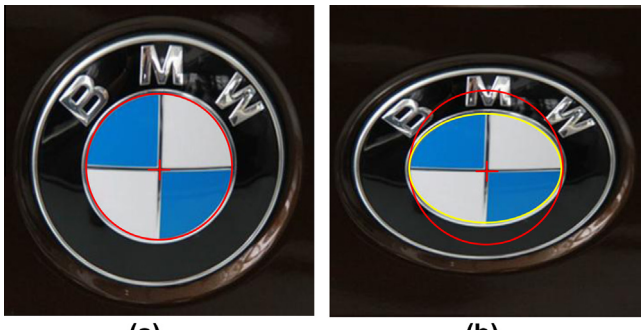

Fig. 1. Circular and elliptical neighboring regions. (a) The image structure in the circle. (b) The mismatched circular region and the matched elliptical region corresponding to the circle in (a) after an affine transformation.

matrix corresponding to the integration scale. Here, $g(x, \sum)$ is the non-uniform Gaussian kernel (Baumberg, 2000):

$$g\left(x, \sum\right) = \frac{1}{2\pi\sqrt{\det \sum}} \exp\left(-\frac{x^T \sum^{-1} x}{2}\right) \tag{2}$$

where $x \in R^2$, and $\sum$ is a symmetric positive definite covariance matrix corresponding to the scale. $L(x, \sum)$ is the affine Gaussian scale-space representation for an image $I(x)$, and it can be generated by a convolution with a non-uniform Gaussian kernel $g(x, \sum)$.

$$L\left(x, \sum\right) = g\left(x, \sum\right) * I(x) \tag{3}$$

$$\nabla L\left(x, \sum\right) = \begin{pmatrix} L_x(x, \sum) \\ L_y(x, \sum) \end{pmatrix} \tag{4}$$

The second moment matrix describes the gradient distribution in a local neighborhood of a point. The gradient derivatives are determined by the matrix $\sum_D$. The derivatives are averaged in the neighborhood of the point by smoothing with a Gaussian which is determined by the matrix $\sum_I$. The eigenvalues of the second moment matrix represent two principal curvatures of a point, which makes it particularly useful for estimating an anisotropic shape of a local image structure.

We can use the second moment matrix to determine a stretch and skew normalized image patch for further processing. The key point is to adapt the shape of a window function based on local image data. Lindeberg and Gårding (1997) described an iterative procedure to adapt the covariance matrix such that the following "fixed point" property holds for the second moment matrix,

$$M = \mu\left(x, \sum_I, \sum_D\right) \tag{5}$$

$$\sum_I = \sigma_I M^{-1} \tag{6}$$

$$\sum_D = \sigma_D M^{-1} \tag{7}$$

In practice for this part of the algorithm, we use the Lindeberg's shape adaptation scheme. For convenience, the iterative adaptation scheme works in the transformed image domain. We first calculate the second moment matrix. We then transform the local image structure using the square root of this second moment matrix which is scaled to have unit determinant. This process is repeated until the second moment matrix is sufficiently close to the identity. Thus, we can estimate the affine transformation between two corresponding points without any prior knowledge about this transformation.

To ensure that each pixel in the elliptical neighboring region of an interest point maps to the correct bin, we normalize the elliptical region to a circular one by using the ellipse parameters from the point's second moment matrix (see Figs. 2(b) and (c)). We can transform the pixel in the elliptical region to a normalized frame using the square root of the second moment matrix $M^{1/2}$. The position x of each pixel which falls within the elliptical region is then mapped to the normalized position x' within the circle region as follows:

$$x' = M^{1/2} x \tag{8}$$

We can obtain a normalized image patch by transforming using the square root of second moment matrix $M^{1/2}$. Mikolajczyk et al. (2005) have shown that any two such normalized patches originating from an image and a linearly distorted copy are related by a rotation. A mathematical proof about affine transformation of a point is given in Appendix A.

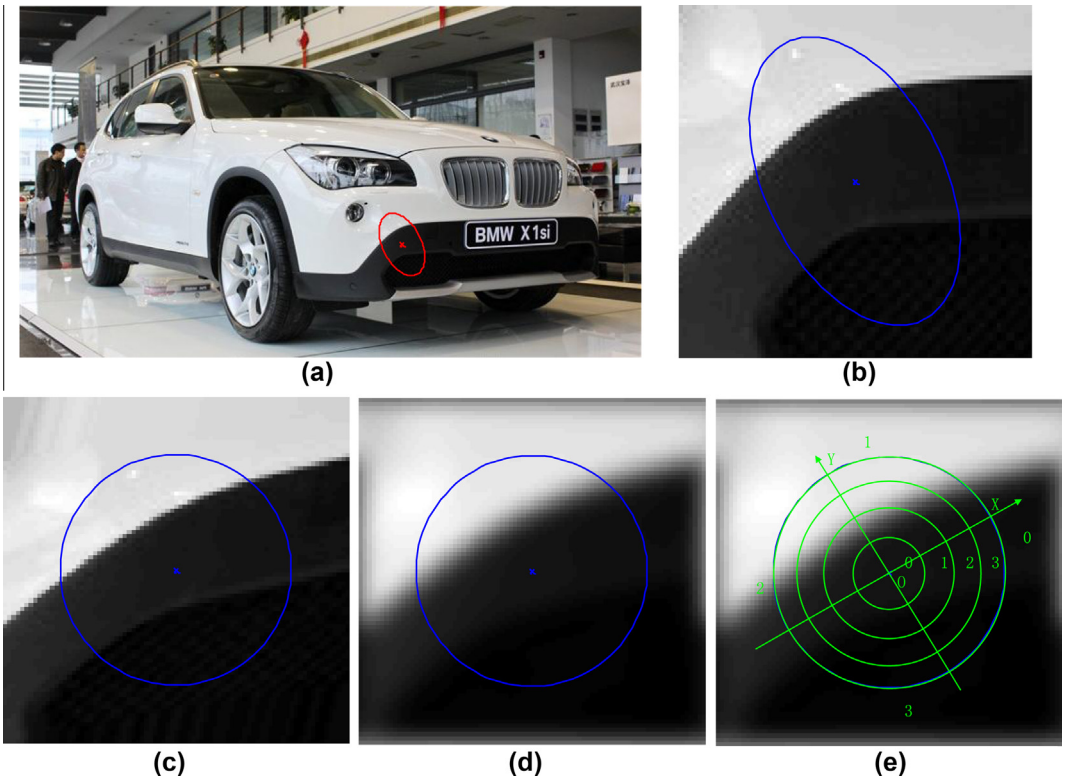

Fig. 2. Illustration of the process in descriptor generating. (a) Estimated interest point neighboring region shape; (b) estimated region shape; (c) normalized region shape; (d) scale-space region; (e) polar histogram orientation bins.

2.2. Affine scale-space based on elliptical neighboring region

The scale parameter in the discrete domain of digital images is also discretized. Thus, the scale-space representation is a set of images represented at different discrete levels of resolution. The scale-space must satisfy the diffusion equation for which the solution is a convolution with the Gaussian kernel. Furthermore this kernel is unique for generating a scale-space representation. The uniqueness of the Gaussian kernel was confirmed by Lindeberg and Gårding (1997), he showed that the convolution with the Gaussian kernel is the best solution to the problem of constructing a multi-scale representation.

For each interest point we create an elliptical neighboring region and we can obtain a normalized circular neighboring region for the point by virtue of the square root of the second moment matrix. Given the normalized circular neighboring region $I'(x)$ of an interest point, the affine scale space representation L' with the scale σ of the point, can be generated by convolution with the Gaussian kernel $g(x, \sigma)$:

$$g(x,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^T x}{2\sigma^2}\right) \tag{9}$$

$$L'(x,\sigma) = g(x,\sigma) * I'(x) \tag{10}$$

The Gaussian kernel is parameterized by one scale factor σ and circularly symmetric. A scale image is obtained by smoothing the image. To obtain the multi-scale representation, the procedure is repeated on the consecutive coarser levels. One can sample the coarser scale image to accelerate the process by the corresponding scale factor after every smoothing operation. In this way, all the computations can be performed in a scale-invariant way by virtue of the Gaussian smoothed image (see Fig. 2(d)).

2.3. Improving SIFT descriptor with the polar histogram orientation bin

Our work is based on an approach similar to SIFT by modification to the neighboring region of the local descriptor. We propose to use a circular region instead of a rectangular region, for the circular region has better rotational invariance.

Assume that we have obtained the affine and scale normalized regions. Firstly, in order to compute the local descriptor we have to compute the derivatives I_{x_1} and I_{x_2} of the normalized region image $I(x)$ with pixel differences:

$$\begin{aligned} I_{x_1}(x_1,x_2) &= I(x_1,x_2+1) - I(x_1,x_2-1),\\ I_{x_2}(x_1,x_2) &= I(x_1+1,x_2) - I(x_1-1,x_2). \end{aligned} \tag{11}$$

The magnitude and orientation of the image gradient are computed in the selected region:

$$M(x_1,x_2) = \sqrt{I_{x_1}(x_1,x_2)^2 + I_{x_2}(x_1,x_2)^2} \tag{12}$$

$$\theta(x_1,x_2) = \tan^{-1}(I_{x_2}(x_1,x_2)/I_{x_1}(x_1,x_2)) \tag{13}$$

The interest region is then divided into fan-shaped sub-regions. We then compute a histogram of the distribution of these vectors in a polar space with equally spaced radial bins (see Fig. 2(e)). In detail, if $x^0 = (x_1^0, x_2^0)^T, x = (x_1, x_2)^T$ and α is the orientation, which is determined by the highest peak of the gradient orientation histogram of the image region at x^0. Define

$$r = \sqrt{(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2} \tag{14}$$

$$\rho = \lceil 4r/r_{\max} \rceil \tag{15}$$

and

$$\varphi = \left\lceil \frac{4}{2\pi}\left(\tan^{-1}\left(\frac{x_2 - x_2^0}{x_1 - x_1^0}\right) - \alpha\right)\right\rceil \tag{16}$$

where ρ and φ are the radial and angular bin indices for the vector $x - x^0$, respectively, and $r_{\max}$ is the radius of the normalized circular neighboring region of the interest point.

The next step is to compute the histogram of the gradient orientation. Each pixel's gradient magnitude is weighted by a uniform Gaussian weighting function $w(x_1, x_2)$ and then added to the corresponding orientation bin k. Here

$$k = \left\lceil \frac{8}{2\pi}(\theta(x_1,x_2) - \alpha)\right\rceil \tag{17}$$

$$w(x_1,x_2) = \exp\left(-3\sqrt{(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2}/r_{max}\right) \tag{18}$$

$$h_{c(\rho,\varphi)}(k) = \sum_{(x_1,x_2)\in c(\rho,\varphi)} M(x_1,x_2)w(x_1,x_2), \quad \theta(x_1,x_2) \in \text{bin } k \tag{19}$$

where (x_1, x_2) is the pixel coordinates in the cell $c(\rho, \varphi)$ and its orientation belongs to bin k. The improved SIFT local region descriptor is a concatenation of the gradient orientation histograms for all the cells:

$$u = (h_{c(0,0)}, \ldots h_{c(\rho,\varphi)}, \ldots h_{c(3,3)}) \tag{20}$$

Finally, in order to reduce the effects of uniform illumination changes, the descriptor in Eq. (20) has to be normalized to unit norm. Here, we have set the gradient orientation relative to the region's orientation in Eq. (16) and Eq. (17), so the descriptor is invariant to rotations of the image region.

2.4. Mirror reflection invariance

2.4.1. Mirror reflection

Mirror reflection can be divided into two types: horizontal and vertical reflection. However, a horizontal reflection image can be obtained by first rotating the original image to 180° and then a vertical reflection; and in the same way we can obtain the vertical reflection. So the horizontal reflection and the vertical reflection are equivalent by rotating the coordinate system. Through the above analysis it is clear that a combined horizontal and vertical reflection is equal to the original image rotated to 180°. The relationship between the descriptors of the same regions after specifying the dominant orientations in the original image and in the vertically reflected image is that the row order of the cells is identical but the order of some columns and all the orientations is inversed. Guo et al. (2010) presented a MIFT descriptor to obtain invariance to mirror transformations by reorganizing the component order of vectors of the SIFT descriptor. We also integrate the mirror reflection invariance to the proposed descriptor similar in spirit to MIFT, but the proposed descriptor is based on the polar histogram orientation bin. It not only provides more distinctive features but also reduces calculations. Only two columns' order needs to be changed in the proposed descriptor, but all the columns' order needs to be changed in the MIFT.

2.4.2. Descriptor reconstruction

As in the original SIFT, a $128 = 4 \times 4 \times 8$ dimensional vector is chosen to define the proposed descriptor. Fig. 3(a) shows an interest point with its interest region in the original image, and Fig. 3(b) is the vertically reflected image, their dominant orientations are specified the positive X-axis.

2.4.2.1. Reconstruction of the 16 cells. As shown in Fig. 3(c), the proposed descriptor adopts the column first order encoding. Consequently, the 16 cells are ordered as Fig. 3(g). However, the

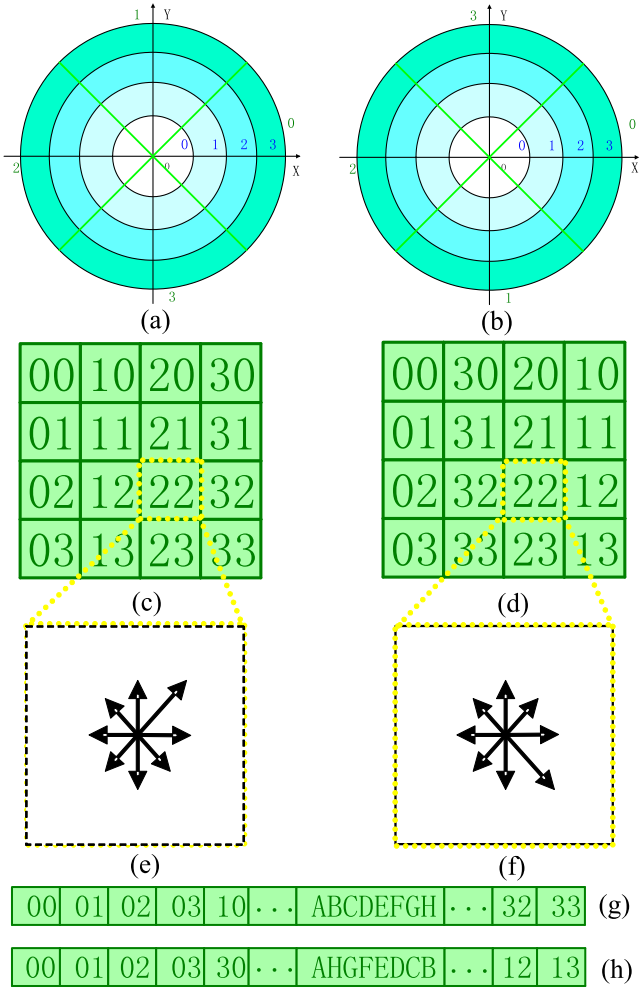

Fig. 3. Illustration of the descriptor organizations of SIFT in the situations with and without mirror reflection. (a) is a feature point with its interest region in the original image; and (b) is (a) in the vertically reflected image; (c) the descriptor organizations in the original image; (d) is (c) in the vertically reflected image; (e) distribution of eight orientations in the 22 cell of (a); (f) distribution of eight orientations in the 22 cell of (b). (g) SIFT descriptor for (a). (f) SIFT descriptor for (b).

column first order is reversed after mirror reflection as Fig. 3(d) shows. The original fixed encoding strategy that was used in SIFT would arrange the 16 cells as shown in Fig. 3(h). Although this encoding strategy is invariant to rotation and scale transformations, and even slightly tolerant to affine transformation, it cannot result in the same order in the situation of mirror reflection. Here, we introduce an adaptive encoding technique that is able to fix the order of the 16 cells irrespective of a mirror reflection. The two summations m_u and m_d of the magnitudes of all the up-pointing (with orientation between 0 and π) and all the down-pointing (with orientation between 0 and $-\pi$) gradients can be used to decide which direction goes first:

$$m_u = \sum_{\rho=0}^{Rbin-1}\sum_{\varphi=0}^{Obin-1}\sum_{k=Nbin/2}^{Nbin-1} h_{c(\rho,\varphi)}(k) \tag{21}$$

$$m_d = \sum_{\rho=0}^{Rbin-1}\sum_{\varphi=0}^{Obin-1}\sum_{k=0}^{Nbin/2-1} h_{c(\rho,\varphi)}(k) \tag{22}$$

where *Nbin* is the number of orientation bins, which is 8 for the proposed descriptor, and *Rbin* and *Obin* are the numbers of radial and angular bins, respectively, which are 4 for the proposed descriptor. According to this measurement, we adaptively change the encoding strategy from the fixed order to the one indicated by the winner of m_u and m_d. From the Fig. 3(c) and (d) we can see that two columns' order needs to be changed in the proposed descriptor.

2.4.2.2. Reconstruction of the 8 orientations. As shown in Fig. 3(e) and (f), every gradient in each cell is specified into its nearest bin of the eight directions. Fig. 3(f) and (e) present the same cell in the circumstances with and without mirror reflection. Consequently, we encode them in the anticlockwise order beginning with 'A' in the case of Fig. 3(a), and in clockwise beginning with 'A' in the case of Fig. 3(b) depending on the relative values of m_u and m_d. As a result, we obtain a unique descriptor for the same interest point in the different mirror reflection cases.

3. Experimental setup

In this section, we first discuss the evaluation metrics used to quantify our results. Then, we introduce the image data used in the experiments.

3.1. Performance evaluation

Some of the results are presented with recall versus 1-precision. The recall is the number of the correctly matched interest regions with respect to the number of the corresponding interest regions between two images of the same scene:

$$recall = \frac{number\ \ of\ \ correct\ \ matches}{number\ \ of\ \ correspondences} \tag{23}$$

1-precision is the number of the false matches relative to the total number of matches:

$$1 - precision = \frac{number\ \ of\ \ false\ \ matches}{number\ \ of\ \ total\ \ matches} \tag{24}$$

For evaluation we use another intuitive evaluation index, the matching score:

$$matching\ \ \ score = \frac{number\ \ of\ \ correct\ \ matchs}{MIN(F_1, F_2)} \times 100\% \tag{25}$$

where F_1 and F_2 are the numbers of interest points on the two images, respectively. A higher score implies that a descriptor is more distinctive and robust.

The average precision (AP) for a single query q is the mean of the precision scores over all the relevant items:

$$AP(q) = \frac{1}{g(q)}\sum_{k=1}^{g(q)}\frac{k}{r(k)} \tag{26}$$

where $g(q)$ is the total number of the ground truth images for the query q. Consider a query q and assume that the kth ground truth image is found to be the Rth result of the retrieval. Then $r(k) = R$. Consequently, the mean average precision (MAP) is the mean of the average precision scores over all the queries:

$$MAP = \frac{1}{Q}\sum_{q\in Q} AP(q) \tag{27}$$

where Q is the set of query q. In the perfect retrieval case with MAP = 1, as the number of the nonrelevant images in the retrieved list increases, the MAP approaches the value 0.

3.2. Datasets

We will compare the descriptors on two image datasets, namely the INRIA dataset (Mikolajczyk and Schmid, 2005) with five geometric and photometric transformations for different scene types with Fig. 4 showing some example images, and the Oxford 5K dataset (Philbin et al., 2007) with some transformations to the images.

3.2.1. INRIA dataset

There are five different changes in imaging conditions: viewpoint changes, scale changes, image blur, JPEG compression, and illumination. We used the images (Fig. 4(a) and (b)) with a rotation angle of about 45° that represents the most difficult case of image rotation, and the scale changes are approximately of a factor 2.0. The images with the viewpoint of the camera are changed by 45° as shown in Fig. 4(c) and (d). As shown in Fig. 4(e) and (f), the images are transformed with significant blur. As shown in Fig. 4(g), the images are changed in JPEG compression for a structured scene. The quality of the image with JPEG compression is 5% of the reference one. Fig. 4(h) shows the images with the occurrence of illumination changes that have been obtained by changing the brightness and the position of the light source.

3.2.2. Oxford 5K dataset

The Oxford Buildings Dataset collected from Flickr by searching for particular Oxford landmarks and is available at http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. In the experiments, we add some deformed images to test the performance of descriptors. The transformation types considered here contain bitrate, noise, change of gamma, mirror reflected, letterbox, logo/text inserted, picture in picture.

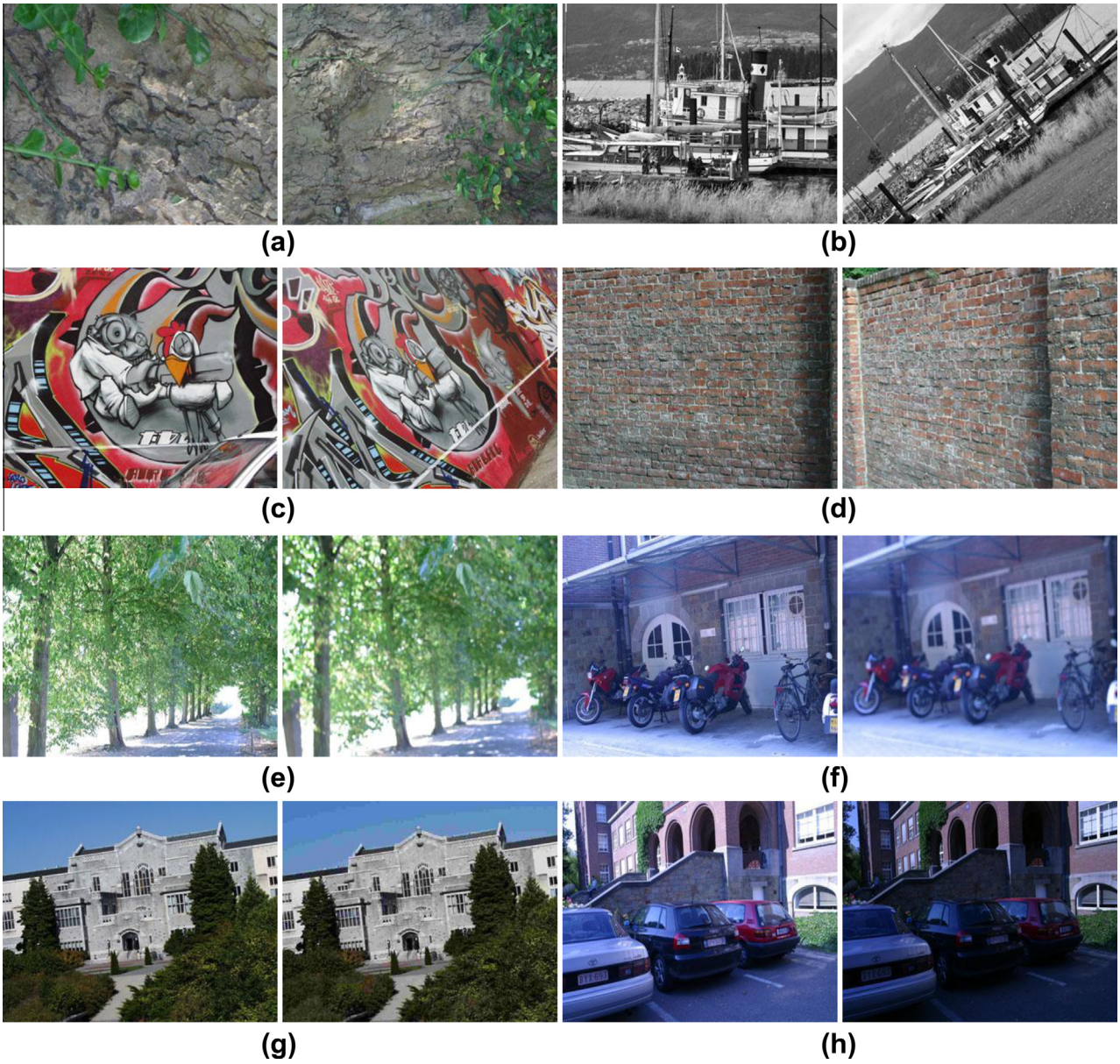

Fig. 4. Sample images of the INRIA dataset used in the experiments: (a) and (b) zoom + rotation, (c) and (d) viewpoint change, (e) and (f) image blur, (g) JPEG compression, (h) light change.

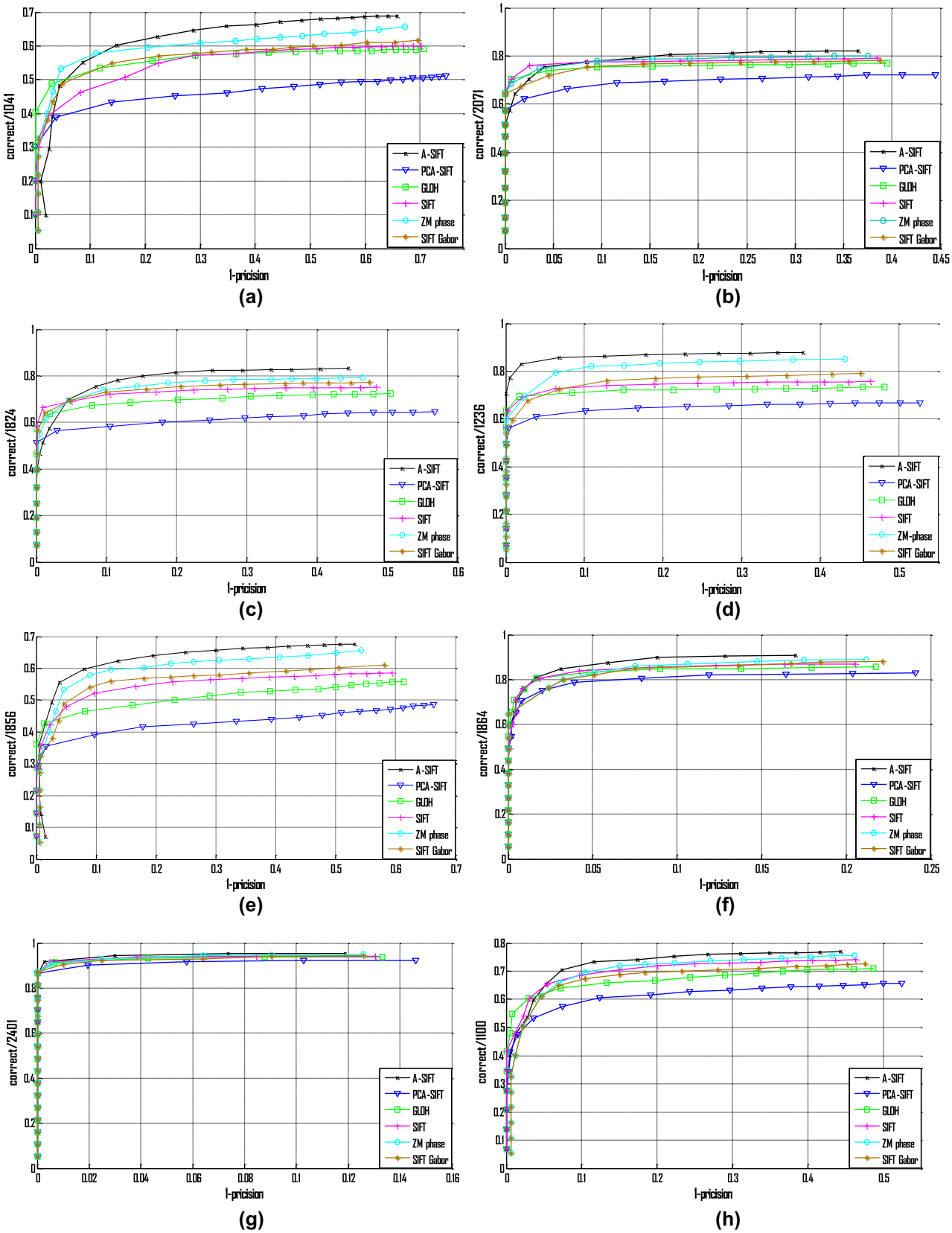

Fig. 5. Illustration of the performance of Recall vs. 1−precision curves; (a) and (b) zoom + rotation, (c) and (d) viewpoint change, (e) and (f) image blur, (g) JPEG compression, (h) light change.

4. Experimental results and analysis

We evaluate the performance of our proposed approach in two related tasks: (i) image region matching, and (ii) object retrieval. In the experiments, all the descriptors are computed on the regions detected with the Hessian Affine detector (Mikolajczyk and Schmid, 2004) for its higher accuracy. There are more than 200,000 region pairs that are detected on the INRIA dataset and over 10 million region pairs are detected on the Oxford 5K dataset.

4.1. Image region matching

The proposed descriptor is compared with five popular descriptors, i.e. SIFT, PCA-SIFT, GLOH, ZM phase and SIFT Gabor, based on the precision-recall criterion with respect to a number of important system parameters. We generate the recall vs. 1−precision graphs for our experiments by varying the threshold for each algorithm. Fig. 5(a)–(h) show the descriptors' performance for the cases of image rotation viewpoint change, image blur, JPEG compression and light change for Fig. 4(a)–(h).

In order to evaluate the performance for image rotation we used the images (as shown in Fig. 4(a) and (b)) with rotation and the scale changes. Fig. 5(a) and (b) show the Recall vs. 1−precision graphs. Our proposed method performs better than the other descriptors for the zoom + rotation transformation. It happens possibly because our approach provides more distinctive information than all the other five descriptors due to applying affine scale-space and polar histogram orientation bin to it. We evaluate the

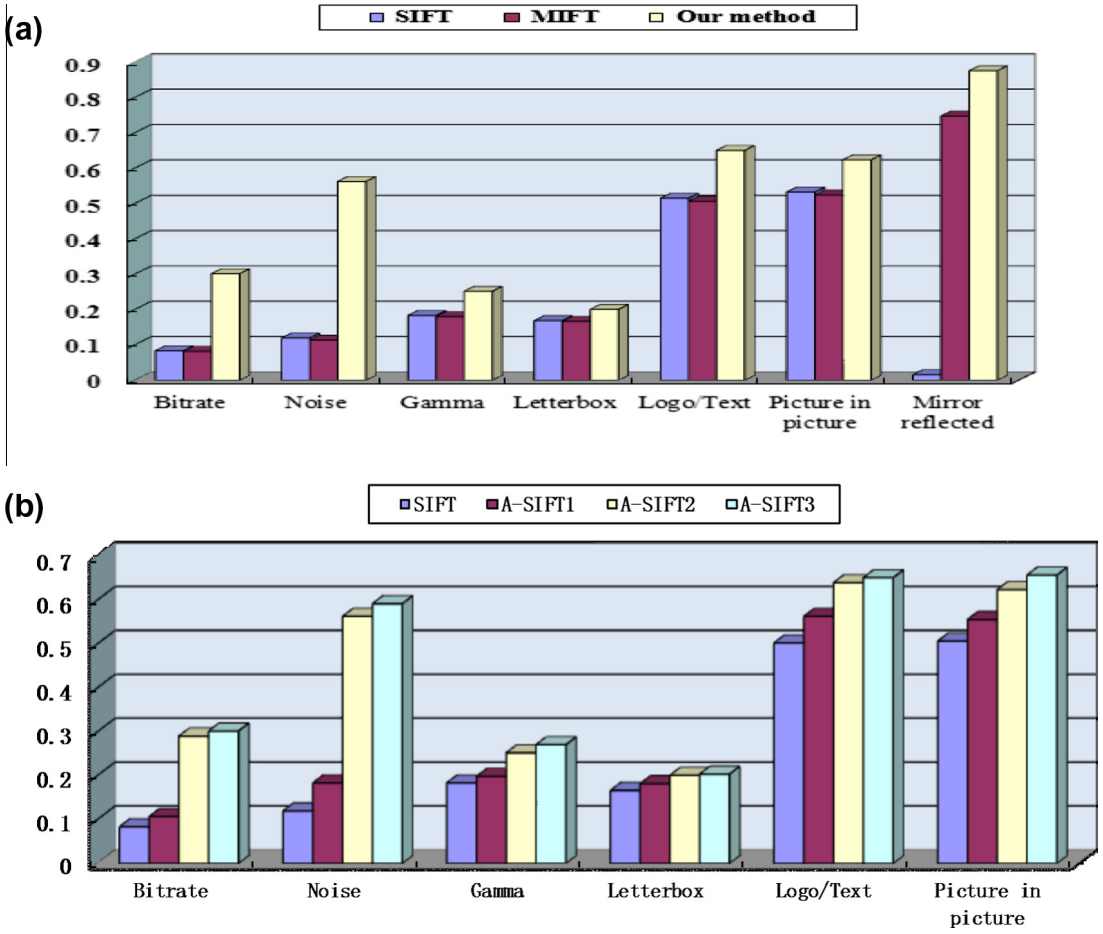

Fig. 6. (a) Illustration the results of matching scores for the seven transformations among our proposed method, MIFT descriptor and SIFT descriptor. (b) Results of matching scores of the SIFT descriptor and the three stage of our proposed method A-SIFT1, 2 and 3 for the six transformations.

Table 1
A summary of MAP performance comparison between our approach and other six approaches for 11 landmarks on the OXFORD 5K dataset.

Landmarks	SIFT	MIFT	PCA-SIFT	GLOH	SIFT Gabor	ZM phase	Our approach
All Souls	0.651	0.65	0.649	0.651	0.649	0.657	**0.674**
Ashmolean	0.662	0.636	0.652	0.664	0.678	0.684	**0.693**
Balliol	0.52	0.508	0.523	0.526	0.524	0.528	**0.537**
Bodleian	0.471	0.441	0.469	0.475	0.488	0.495	**0.516**
Christ Church	0.448	0.425	0.442	0.457	0.465	0.466	**0.488**
Cornmarket	0.605	0.563	0.584	0.604	0.619	0.618	**0.627**
Hertford	0.683	0.652	0.676	0.682	0.709	0.713	**0.741**
Keble	0.593	0.575	0.591	0.606	0.618	0.614	**0.622**
Magdalen	0.571	0.543	0.567	0.573	0.609	0.611	**0.628**
Pitt Rivers	0.657	0.636	0.653	0.666	0.687	0.698	**0.714**
Radcliffe Cam	0.621	0.605	0.627	0.633	0.648	0.647	**0.661**
Average	*0.5893*	*0.5667*	*0.5848*	*0.5943*	*0.6085*	*0.6117*	***0.6274***

performance for images with viewpoint changes of about 45°. There is also some brightness and scale changes in the test images (see Fig. 4(c) and (d)). Fig. 5(c) and (d) shows that our proposed method outperforms the other five approaches. The explanation may be that the descriptor's invariance to the viewpoint change is improved by using the elliptical neighboring region for every feature point. As shown in Fig. 5(e) and (f), the performance is measured for images with a significant amount of blur (see Fig. 4(e) and (f)). In Fig. 5(g), we evaluate the influence of JPEG compression for a structured scene (see Fig. 4(g)). Fig. 5(h) shows the results in the presence of illumination changes (see Fig. 4(h)). Note that the matching performance using our approach is slightly better than those of the other descriptors. The reason may be that our proposed method provides more distinctiveness information by applying the polar histogram orientation bin to it.

The results show that the proposed descriptor has the leading performance under all photometric and geometric transformations. And also the proposed descriptor has the best performances under image blur, JPEG compression and nonlinear lighting.

There are only a few images in the INRIA dataset, and it cannot be used to test the mirror invariance of descriptors. So we added 7 transformations (every transformation contains 10 images) for each query image in the Oxford 5K dataset, and these transformations are not in the INRIA dataset. In this experiment we are test the mirror invariance of the proposed descriptor with SIFT and MIFT in the OXFORD 5K dataset. The other descriptors used in above are not tested for the reason that they are no mirror invariance and the results in above show that the proposed descriptor has better performance than the others. In our proposed method, the descriptors are computed on the normalized image patches. In SIFT and MIFT the descriptors are computed on original image patches. For a given descriptor, we match each feature point in the original image with feature points in the transformed image using nearest neighbor (NN). Fig. 6(a) shows the results of the average match score for seven transformations. It is obvious that the MIFT descriptor performs much better than the SIFT descriptor for the mirror reflected images but has lower scores in the other transformations. This may be caused by the lost accuracy of MIFT in rearranging the component of order of descriptors. Our proposed method performs much better than SIFT and MIFT in all the transformations. The reason is that we not only improve the descriptor's invariance to mirror transformation but also improve the invariance to the scale, rotation and viewpoint change transformations.

In this experiment we aim at studying the influence of the methods (normalizing elliptical neighboring region, transforming to affine scale-space and improving SIFT descriptor with polar histogram orientation bin) on the matching performances. We compare our proposed method with the existing SIFT descriptors on the OXFORD 5K dataset. A-SIFT1, A-SIFT2, A-SIFT3 are our method in the stage of normalizing elliptical neighboring region, transforming to affine scale-space and improving SIFT descriptor with polar histogram orientation bin, respectively. Fig. 6(b) shows the results of the average match scores for the six transformations. We can see that the matching performance of the proposed descriptor is improved at every stage. And we find that the proposed methods are complementary to each other. Moreover, the best result is obtained by using all of them together.

4.2. Object retrieval

In this group of experiments, we perform a comparative study on the performances of our approach and several other descriptors involving SIFT, applied to object retrieval in images. The experiments is performed over a set of datasets provided by OXFORD 5K.

Our object retrieval engine uses the vector space model (Baeza-Yates and Ribeiro-Neto, 1999) of information retrieval. Firstly, the descriptors are extracted from the images in the OXFORD 5K dataset. Then the visual vocabularies with size of 1M words are generated by HKM (Nister and Stewenius, 2006) clustering the descriptors. Finally, all the descriptors are quantized according to the visual vocabularies and are stored in an inverted file (Baeza-Yates and Ribeiro-Neto, 1999). The query and each document in the inverted file are represented as a sparse vector of term (visual word) occurrences, and the retrieval process is calculating the similarity between the query vector and each document vector by use of the Euclidean distance. For computational speed, the engine stores visual word occurrences in the inverted file, which maps individual words to the documents in which they occur. This can result in a substantial speedup, as only the documents which contain certain vectors presented in the query need to be examined. The scores for each document are accumulated so that they are identical to explicitly calculating the similarity.

For each image in the OXFORD 5K dataset, we detect 7 types of descriptors (SIFT, MIFT, PCA – SIFT, GLOH, SIFT Gabor, ZM phase, and our approach). For each type of these descriptors, they are quantized and then used to index the images for the search engine. We compute an AP score for each of the 5 queries for a landmark, averaging these to obtain a MAP score. A summary of performance comparison between our approach and other six approaches is shown in Table 1, which shows the MAP scores for each experiment on the 11 landmarks. From Table 1, we observe that the results obtained with our approach are consistently higher than those with other approaches on all datasets. Some landmarks like Bodleian, the separation between descriptors is larger than that for other landmarks like Cornmarket. The explanation may be related to the transformations (viewpoint, rotation and scale changes) contained in the landmarks like Bodleian. For the transformations in the images, our approach provides more distinctive information than other six descriptors. The results of the datasets average MAP scores for various descriptor methods are shown in the last row of Table 1. We can see that our approach obtains the highest average MAP scores; slightly lower scores are obtained by the ZM phase and the SIFT Gabor.

5. Conclusions

This paper presents a modification to the SIFT descriptor. The proposed descriptor framework consists of the following steps: normalizing elliptical neighboring regions, transforming to affine scale-space, improving SIFT descriptor with polar histogram orientation bin, as well as integrating mirror reflection invariance. In order to evaluate the performance of our descriptor we use the framework of Mikolajczyk and Schmid (2005). Experimental comparisons on seven different feature descriptors involving SIFT have been carried out to evaluate the proposed descriptor, showing that our proposed method is more robust and distinctive than the other descriptors in image match and retrieval.

Acknowledgments

This work was partly supported by funds from National Basic Research Program of China (973 Program No. 2007CB311002), and National High Technology Research and Development Program of China (863 Program No. 2009AA01Z409). We are grateful to David Lowe, Krystian Mikolajczyk and Cordelia Schmid for providing the code for their detectors/descriptors.

Appendix A

The second moment matrix has a property that makes it particularly useful to estimate an anisotropic shape of a local image structure. This property was explored by Lindeberg and Gårding

(1997) and later on by Baumberg (2000), Mikolajczyk and Schmid (2004) to find the affine deformation of an isotropic structure. In the following we present how to determine an anisotropic shape. Consider a point x_i transformed by a linear transformation $x_j = Bx_i$. The second moment matrix μ_i computed in the point x_i is then transformed in the following way:

$$\mu\left(x_i, \sum_{I,i}, \sum_{D,i}\right) = B^T \mu\left(Bx_i, B\sum_{I,i}B^T, B\sum_{D,i}B^T\right)B = B^T\mu\left(x_j, \sum_{I,j}, \sum_{D,j}\right)B \quad (A.1)$$

If we denote the corresponding matrices by:

$$\mu\left(x_i, \sum_{I,i}, \sum_{D,i}\right) = M_i \quad \mu\left(x_j, \sum_{I,j}, \sum_{D,j}\right) = M_j \quad (A.2)$$

these matrices are then related by:

$$M_i = B^T M_j B \quad M_j = B^{-T} M_i B^{-1} \quad (A.3)$$

The derivation and the integration kernels are in this case transformed by:

$$\sum_j = B\sum_i B^T \quad (A.4)$$

Let us suppose that the matrix M_i is computed in such a way that:

$$\sum_{I,i} = \sigma_I M_i^{-1} \quad \sum_{D,i} = \sigma_D M_i^{-1} \quad (A.5)$$

where the scalars σ_I and σ_D are the integration and derivation scales, respectively. We can then derive the following relation:

$$\begin{aligned} \sum_{I,j} &= B\sum_{I,i}B^T = \sigma_I\left(BM_i^{-1}B^T\right) = \sigma_I\left(B^{-T}M_iB^{-1}\right)^{-1} = \sigma_I M_j^{-1} \\ \sum_{D,j} &= B\sum_{D,i}B^T = \sigma_D\left(BM_i^{-1}B^T\right) = \sigma_D\left(B^{-T}M_iB^{-1}\right)^{-1} = \sigma_D M_j^{-1} \end{aligned} \quad (A.6)$$

The affine transformation can then be defined by:

$$B = M_j^{-1/2} R M_i^{1/2} \quad (A.7)$$

where R represents an arbitrary rotation. If the neighborhoods of points x_j and x_i are normalized by transformations $x'_j = M_j^{1/2}x_j$ and $x'_i = M_i^{1/2}x_i$, respectively, the normalized regions are related by a simple rotation $x'_i = Rx'_j$.

$$x_j = Bx_i = M_j^{-1/2}RM_i^{1/2}x_i, \quad M_j^{1/2}x_j = RM_i^{1/2}x_i \quad (A.8)$$

References

Abdel-Hakim, A.E., Farag, A.A., 2006. CSIFT: a SIFT descriptor with color invariant characteristics. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1978–1983.

Arican, Z., Frossard, P., 2012. Scale-invariant features and polar descriptors in omnidirectional imaging. IEEE Transactions on Image Processing 21 (5), 2412–2423.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval, vol. 463. ACM press, New York.

Baumberg, A., 2000. Reliable feature matching across widely separated views. IEEE Conference on Computer Vision and Pattern Recognition, Proceedings, 774–781.

Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4), 509–522.

Brown, M., Hua, G., Winder, S., 2011. Discriminative learning of local image descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (1), 43–57.

Carneiro, G., Jepson, A., 2002. Phase-based local features. Computer Vision-ECCV, 282–296.

Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., et al., 2010. WLD: a robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), 1705–1720.

Chen, Z., Sun, S.K., 2010. A Zernike moment phase-based descriptor for local image representation and matching. IEEE Transactions on Image Processing 19 (1), 205–219.

Diplaros, A., Gevers, T., Patras, I., 2006. Combining color and shape information for illumination-viewpoint invariant object recognition. IEEE Transactions on Image Processing 15 (1), 1–11.

Florack, L., Romeny, B.H., Koenderink, J., Viergever, M., 1991. General intensity transformations and second order invariants. Proceedings of the Seventh Scandinavian Conference Image Analysis, 338–345.

Florindo, J., Backes, A., de Castro, M., Bruno, O., 2012. A comparative study on multiscale fractal dimension descriptors. Pattern Recognition Letters.

Freeman, W.T., Adelson, E.H., 1991. The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9), 891–906.

Gómez, F., Romero, E., 2011. Rotation invariant texture characterization using a curvelet based descriptor. Pattern Recognition Letters.

Gevers, T., Smeulders, W., 1999. Color based object recognition. Pattern Recognition 32 (3), 453–464.

Guo, X., Cao, X., Zhang, J., Li, X., 2010. Mift: a mirror reflection invariant feature descriptor. Computer Vision–ACCV 2009, Part II, LNCS 5995, 536–545.

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: a more distinctive representation for local image descriptors. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 502, II-506–II-513.

Lee, T.S., 1996. Image representation using 2D Gabor wavelets. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (10), 959–971.

Li, C., Ma, L., 2009. A new framework for feature descriptor based on SIFT. Pattern Recognition Letters 30 (5), 544–557.

Li, J., Allinson, N.M., 2008. A comprehensive review of current local features for computer vision. Neurocomputing 71 (10), 1771–1787.

Lindeberg, T., Gårding, J., 1997. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. Image and Vision Computing 15 (6), 415–434.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.

Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. International Journal of Computer Vision 60 (1), 63–86.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (10), 1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., et al., 2005. A comparison of affine region detectors. International Journal of Computer Vision 65 (1–2), 43–72.

Morel, J.M., Yu, G., 2009. ASIFT: a new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences 2 (2), 438–469.

Moreno, P., Bernardino, A., Santos-Victor, J., 2009. Improving the SIFT descriptor with smooth derivative filters. Pattern Recognition Letters 30 (1), 18–26.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2161–2168.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7), 971–987.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching. IEEE Conference on Computer Vision and Pattern Recognition, 1–8.

Schmid, C., Mohr, R., 1997. Local gray value invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (5), 530–535.

Shin, D., Tjahjadi, T., 2010. Clique descriptor of affine invariant regions for robust wide baseline image matching. Pattern Recognition 43 (10), 3261–3272.

Stokman, H., Gevers, T., 2007. Selection and fusion of color models for image feature detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3), 371–381.

Tuytelaars, T., Van Gool, L., 2004. Matching widely separated views based on affine invariant regions. International Journal of Computer Vision 59 (1), 61–85.

Verma, A., Liu, C., Jia, J., 2011. New colour SIFT descriptors for image classification with applications to biometrics. International Journal of Biometrics 3 (1), 56–75.

Wu, J., Rehg, J.M., 2011. CENTRIST: a visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8), 1489–1501.

Zhang, G., Wang, Y., 2011. Robust 3D face recognition based on resolution invariant features. Pattern Recognition Letters 32 (7), 1009–1019.