Rotationally Invariant Descriptors using Intensity Order Pooling

Bin Fan, *Member, IEEE,* Fuchao Wu and Zhanyi Hu

The Authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences.

Digital Object Indentifier 10.1109/TPAMI.2011.277

Abstract

This paper proposes a novel method for interest region description, which pools local features based on their intensity orders in multiple support regions. Pooling by intensity orders is not only invariant to rotation and monotonic intensity changes, but also encodes ordinal information into descriptor. Two kinds of local features are used in this paper, one based on gradients and the other on intensities, hence two descriptors are obtained: MROGH and MRRID. Thanks to the intensity order pooling scheme, the two descriptors are rotation invariant without estimating a reference orientation, which appears to be a major error source for the most of existing methods, such as SIFT, SURF and DAISY. Promising experimental results on image matching and object recognition demonstrate the effectiveness of the proposed descriptors compared to state-of-the-art descriptors.

Index Terms

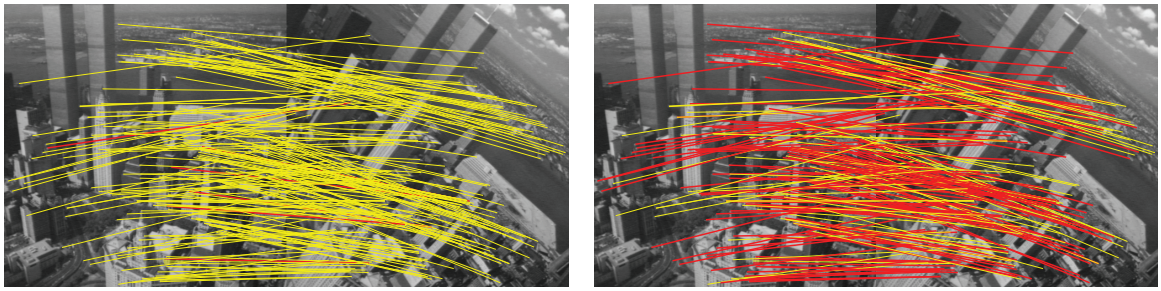
Local Image Descriptor, Rotation Invariance, Monotonic Intensity Invariance, Image Matching, Intensity Orders, SIFT.

I. INTRODUCTION

Local image descriptors computed from interest regions have been widely studied in computer vision. They have become more and more popular and useful for a variety of visual tasks, such as structure from motion [1]–[5], object recognition [6], classification [7] as well as panoramic stitching [8].

A good local image descriptor is expected to have high discriminative ability so that the described point can be easily distinguished from other points. Meanwhile, it should also be robust to a variety of possible image transformations, such as scale, rotation, blur, illumination and viewpoint changes, so that the corresponding points can be easily matched across images which are captured under different imaging conditions. Improving distinctiveness while maintaining robustness is the main concern in the design of local image descriptors. In this paper, we focus on designing local image descriptors for interest regions.

Interest regions are usually detected as affine invariant regions (e.g., Hessian-Affine and Harris-Affine regions [9]) and represented by ellipses, which can be normalized to a canonical region by mapping the corresponding ellipses onto a circle. Such an affine normalization procedure has a rotation ambiguity. In order to alleviate this problem, researchers either designed rotationally invariant descriptors [10]–[13] or rotated the canonical/normalized region according to



(a) By SIFT descriptor (b) By our proposed descriptor (MROGH)

Fig. 1. Matching results of corresponding points which have large orientation estimation errors ($\geq 20^o$) by (a) SIFT and (b) our proposed descriptor. The corresponding points that are matched by their descriptors are marked with red lines while yellow lines indicate those corresponding points that are un-matchable by their descriptors. Most of these corresponding points can be correctly matched by our proposed descriptor (MROGH), but few of them can be correctly matched by SIFT.

an estimated reference orientation of the interest region, such as [6], [14]–[16]. The former achieves rotation invariance at the expense of degrading discriminative ability since this kind of descriptor usually loses some important spatial information, while the latter suffers from the potential instability of orientation estimation.

This paper proposes a novel method for descriptor construction. By using gradient-based and intensity-based local features, two local image descriptors are obtained, MROGH (**M**ulti-Support **R**egion **O**rder-Based **G**radient **H**istogram) and MRRID (**M**ulti-Support **R**egion **R**otation and **I**ntensity Monotonic Invariant **D**escriptor). They are rotation invariant without relying on a reference orientation and still have high discriminative ability. Therefore, they possess the advantages of the existing two kinds of descriptors, but effectively avoid their disadvantages. Our main contributions include:

- A comprehensive reliability analysis of orientation estimation is reported in Section III. We found that the orientation estimation error is one of the major sources of the false negatives (the true corresponding points that are not matched by their descriptors). Since our proposed descriptors calculate local features without resorting to a reference orientation, they are more reliable than those descriptors which rely on a reference orientation for rotation invariance.
- The rotation invariant local features are pooled together based on their intensity orders. Since such a pooling scheme is rotation invariant, the constructed descriptors are rotation

invariant. In addition, the ordinal information is encoded into descriptor by this pooling scheme. Therefore, the constructed descriptors could have high discriminability while be rotation invariant. Fig. 1 shows the matching results of corresponding points whose orientation estimation errors are larger than 20^o (orientations are estimated by the method used in [6]). Although these corresponding points can be hardly matched by SIFT due to large orientation estimation errors, many of them can be correctly matched by our proposed descriptor (MROGH).

- Multiple support regions are used for descriptor construction to further improve its discriminative ability. Although two non-corresponding points may have similar appearances in a certain size of support region, they usually can be easily distinguished in a different size of support region. Therefore, by constructing our descriptors from multiple support regions, their discriminative abilities are further enhanced.

All these factors contribute to the good performance of our proposed descriptors. Their effectiveness and superiorities have been evaluated on image matching and object recognition tasks.

This paper is an extension of our previous work [17]. Specifically, the extensions include:

(1) An in-depth analysis of the rotation invariance of local descriptors.

(2) The MRRID descriptor is introduced to tackle large illumination changes.

(3) A detailed description of the construction of our descriptors as well as a discussion about their properties.

(4) Experiments on object recognition and much more results on image matching.

The rest of this paper is organized as follows: Section II gives a brief overview of related work. Section III presents an experimental study on the rotation invariance of local descriptors. Construction of our proposed descriptors is elaborated in Section IV, followed by the experiments in Section V. Finally, conclusions are presented in Section VI.

II. Related Work

Generally speaking, there are three main steps in matching points by local image descriptors. The first step is to detect points in images. The detected points should be detectable and matchable across images which are captured under different imaging conditions. Such points are called interest points or feature points in the literature. Harris [18] and DoG (Difference of Gaussian) [6]

points are two types of popular interest points. Feature point detection is usually followed by an additional step of detecting affine invariant region around the interest point in order to deal with large viewpoint changes. Many methods in the literature have been proposed for this purpose. Widely used methods include the Harris-Affine/Hessian-Affine detector [9], Maximally Stable Extremal Regions (MSER) [19], intensity-based and edge-based detectors [20]. Please see [21] for a comprehensive study of these affine region detectors. Once interest regions have been extracted, they can be normalized to a canonical region which remains the same under affine transformations of the original local patches. Secondly, feature descriptors are constructed from these interest regions (affine normalized) in order to distinguish them from each other. The final step of point matching is to compute the distance between descriptors of two candidate points and to decide whether they are a match or not. Popular decision-making strategies are nearest neighbor (NN) and nearest neighbor distance ratio (NNDR) [22]. It has been shown in [22] that the rank of different descriptors is not influenced by matching strategies. This makes the comparison of local descriptors convenient as they do not need to be compared with all possible matching strategies.

Local image descriptors have received a lot of attention in the computer vision community. Many local descriptors have been developed since the 1990s [6], [14], [15], [23]–[29]. Perhaps one of the most famous and popular descriptors is SIFT (Scale Invariant Feature Transform) [6]. According to the comparative study of Mikolajczyk and Schmid [22], SIFT and its variant GLOH (Gradient Location and Orientation Histogram) outperform other local descriptors, including shape context [23], steerable filters [30], spin images [10], differential invariants [11], moment invariants [31]. Inspired by the high discriminative ability and robustness of SIFT, many researchers have developed various local descriptors following the way of SIFT. Ke and Sukthankar [24] applied PCA (Principal Component Analysis) [32] to gradient patch of keypoint and introduced the PCA-SIFT descriptor which was said to be more compact and distinctive than SIFT. Bay et al. [27] proposed an effective implementation of SIFT with the integral image technique, and they achieved 3 to 7-fold speed-ups. Tola et al. [16] developed a fast descriptor named DAISY for dense matching. Winder and Brown [33] proposed a framework to learn local descriptors with different combinations of local features and feature pooling schemes. The SIFT and many other descriptors can be incorporated into their framework. A DAISY-like descriptor was reported with the best performance among all configurations. Then, the best DAISY was

picked in [34]. Of course, there are many other local descriptors besides those variants of SIFT. Berg and Malik [35] proposed geometric blur for template matching. Forssen and Lowe [36] used the shape of detected MSER to construct the local descriptor. Chen et al. [26] proposed a descriptor based on the *Weber's Law* [37].

In order to deal with the problem of illumination changes, some researchers utilized intensity orders since they are invariant to monotonic intensity changes. Mittal and Ramesh [38] proposed a change detection method which penalizes an order flip of two pixels according to their intensities. Gupta and Mittal [39] proposed an illumination and affine invariant point matching method which penalizes flips of intensity order of certain point-pairs near keypoints. Gupta and Mittal [25] proposed a monotonic change invariant feature descriptor based on intensity order of point pairs in the interest region. The point pairs are carefully chosen from extremal regions in order to be robust to localization error as well as to intensity noise. Matching this kind of descriptors is based on a distance function that penalizes order flips. Tang et al. [28] used a 2D histogram of position and intensity order to construct a feature descriptor to deal with complex brightness changes. Heikkila et al. [14] proposed to use a variant of LBP (Local Binary Pattern) [40], i.e., CS-LBP which encodes intensity ordinal information locally for feature description, the obtained descriptor was reported to have better performance than SIFT, especially for matching image pairs with illumination changes. Gupta et al. [29] generalized the CS-LBP descriptor with a ternary coding style and proposed to incorporate a histogram of relative intensities in their work. Therefore, their proposed descriptor captures both local orders as well as overall distribution of pixel orders in the entire patch.

Besides low-level features (e.g., histogram of gradient orientation in SIFT) which are used for descriptor construction, choosing an optimal support region size is also critical for feature description. Some researchers reported that a single support region is not enough to distinguish incorrect matches from correct ones. Mortensen et al. [15] proposed combining SIFT with global context computed from curvilinear shape information in a much larger neighborhood to improve the performance of SIFT, especially for matching images with repeated textures. Harada et al. [41] proposed a framework of embedding both local and global spatial information to improve the performance of local descriptors for scene classification and object recognition. Cheng et al. [42] proposed using multiple support regions of different sizes to construct a feature descriptor that is robust to general image deformations. In their work, a SIFT descriptor is computed for each

support region, then they are concatenated together to form their descriptor. Moreover, they further proposed a similarity measure model, Local-to-Global Similarity model, to match points described by their descriptors.

Our work is fundamentally different from the previous ones. We calculate local features in a rotation invariant way, but many previous methods are not strictly rotation invariant since they need to assign a reference orientation for each interest point, such as SIFT, DAISY and CS-LBP. As shown in our experiments in Section III, the orientation estimation is an error-prone process. Since our proposed descriptors do not rely on a reference orientation, they should be potentially more robust. In fact, the need of an orientation for reference is also a drawback and bottleneck of the previous methods which utilize multiple support regions, hence largely differentiates our method from them. Although some local descriptors such as spin image and RIFT (Rotation Invariant Feature Transform) [10] achieve rotation invariance without requiring a reference orientation, they are less distinctive since some spatial information is lost due to their feature pooling schemes. In our work, local features are pooled together according to their intensity orders. Such a feature pooling scheme is inherently rotation invariant, and also invariant to monotonic intensity changes.

III. An Analysis of the Rotation Invariance of Local Descriptors

Local descriptors in the literature can be roughly divided into two categories concerning rotation invariance: 1) Assigning an estimated orientation to each interest point and computing the descriptor relative to the assigned orientation, such as SIFT, SURF, CS-LBP; 2) Designing the local descriptor in an inherently rotation invariant way, such as spin image and RIFT. Currently, it is commonly believed that local descriptors belonging to the second category are less distinctive than those in the first category since their rotation invariance is achieved at the expense of some spatial information loss. Therefore, many existing local descriptors first rotate the support region according to a reference orientation, and then divide the support region into subregions to encode spatial information.

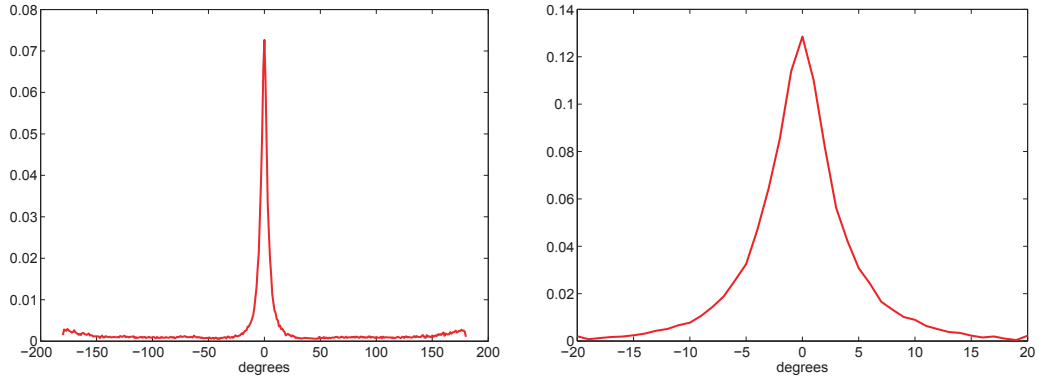
Although SIFT, SURF etc. can achieve satisfactory results in many applications by assigning an estimated orientation for reference, we would claim that the orientation estimation based on local image properties is an error-prone process. We show experimentally that orientation estimation error will make many true corresponding points un-matchable by their descriptors.

To this end, we collected 40 image pairs with rotation transformation (some of them also involve scale changes) from the Internet [43], each of which is related by a homography that is supplied along with the image pair. They are captured from five types of scenes as shown in Fig. 1 in the supplemental material. For each image pair, we extracted SIFT descriptors of the SIFT keypoints and matched them by the nearest neighbor of the distances of their descriptors. We focus on the orientation estimation errors between corresponding points. For a pair of corresponding points (x, y, θ) and (x', y', θ'), the orientation estimation error is computed by:

$$\varepsilon = \theta' - f(\theta; H) \tag{1}$$

where $f(\theta; H)$ is the ground truth orientation of θ when warping from the first image to the second image according to the homography H between the two images. Fig. 2 presents some statistical results. Fig. 2(a) is the histogram of orientation estimation errors among all corresponding points. A similar histogram was obtained by Winder and Brown [33] by applying random synthetic affine warps. Here we use real image pairs with mainly rotation transformation. Fig. 2(b) shows the histogram of orientation estimation errors among those corresponding points that are matched by their SIFT descriptors. Fig. 2(b) indicates that for SIFT descriptor, orientation estimation error of no more than 20^o is required in order to match corresponding points correctly. However, it can be clearly seen from Fig. 2(a) that many corresponding points have errors larger than 20^o. Only 63.77% of the corresponding points have orientation estimation errors in the range of [-20^o,20^o]. Those corresponding points with large orientation estimation errors ($\geq 20^o$) may not be correctly matched by comparing their descriptors. In other words, 36.23% of the corresponding points will be incorrectly matched mainly due to their large orientation estimation errors. Therefore, orientation estimation has a significant impact on distinctive descriptor construction.

In order to give more insight into the influence of the estimated orientation on the matching performance of local descriptor, we conducted some image matching experiments. The experimental images downloaded from the Internet [43] are shown in Fig. 3. The tested image pairs are selected with the most challenging ones among those in the dataset to perform experiments on image blur, JPEG compression and illumination change, i.e., the 1st and the 6th images in the dataset. For each image pair, there exists a global rotation between the two images. Specifically, 0^o for image pairs in Fig. 3(d)-3(f), and a certain degree for image pairs in Fig. 3(a)-3(c). We



(a) (b)

Fig. 2. Distributions of the orientation estimation errors between corresponding points. (a) The distribution among all the corresponding points, only 63.77% of the corresponding points have errors in the range of [-20^o,20^o]. (b) The distribution among those corresponding points that are matched by their SIFT descriptors. See text for details.

used two local descriptors (SIFT and DAISY) with two different orientation assignment methods for image matching:

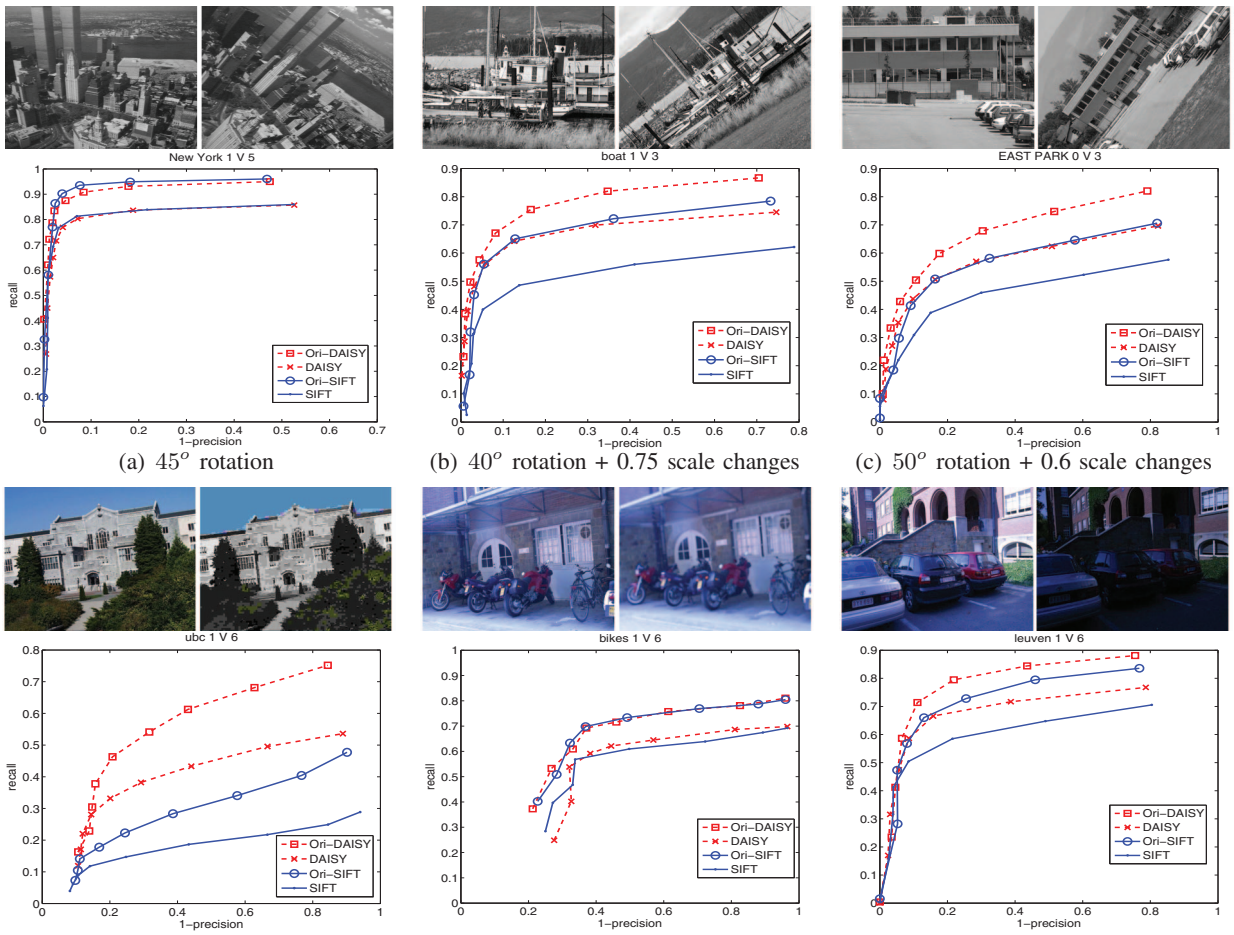
(1) The orientation is assigned by the ground truth orientation, and the constructed local descriptors are denoted as **Ori-SIFT** and **Ori-DAISY**.

(2) The orientation is assigned by the method suggested in [6], and the constructed local descriptors are denoted as **SIFT** and **DAISY**.

By comparing the performance of **Ori-SIFT** (resp. **Ori-DAISY**) with **SIFT** (resp. **DAISY**), we can get a clear understanding of the influence of the orientation estimation on constructing local descriptor for image matching. Experimental results are shown below the tested image pairs in Fig. 3. It can be seen from Fig. 3 that **Ori-SIFT** (**Ori-DAISY**) significantly outperforms **SIFT** (**DAISY**). Since the only difference between **Ori-SIFT** (**Ori-DAISY**) and **SIFT** (**DAISY**) is the orientation assignment method, it demonstrates that a more accurate orientation estimation will largely improve the performance of the local descriptor. The currently used orientation estimation method based on the histogramming technique is still an error-prone procedure and will adversely affect the performance of the local descriptor. Therefore, it leaves a long way for local descriptor to be inherently rotation invariant[1] while maintaining its high distinctiveness. In this paper, we make an effort along this way and propose constructing local descriptors by

[1]It means achieving rotation invariance without resorting to an estimated orientation for reference.

pooling local features according to intensity orders. The proposed two descriptors are not only inherently rotation invariant, but also more distinctive than state-of-the-art local descriptors as shown by experiments in Section V.



(a) 45^o rotation (b) 40^o rotation + 0.75 scale changes (c) 50^o rotation + 0.6 scale changes

(d) JPEG compression (e) image blur (f) illumination change

Fig. 3. Image matching results of SIFT and DAISY with two different orientation assignment methods. See text for details.

IV. The Proposed Method

The key idea of our method is to pool rotation invariant local features based on intensity orders. Instead of assigning a reference orientation to each interest point to make the computation of local features rotation invariant, we calculate local features in a locally rotation invariant coordinate system. Thus they are inherently rotation invariant. Meanwhile, sample points are adaptively partitioned into several groups based on their intensity orders. Then the rotation invariant local

features of sample points in these groups are pooled together separately to construct a descriptor. Since the intensity orders of sample points are rotation invariant, such a feature pooling scheme is also rotation invariant. Therefore, no orientation is required for reference in our method. In this paper, we pool two kinds of local features based on intensity orders to construct local image descriptors: one based on gradients and the other on intensities, hence two descriptors are obtained, named MROGH and MRRID respectively. The details will be described next.

A. Affine Normalized Regions

The detected regions for calculating descriptors are either circular or elliptical regions of different sizes based on the used region detectors. For example, the Hessian-Affine/Harris-Affine detector detects elliptical regions that are affine invariant up to a rotation transformation, while the Hessian-Laplace/Harris-Laplace detects circular regions. Please see [21], [22] for more details about region detectors. To obtain scale or affine invariance, the detected region is usually normalized to a canonical region. Similar to many other local descriptors, this work is to design local descriptor of the normalized region, which is a circular region of radius 20.5 pixels. Thus the minimal patch that contains the normalized region is in size of 41×41 pixels. A similar patch size is also used in [22], [24]. If the detected region is larger than the normalized region, the image of the detected region is smoothed by a Gaussian kernel before region normalization. The standard derivation of Gaussian used for smoothing is set to be the size ratio of the detected region and the normalized region [22].

Given a detected region denoted by a symmetrical matrix $\mathbf{A} \in \Re^{2\times 2}$, for any point $\mathbf{X}$ in the region, it satisfies:

$$\mathbf{X}^T\mathbf{A}\mathbf{X} \leq 1 \tag{2}$$

If $\mathbf{A} = \frac{1}{c^2}\mathbf{E}$ where $\mathbf{E}$ is the identity matrix, then the region is a circular one and c is its radius, otherwise it is an elliptical region. The normalization aims to warp the detected region into a canonical circular region as shown in Fig. 4. The sample point $\mathbf{X}'$ belonging to the normalized region satisfies:

$$\mathbf{X}'^T\mathbf{X}' \leq r^2 \tag{3}$$

where r is the radius of the normalized region, which is set to 20.5 pixels in this paper. Combining

Eq. (2) and Eq. (3), we have

$$\mathbf{X} = \frac{1}{r}\mathbf{A}^{-\frac{1}{2}}\mathbf{X}' = \mathbf{T}^{-1}\mathbf{X}' \tag{4}$$

Therefore, for each sample point $\mathbf{X}'$ in the normalized region, we calculate its corresponding point $\mathbf{X}$ in the detected region and take the intensity of $\mathbf{X}$ as the intensity of $\mathbf{X}'$ in the normalized region, i.e., $I(\mathbf{X}') = I(\mathbf{X})$. Usually, $\mathbf{X}$ is not exactly located at a grid point, so $I(\mathbf{X})$ is obtained by bilinear interpolation. Fig. 4 gives an example of the normalized region.

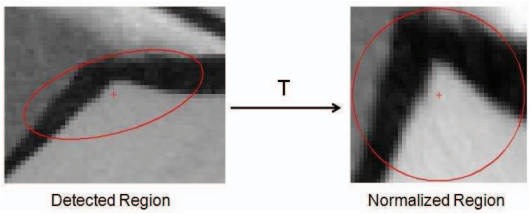


Fig. 4. The affine normalization of a detected region to the canonical circular region (normalized region).

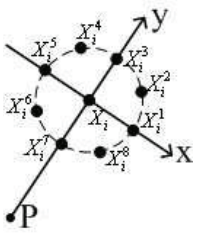


Fig. 5. The locally rotation invariant coordinate system used for calculating local feature of a sample point X_i. P is the interest point (center of the normalized region).

In the reminder of this paper, support regions are such normalized circular regions and intensities of points in the support regions are obtained by bilinear interpolation as described above.

B. Support Region Partition Based on Intensity Orders

Given a support region, one can divide it into several rings and pool together the calculated local features of sample points in each ring in a similar way as spin image or RIFT [10] does, so as to achieve a rotation invariant description of the region. However, such an undertaking will reduce the distinctiveness of descriptor since some spatial information is lost. In other words, pooling local features circularly achieves a rotation invariant representation at the expense of descriptor's discriminative ability degradation. Therefore, many popular and state-of-the-art methods divide the support region into subregions in order to encode more spatial information, such as SIFT [6], DAISY [16], CS-LBP [14], OSID [28] and so on [15], [27], [29]. Unfortunately, these pre-defined subregions need to assign a reference orientation in order to be rotation invariant. As we analyzed in Section III, the orientation estimation is not stable enough, and we adopt here a different

approach. Rather than geometrically dividing the support region into subregions, we partition sample points into different groups based on their intensity orders. In our case, sample points in each group are not necessarily spatial neighbors, and such an adaptive partition approach does not require assigning an orientation for reference.

We denote by $R = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ a support region with n sample points, and $I(\mathbf{X}_i)$ the intensity of sample point $\mathbf{X}_i$. Our goal is to partition R into k groups according to the intensity orders of sample points. Firstly, intensities of sample points are sorted in non-descending order and a set of sorted sample points is obtained as:

$$\{\mathbf{X}_{f(1)}, \mathbf{X}_{f(2)}, \cdots, \mathbf{X}_{f(n)} : I(\mathbf{X}_{f(1)}) \leq I(\mathbf{X}_{f(2)}) \leq \cdots \leq I(\mathbf{X}_{f(n)})\}$$

where $f(1), f(2), \cdots, f(n)$ is a permutation of $1, 2, \cdots, n$. Then we can take $k+1$ intensities from them as follows:

$$t_i = I(\mathbf{X}_{f(s_i)}) : t_0 \leq t_1 \leq \cdots \leq t_k \tag{5}$$

where

$$s_i = \begin{cases} \left\lceil \frac{n}{k} i \right\rceil, i = 1, 2, \cdots, k \\ 1, \qquad i = 0 \end{cases} \tag{6}$$

Finally, the n sample points are partitioned into k groups as:

$$R_i = \{\mathbf{X}_j \in R : t_{i-1} \leq I(\mathbf{X}_j) \leq t_i\}, i = 1, 2, \cdots, k \tag{7}$$

It is worth noting that only the intensity orders of sample points are used. Therefore, the partitioned sample point groups are invariant to monotonic intensity changes. The top of Fig. 7 gives an example of such intensity order based partition for a support region, where different point groups are indicated by different colors.

C. The Computation of Rotation Invariant Local Features

1) The Rotation Invariant Coordinate System: The key for the computation of rotation invariant local features is to construct a rotation invariant coordinate system for each sample point. As shown in Fig. 5, suppose $\mathbf{P}$ is an interest point and $\mathbf{X}_i$ is one of the sample points in its support region. Then a local xy coordinate system can be established by $\mathbf{P}$ and $\mathbf{X}_i$ by setting $\overrightarrow{\mathbf{PX}_i}$ as the positive y-axis for the sample point $\mathbf{X}_i$. Since such a local coordinate system is rotation invariant, we calculate local features in this coordinate system to obtain rotation invariance and

accumulate them in the partitioned point groups to construct our descriptors. Here, we propose two kinds of local features: one based on gradients and the other on intensities.

2) Gradient-Based Local Feature: For each sample point $\mathbf{X}_i$, its rotation invariant gradient can be calculated in this local coordinate system by using pixel differences as follows:

$$Dx(\mathbf{X}_i) = I(\mathbf{X}_i^1) - I(\mathbf{X}_i^5) \tag{8}$$

$$Dy(\mathbf{X}_i) = I(\mathbf{X}_i^3) - I(\mathbf{X}_i^7) \tag{9}$$

where $\mathbf{X}_i^j, j = 1, 3, 5, 7$ are $\mathbf{X}_i$'s neighboring points along x-axis and y-axis in the local xy coordinate system as shown in Fig. 5, and $I(\mathbf{X}_i^j)$ stands for the intensity at $\mathbf{X}_i^j$. Then, the gradient magnitude $m(\mathbf{X}_i)$ and orientation $\theta(\mathbf{X}_i)$ can be computed as:

$$m(\mathbf{X}_i) = \sqrt{Dx(\mathbf{X}_i)^2 + Dy(\mathbf{X}_i)^2} \tag{10}$$

$$\theta(\mathbf{X}_i) = \tan^{-1}(Dy(\mathbf{X}_i)/Dx(\mathbf{X}_i)) \tag{11}$$

Note that $\theta(\mathbf{X}_i)$ is then transformed in the range of $[0, 2\pi)$ according to the signs of $Dx(\mathbf{X}_i)$ and $Dy(\mathbf{X}_i)$. In order to make an informative representation of $\mathbf{X}_i$, its gradient is transformed to a d dimensional vector, denoted by $F_G(\mathbf{X}_i) = (f_1^G, f_2^G, \cdots, f_d^G)$. To this end, $[0, 2\pi)$ is split into d equal bins as $\text{dir}_i = (2\pi/d) \times (i-1), i = 1, 2, \cdots, d$, then $\theta(\mathbf{X}_i)$ is linearly allocated to the two adjacent bins according to its distances to them weighted by $m(\mathbf{X}_i)$:

$$f_j^G = \begin{cases} m(\mathbf{X}_i)\frac{(2\pi/d - \alpha(\theta(\mathbf{X}_i), \text{dir}_j))}{2\pi/d}, & \text{if } \alpha(\theta(\mathbf{X}_i), \text{dir}_j) < 2\pi/d \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $\alpha(\theta(\mathbf{X}_i), \text{dir}_j)$ is the angle between $\theta(\mathbf{X}_i)$ and dir_j.

Note that in RIFT [10], it uses a similar way to calculate the rotation invariant gradient for its feature description. However, since it accumulates the cells of the histogram of gradient orientation in rings around the interest point to achieve rotation invariance, some spatial information is lost hence degrading its distinctiveness. Our proposed descriptors are constructed using the intensity order pooling, which encodes ordinal information into descriptors while maintaining descriptors' rotation invariance. Experiments (Section V) show that pooling by intensity orders is more informative than by rings.

3) Intensity-Based Local Feature: Our intensity-based feature is similar to CS-LBP [14], but it is calculated in a rotation invariant way, i.e., calculated in the locally rotation invariant coordinate system as shown in Fig. 5. Since our method aims to skip the rotation alignment of support region to achieve rotation invariance without resorting to an orientation for reference, such a modification is essential. For each sample point $\mathbf{X}_i$, suppose that $\mathbf{X}_i^j, j = 1, 2, \cdots, 2m$ are its $2m$ neighboring points regularly sampled along a circle. To obtain rotation invariance, $\mathbf{X}_i^1$ is located at the intersection of this circle and the positive x-axis in the local xy coordinate system. By comparing the intensities of opposite sample points, we get a m dimensional binary vector: $(\text{sign}(I(\mathbf{X}_i^{m+1}) - I(\mathbf{X}_i^1)), \text{sign}(I(\mathbf{X}_i^{m+2}) - I(\mathbf{X}_i^2)), \cdots, \text{sign}(I(\mathbf{X}_i^{m+m}) - I(\mathbf{X}_i^m)))$. Then a local feature $F_I(\mathbf{X}_i) = (f_1^I, f_2^I, \cdots, f_{2^m}^I)$ of $\mathbf{X}_i$ can be obtained by mapping the m dimensional binary vector into a 2^m dimensional vector:

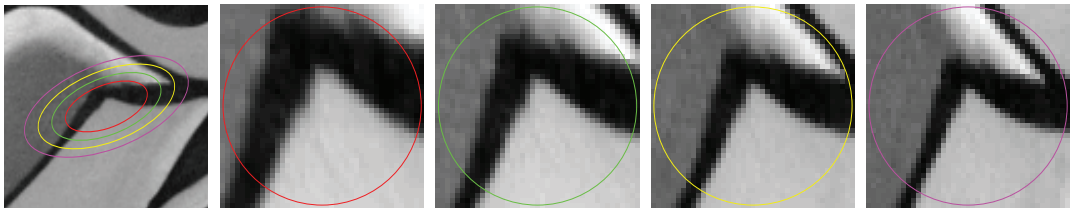
$$f_j^I = \begin{cases} 1, \text{if } \sum_{k=1}^{m} \text{sign}(I(\mathbf{X}_i^{k+m}) - I(\mathbf{X}_i^k)) \times 2^{k-1} = (j-1) \\ 0, \qquad\qquad \text{otherwise} \end{cases}, \text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, \text{otherwise} \end{cases} \tag{13}$$

Note that $F_I(\mathbf{X}_i)$ has only one element to be 1, all the others are 0.

D. Local Image Descriptor Construction

As our experiments show, one single support region is not enough to distinguish incorrect matches from correct ones in general. Two non-corresponding interest points may accidently have similar appearances in a certain local region. However, it is less likely that two non-corresponding interest points have similar appearances in several local regions of different sizes. In contrast, two corresponding interest points should have similar appearances in a local region of any size, although some small differences may exist due to localization error of interest point and region detection. That is to say, using multiple support regions, one can handle the mismatching problem better than using a single support region. Therefore, we utilize multiple support regions to construct our descriptors.

As shown in Fig. 6, we choose support regions as the N nested regions centered at the interest point with an equal increment of radius. Suppose that the detected region is denoted by $\mathbf{A} \in \Re^{2\times 2}$. It is used as the minimal support region and so other support regions are defined as $\mathbf{A}_i = \frac{1}{r_i^2}\mathbf{A}, i = 1, 2, \cdots, N$ in which r_i indicates the size of the ith support region. In this



(a) (b) (c) (d) (e)

Fig. 6. The selection of 4 support regions and their normalization. Each color corresponds to a boundary of one support region and all the support regions are normalized to a circular region with unified radius. (a) shows the selected support regions and (b)-(e) are the normalized regions.

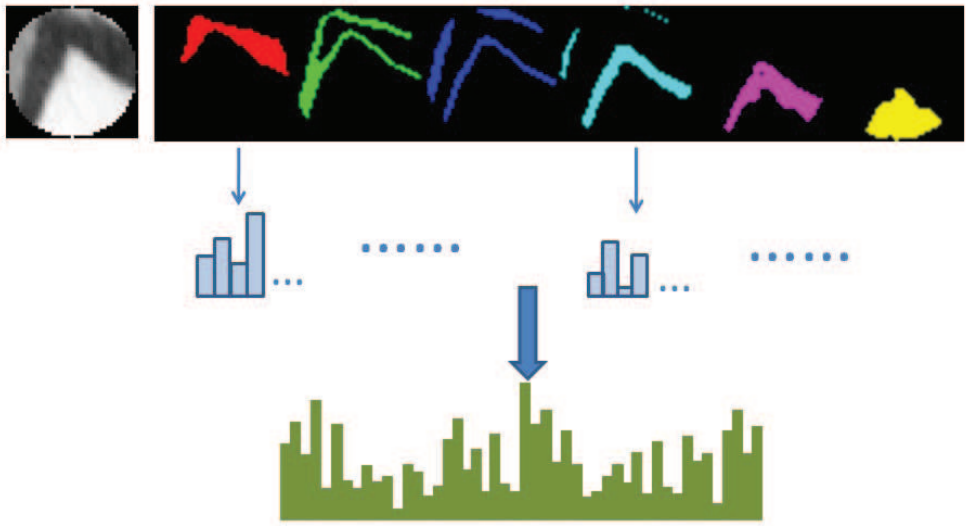


Fig. 7. The procedure of pooling local features based on their intensity orders for a support region. See text for details.

paper, we set $r_i = 1 + 0.5 \times (i - 1)$ such that the support regions have an equal increment of radius as plotted in Fig. 6(a). In each support region, the rotation invariant local features of all the sample points are then pooled by their intensity orders. As shown in Fig. 7, first they are pooled together to form a vector in each partition which is obtained based on the intensity orders of sample points and then the accumulated vectors of different partitions are concatenated together to represent this support region. We denote it as $D(R) = (F(R_1), F(R_2), \cdots, F(R_k))$ and $F(R_i)$ is the accumulated vector of partition R_i, i.e.,

$$F(R_i) = \sum_{\mathbf{X} \in R_i} F_G(\mathbf{X}) \tag{14}$$

if the gradient-based feature is used (the constructed descriptor is then called MROGH) or

$$F(R_i) = \sum_{\mathbf{X} \in R_i} F_I(\mathbf{X}) \tag{15}$$

if the intensity-based feature is used (the constructed descriptor is then called MRRID). Finally, all the vectors calculated from the N support regions are concatenated together to form our final descriptor: $\{D_1 D_2 \cdots D_N\}$. Fig. 8 gives an overview of our proposed method.

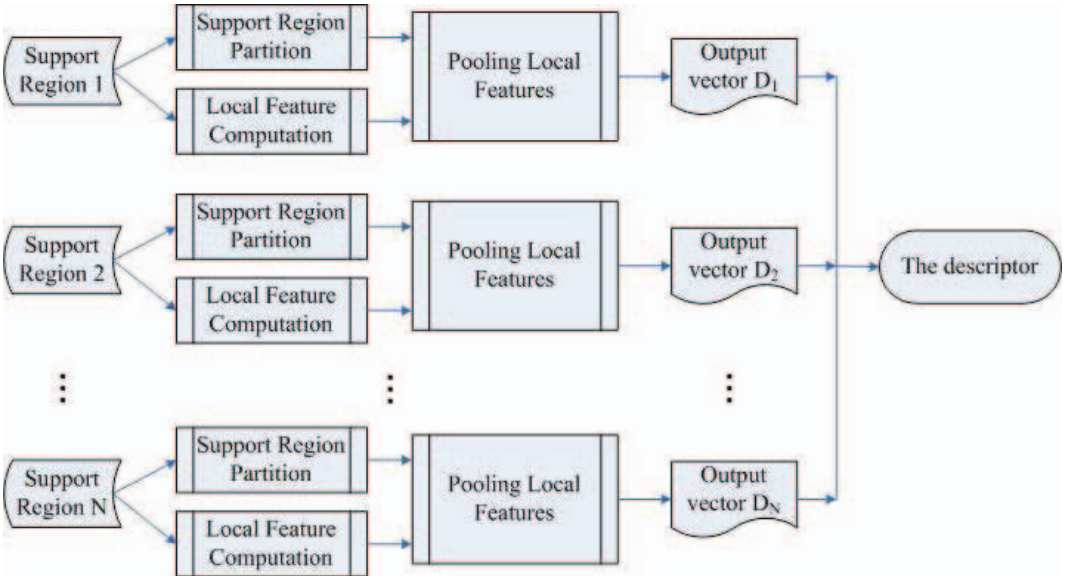


Fig. 8. The workflow of our proposed method.

E. Properties

Our proposed two descriptors (MROGH and MRRID) have the following characteristics making them both distinctive and robust to many image transformations.

(1) Local features are calculated in a rotation invariant way without resorting to an estimated orientation for reference. Therefore, the resulting rotation invariance is more stable than that obtained by aligning the positive x-axis with an estimated orientation. The estimated orientation tends to be unreliable according to our experimental study (see Section III).

(2) By partitioning sample points in the support region according to their intensity orders, the obtained partitions are rotation invariant and so they are suitable for pooling rotation invariant local features. Unlike the ring-shaped partitions which lose some spatial information, the intensity

order based partitions can encode ordinal information into descriptor. Therefore, the proposed descriptors could have higher discriminative ability.

(3) Since intensity orders are invariant to monotonic intensity changes, our proposed descriptors provide a higher degree of illumination invariance, not merely to the linear illumination change. Thus they can deal with large illumination changes, especially for MRRID, it has much better results than MROGH and other evaluated descriptors when matching images exhibit large illumination changes (see Section V-C.2). This is because for MRRID, not only its feature pooling scheme is based on intensity orders, its local feature is also based on the relative intensity relationship of sample points.

(4) The proposed descriptors are constructed on the basis of multiple support regions, further enhancing their discriminative ability. By utilizing multiple support regions, it also avoids the problem of selecting an optimal region size to construct descriptor for a detected interest region to some extent.

V. EXPERIMENTS

A. Parameters Evaluation

There are several parameters in the proposed descriptors: the number of spatial partitions k, the number of support regions N, the number of orientation bins d, and the number of binary codes m. As listed in Table I, MROGH and MRRID share two parameters: the number of spatial partitions and the number of support regions, while the number of orientation bins is needed in MROGH and the number of binary codes is needed in MRRID. In order to evaluate their influences on the performance of the proposed descriptors, we conducted image matching experiments on 142 pairs of images[2] with different parameter settings as listed in Tab. I. These 142 image pairs are mainly selected from the dataset of zoom and rotation transformations. Note that they do not contain image pairs in the standard Oxford dataset [44] because those image pairs are used for the descriptors evaluation in the later stage.

Fig. 9 shows the *average recall* vs. *average 1-precision* curves of MROGH and MRRID with different parameter settings. The definition of a correct match and a correspondence is the same as [22] which is determined with overlap error [9]. The matching strategy used here is the nearest

[2]They are real images downloaded from [43]. The ground truth homography is supplied along with the image pair.

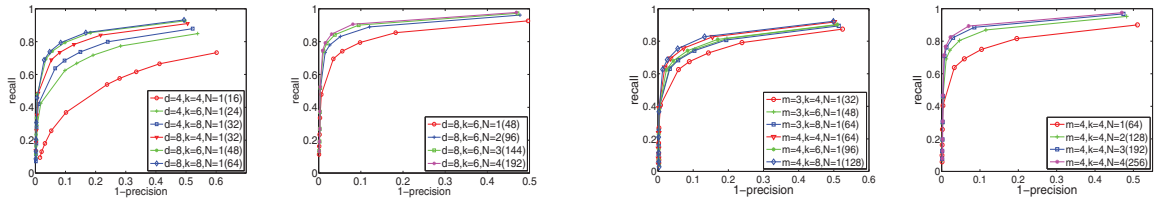
TABLE I

PARAMETERS OF OUR PROPOSED DESCRIPTORS

	denotation	parameter settings	description
MROGH	k	4,6,8	number of spatial partitions
	d	4,8	number of orientation bins
	N	1,2,3,4	number of support regions
MRRID	k	4,6,8	number of spatial partitions
	m	3,4	number of binary codes
	N	1,2,3,4	number of support regions

neighbor distance ratio [22]. In the remaining experiments, we used the same definitions of a match, a correct match and a correspondence.

It can be seen from Fig. 9 that the performances of MROGH and MRRID are improved when the number of support regions is increased. When a single support region is used, MROGH performs the best with the parameter setting of 'd=8,k=8', closely followed by the setting of 'd=8,k=6'. While MRRID performs the best with the parameter setting of 'm=4,k=8', followed by the setting of 'm=4,k=4'. Taking into consideration of performance and complexity (dimension), we use the parameter setting of 'd=8,k=6,N=4' for MROGH and 'm=4,k=4,N=4' for MRRID in the remaining experiments. Thus the dimensionality is 192 for MROGH and 256 for MRRID.



(a) Gradient-Based Descriptor (MROGH) (b) Intensity-Based Descriptor (MRRID)

Fig. 9. The average performance of the proposed descriptors with different parameter settings.

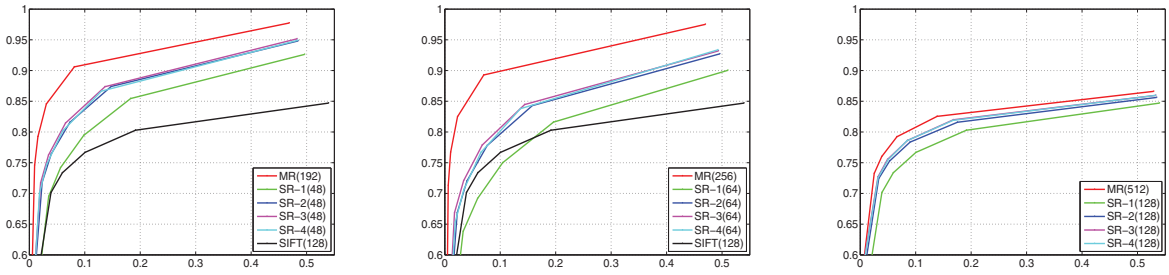
B. Multi-Support Regions vs. Single Support Region

This experiment aims to show the superiority of using multi-support regions over a single support region. We used the same dataset as in the experiment of parameters evaluation (Section V-A). Since four support regions are used in our method, for each of the proposed descrip-

tors (MROGH or MRRID), we can obtain five descriptors for one interest point: **SR-i** denotes the descriptor calculated from the ith support region and **MR** is the descriptor concatenating **SR-1**, **SR-2**, **SR-3**, **SR-4**. For each image pair, we respectively calculated these five descriptors to perform point matching and obtained the *average recall* vs. *average 1-precision* curves as before. The comparative results are shown in Fig. 10, in which Fig. 10(a) is the results of MROGH while Fig. 10(b) shows the results of MRRID. It can be found that **SR-2**, **SR-3**, **SR-4** perform better than **SR-1**, but **SR-2**, **SR-3**, **SR-4** perform almost as well. Thus when using a single support region, the performance of descriptor improves when the size of support region increases, but the effects become negligible after reaching a certain size. Fig. 10(a) and Fig. 10(b) demonstrate that by combining multiple support regions for descriptor construction, the performance of MROGH and MRRID improves significantly over the best performance of using a single support region. This indicates that using multiple support regions improves the discriminative ability of our proposed descriptors. For comparison the performance of SIFT is plotted too. It is worth noting that even a single region based MROGH (**SR-1**) is superior to SIFT while the dimension of **SR-1** is much lower (48 vs. 128), demonstrating the effectiveness of the proposed method. Fig. 10(c) shows the results of calculating SIFT with multiple support regions. Although the performance of SIFT is improved with multiple support regions, the improvement is much less significant than MROGH and MRRID. This is because the error and uncertainty induced by incorrect orientation estimation hamper its further improvement. In contrast, MROGH and MRRID do not have such limitations as they do not involve any orientation estimation.

C. Image Matching

To evaluate the performance of the proposed descriptors, we conducted extensive experiments on image matching. We followed the evaluation procedure proposed by Mikolajczyk and Schmid [22]. The codes for evaluation were downloaded from their website [44]. Three other local descriptors were evaluated in our experiments for comparison: RIFT, SIFT and DAISY. In our experiments, SIFT was downloaded from the Oxford University website [44] which is the same as the one used by Mikolajczyk and Schmid [22] while RIFT and DAISY (The **T1-8-2r8s** configuration [34] of DAISY is used.) are our own implementations. Note that the downloaded SIFT only uses one dominant orientation (the orientation corresponds to the largest peak of the orientation histogram) for each interest point. In order to be consistent with it, our implementation

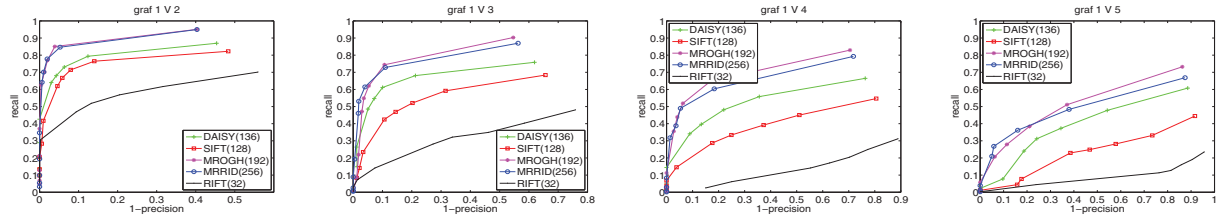


(a) MROGH (b) MRRID (c) SIFT

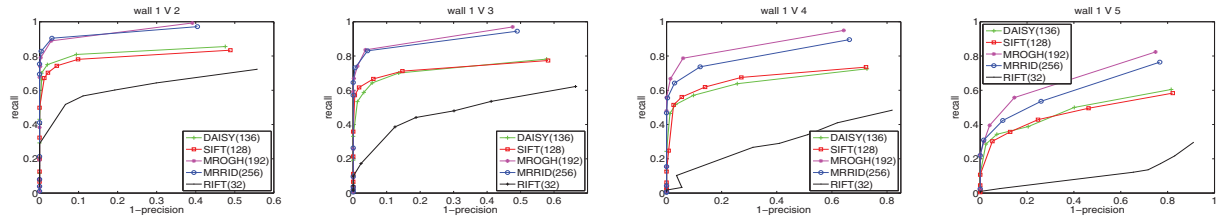
Fig. 10. Performance comparison between multi-support regions and a single support region when using (a) MROGH, (b) MRRID and (c) SIFT respectively. **SR-i** is the result using the ith support region and **MR** is the result using multi-support regions. The performances of MROGH and MRRID are significantly improved when using multi-support regions, while the improvement of SIFT is much less significant.

of DAISY also uses one dominant orientation for each interest point. Therefore, all the evaluated methods (DAISY, SIFT, RIFT, MROGH and MRRID) have the same number of features. So the obtained numbers of correspondences are the same too. In our experiments (Intel Core2 CPU 1.86GHz), the time of constructing a descriptor for a feature point is: 1.6ms for RIFT, 12.3ms for DAISY, 4.6ms for SIFT, 8.3ms for MROGH, 1.9 ms for MROGH with a single region, 18.5ms for MRRID and 4.8ms for MRRID with a single region.

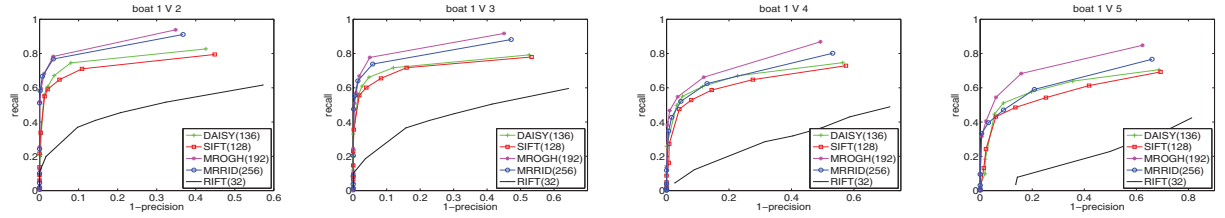
1) Performance on the Oxford Dataset: The proposed descriptors were evaluated on the standard Oxford dataset, in which image pairs are under various image transformations, including viewpoint change, scale and rotation changes, image blur, JPEG compression and illumination change. They are shown in Fig. 2 in the supplemental material. In each row, the images from the 2nd to the 5th columns are matched to the 1st image separately. As in [22], we evaluated the performance of descriptors using both Hessian-Affine and Harris-Affine detectors [9]. Due to paper length limit, we only give the results with Hessian-Affine regions in Fig. 11. Please see Fig. 3 in the supplemental material for the results with Harris-Affine regions. It can be seen from these results that although the performance of each descriptor varied with different feature detectors (Hessian-Affine or Harris-Affine in this experiment), the relative performance among different descriptors is consistent with different feature detectors. The results show that MROGH and MRRID outperform the other evaluated local descriptors in all the tested cases. Such a good performance of MROGH and MRRID may be attributed to their well-designed



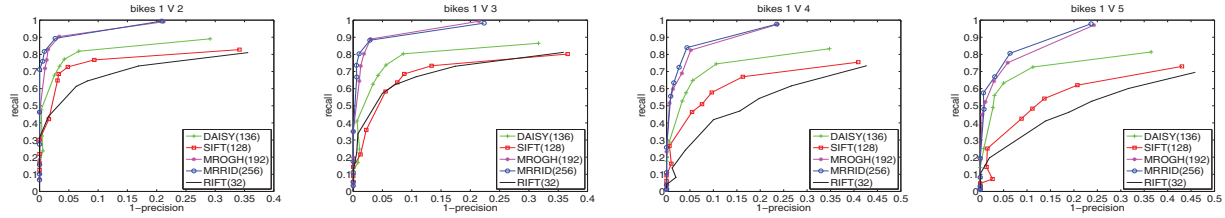
(a) viewpoint changes (graf)



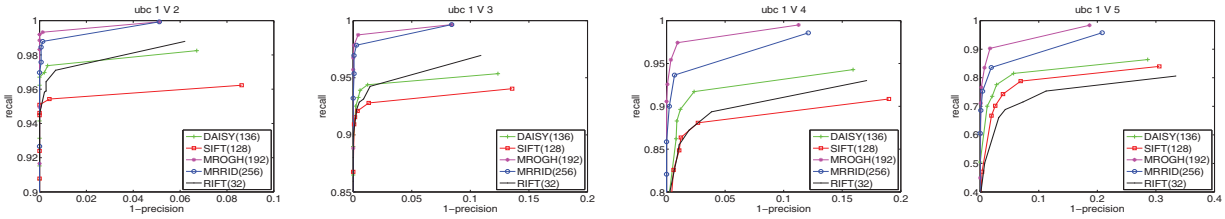
(b) viewpoint changes (wall)



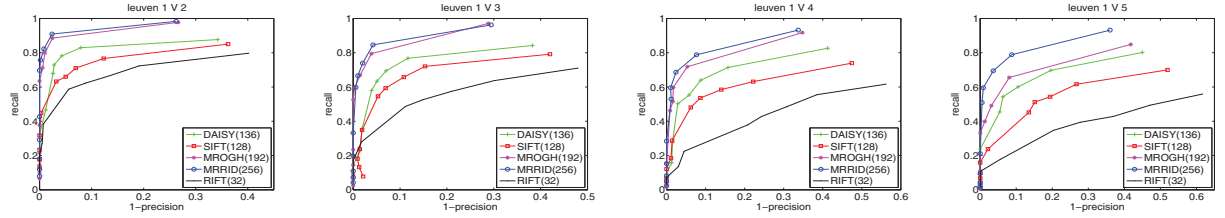
(c) rotation and scale changes (boat)



(d) image blur (bikes)



(e) JPEG compression (ubc)



(f) illumination changes (leuven)

Fig. 11. Experimental results under various image transformations in the Oxford dataset for Hessian-Affine Region.

properties. They not only use multiple support regions to improve discriminative ability, but also combine local features (histogram of gradient orientation in MROGH and CS-LBP in MRRID) with ordinal information. Moreover, they are rotation invariant without relying on a reference orientation, further improving their robustness. Since MROGH uses a similar local feature to the one used in RIFT, the significant performance improvement of MROGH over RIFT demonstrates the effectiveness and advantage of our proposed feature pooling scheme, i.e. pooling by intensity orders is more informative than by rings. In most cases, MROGH performs better than MRRID. When images exhibit blur or illumination changes, MRRID is better, especially when encountering large illumination changes.

2) Performance on Images with Large Illumination Changes: In order to investigate the robustness of the proposed descriptors to monotonic intensity changes, we conducted image matching on three additional image pairs with large illumination changes as shown in the top of Fig. 12. They were captured by changing the exposure time of camera, the matching results are shown below the tested images. The Hessian-Affine detector was used for interest region detection. It can be seen from Fig. 12 that MRRID performs the best in such cases and achieves much higher performance than the other evaluated descriptors due to its invariance to monotonic intensity changes. The more severe the brightness change is, the larger the performance gap between MRRID and the other evaluated descriptors is.

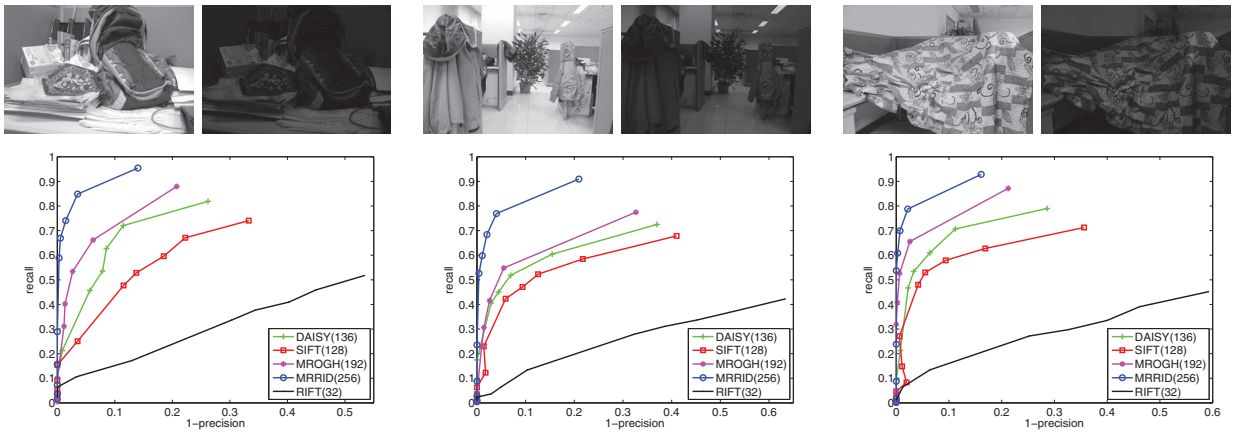
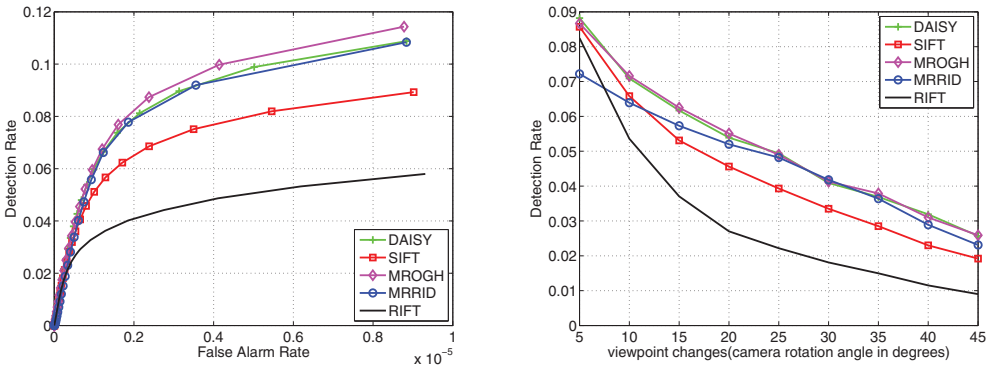


Fig. 12. Matching results of images with large illumination changes, which were captured by changing the exposure time of camera. The more severe the brightness change is, the larger the performance gap between MRRID and the other evaluated descriptors is.

3) Performance Evaluation based on 3D Objects: Moreels and Perona [45] evaluated different combinations of feature detectors and descriptors based on 3D objects. We evaluated the proposed descriptors following their work[3]. The Hessian-Affine detector was used for feature extraction since it was reported with the best results combined with SIFT in [45]. As in [45], interest point matching was conducted on a database containing both the target features and a large amount of features from unrelated images so as to mimic the process of image retrieval/object recognition. 10^5 features were randomly chosen from 500 unrelated images obtained from Google by typing 'things'. Please refer to [45] for more details about the dataset and experimental setup.



(a) (b)

Fig. 13. Experimental results on the dataset of 3D objects. (a) Detection Rate vs. False Alarm Rate; (b) Detection Rate vs. Viewpoint Changes at false alarm rate of 10^{-6}.

Fig. 13 shows the comparative results. In Fig. 13(a), the ROC curves of 'detection rate vs. false alarm rate' were obtained by varying the threshold of nearest neighbor distance ratio which defines a match. The detection rate is the number of detections against the number of tested matches, while the false alarm rate is the number of false alarms divided by the number of tested matches. A tested match is classified as a non-match, a false alarm or a correct match (a detection) according to the distance between their descriptors and whether the geometric constraints are satisfied or not [45]. In Fig. 13(b), it shows the detection rate as a function of the viewpoint changes at a fixed false alarm rate. The false alarm rate 10^{-6} implies that one false alarm

[3]The dataset was downloaded from http://www.vision.caltech.edu/pmoreels/Datasets/TurntableObjects. The downloadable dataset actually contains 77 different objects, so in our experiments it is 77 different objects used for evaluation not 100 as described in [45].

over 10 attempts in average since the false alarm rate is normalized by the number of database features (10^5). It can be seen from Fig. 13 that the proposed descriptors and DAISY outperform SIFT and RIFT, while MROGH is sightly better than DAISY.

4) Performance on the Patch Dataset: The proposed descriptors were also evaluated on the Patch Dataset [34], [46]. Since the matched patches in this dataset only have very small rotation, we varied the rotation of patches before constructing descriptors in order to mimic the real situation of image matching in which the rotation between the matched images is usually unknown. The matching results are shown in Fig 14. From Fig. 14(a) to Fig. 14(c) and Fig. 14(e) to Fig. 14(g), MROGH and MRRID outperform SIFT, DAISY and RIFT on these image patches taken from the same objects under different lighting and viewing conditions. Fig. 14(d) and Fig. 14(h) show the performance degradation of SIFT and DAISY when they need to assign a reference orientation. OriDAISY and OriSIFT refer to the method that construct descriptors on the patches without rotation, while DAISY and SIFT refer to the method that construct descriptors on the patches according to a reference orientation. It can be seen from Fig. 14(d) and Fig. 14(h) that the matching performances of SIFT and DAISY degrade when the orientation is unknown (In this case, a reference orientation is required for descriptor construction.).

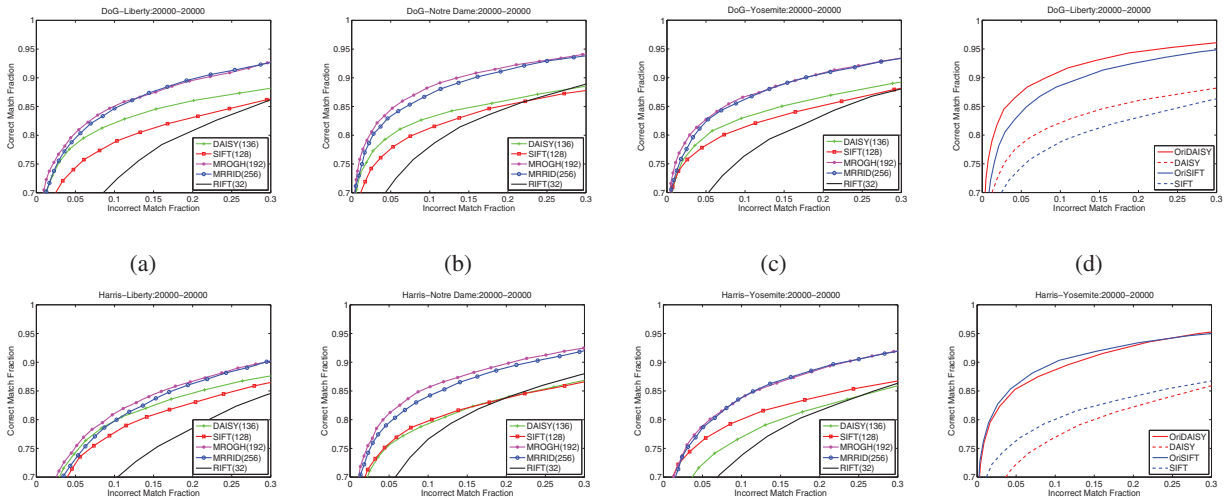


Fig. 14. ROC curves of 5 different local descriptors on the Liberty, Notre Dame and Yosemite dataset (20000 matches and 20000 non-matches are used). (a)-(c) are the results on DoG interest point and (e)-(g) are the results on multi-scale Harris interest point. (d) and (h) show the performance degradation of SIFT and DAISY when they need to assign a reference orientation for rotation invariance.

D. Object Recognition

We further conducted experiments on object recognition to show the effectiveness of the proposed descriptors. In our experiments, the similarity between images is defined as follows. Suppose that I_1 and I_2 are two images, and $\{f_1^1, f_2^1, \cdots, f_m^1\}$, $\{f_1^2, f_2^2, \cdots, f_n^2\}$ are two sets of feature descriptors extracted from I_1 and I_2. Then the similarity between I_1 and I_2 is defined as:

$$\text{Sim}(I_1, I_2) = \frac{\sum_{i,j} g(f_i^1, f_j^2)}{m \times n} \tag{16}$$

where

$$g(f_i^1, f_j^2) = \begin{cases} 1, & \text{if } \text{dist}(f_i^1, f_j^2) \leq T \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

in which $\text{dist}(f_i^1, f_j^2)$ is the Euclidean distance of descriptors, and T is a threshold which is tuned to give the best result for each evaluated descriptor. Such a definition of similarity only relies on the distinctiveness of the local descriptor.

Three tested datasets were downloaded from the Internet [47], [48]. The first dataset (named 53 Objects) contains 53 objects, each of which has 5 images taken from different viewpoints. The second dataset (named ZuBuD) contains 201 different buildings, and each building has 5 images taken from different viewpoints. The third dataset is the Kentucky dataset [48]. It contains 2550 objects, each of which has 4 images taken from different viewpoints. We used the first 4000 images from 1000 objects in our experiments. Some images in these datasets are shown in Fig. 4 in the supplemental material. For each image in the 53 Objects and ZuBuD datasets, we calculated its similarities to the remaining images in the dataset, and returned the top 4 images with the largest similarities. While for each image in the Kentucky dataset, we returned the top 3 images. We recorded the ratio of the number of returned correct images to the number of total returned images as the recognition accuracy. Table II gives the recognition results and some recognition examples are shown in Fig. 5 in the supplemental material. MROGH performs the best among all the evaluated descriptors. Since MRRID is primarily designed to deal with large illumination changes and the tested datasets mainly contain viewpoint changes, MRRID does not perform as good as MROGH. However, it is still better than SIFT and RIFT.

TABLE II

RECOGNITION RESULTS ON THE THREE TESTED DATASETS WITH DIFFERENT LOCAL DESCRIPTORS

53 Objects	Descriptor	RIFT	SIFT	DAISY	MRRID	MROGH
	Accuracy	37.0%	52.2%	61.2%	57.4%	**72.5%**
ZuBuD	Descriptor	RIFT	SIFT	DAISY	MRRID	MROGH
	Accuracy	66.8%	75.5%	83.1%	78.6%	**88.1%**
Kentucky	Descriptor	RIFT	SIFT	DAISY	MRRID	MROGH
	Accuracy	34.0%	48.2%	58.3%	57.5%	**74.0%**

VI. CONCLUSION

This paper presents a novel method for constructing interest region descriptors. The key idea is to pool rotation invariant local features based on intensity orders. It has the following important properties:

a. Unlike the undertaking in many popular descriptors where an estimated orientation is assigned to the descriptor for obtaining rotation invariance, our proposed descriptors are inherently rotation invariant thanks to a rotation invariant way of local feature computation and an adaptive feature pooling scheme.

b. The proposed pooling scheme partitions the sample points into several groups based on their intensity orders, rather than their geometric locations. Pooling by intensity orders makes our two descriptors rotation invariant without resorting to the conventional practice of orientation estimation, which is an error-prone process according to our experimental analysis. In addition, the intensity order pooling scheme can encode ordinal information into descriptor.

c. Multiple support regions are used to further improve discriminative ability.

By pooling two different kinds of local features (gradient-based and intensity-based) based on intensity orders, this paper obtained two descriptors: MROGH and MRRID. The former combines information of intensity orders and gradient, while the latter is completely based on intensity orders which makes it particularly suitable to large illumination changes. Extensive experiments show that both MROGH and MRRID could achieve better performance than state-of-the-art descriptors.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China under the grant No.60835003 and No.61075038. Thanks for the anonymous suggestions by the reviewers. Thanks for the helpful discussions with Prof. Shiming Xiang and Prof. Michael Werman.

REFERENCES

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *International Conference on Computer Vision*, 2009, pp. 72–79.

[2] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[3] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building rome on a cloudless day," in *European Conference on Computer Vision*, 2010, pp. 368–381.

[4] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Transactions on Graphics (TOG)*, vol. 25, pp. 835–846, 2006.

[5] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski, "Reconstructing rome," *Computer*, vol. 43, pp. 40–47, 2010.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[8] R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, pp. 1–104, 2006.

[9] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *European Conference on Computer Vision-Part I*, 2002, pp. 128–142.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1265–1278, 2005.

[11] J. J. Koenderink and A. J. V. Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, pp. 367–375, 1987.

[12] A. Baumberg, "Reliable feature matching across widely separated views," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2000.

[13] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets, or how do i organize my holiday snaps?," in *European Conference on Computer Vision*, 2002, pp. 414–431.

[14] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognition*, vol. 42, pp. 425–436, 2009.

[15] E. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 184–190.

[16] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[17] B. Fan, F. Wu, and Z. Hu, "Aggregating gradient distributions into intensity orders: A novel local image descriptor," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

[18] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Fourth Alvey Vision Conf.*, 1988, pp. 147–151.

[19] J. Matas, O. Chum, M. Urba, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *British Machine Vision Conference*, 2002, pp. 384–396.

[20] T. Tuytelaars and L. V. Gool, "Matching widely separated views based on affine invariant regions," *International Journal of Computer Vision*, vol. 59, no. 1, pp. 61–85, 2004.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[22] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[23] S.Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *NIPS*, 2001, pp. 831–837.

[24] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. I, 2004, pp. 511–517.

[25] R. Gupta and A. Mittal, "SMD: A locally stable monotonic change invariant feature descriptor," in *European Conference on Computer Vision*, 2008, pp. 265–277.

[26] J. Chen, S. Shan, G. Zhao, X. Chen, W. Gao, and M. Pietikainen, "A robust descriptor based on weber's law," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[27] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[28] F. Tang, S. H. Lim, N. L. Change, and H. Tao, "A novel feature descriptor invariant to complex brightness changes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2631–2638.

[29] R. Gupta, H. Patil, and A. Mittal, "Robust order-based methods for feature description," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 334–341.

[30] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 891–906, 1991.

[31] L. V. Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," in *European Conference on Computer Vision*, 1996, pp. 642–651.

[32] I. T. Joliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[33] S. Winder and M. Brown, "Learning local image descriptors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[34] S. Winder, G. Hua, and M. Brown, "Picking the best DAISY," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 178–185.

[35] A. C. Berg and J. Malik, "Geometric blur for template matching," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, pp. 607–614.

[36] P.-E. Forssen and D. G. Lowe, "Shape descriptor for maximally stable extremal regions," in *International Conference on Computer Vision*, 2007.

[37] A. K. Jain, *Fundamentals of Digital Image Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[38] A. Mittal and V. Ramesh, "An intensity-augmented ordinal measure for visual correspondence," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. I, 2006, pp. 849–856.

[39] R. Gupta and A. Mittal, "Illumination and affine- invariant point matching using an ordinal approach," in *International Conference on Computer Vision*, 2007.

[40] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, pp. 51–59, 1996.

[41] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Improving local descriptors by embedding global and local spatial information," in *European Conference on Computer Vision*, 2010, pp. 736–749.

[42] H. Cheng, Z. Liu, N. Zheng, and J. Yang, "A deformable local image descriptor," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[43] http://lear.inrialpes.fr/people/mikolajczyk/.

[44] http://www.robots.ox.ac.uk/˜vgg/research/affine/.

[45] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 263–284, 2007.

[46] http://www.cs.ubc.ca/˜mbrown/patchdata/patchdata.html.

[47] http://www.vision.ee.ethz.ch/datasets/index.en.htm.

[48] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2161–2168.

Bin Fan received his B.S. degree in Automation from Beijing University of Chemical Technology, Beijing, China, in 2006, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. In July 2011, he joined the National Laboratory of Pattern Recognition, where he works as an assistant professor. His research interest covers Image Matching and Feature Extraction.

Fuchao Wu received his B.S. degree in Mathematics from Anqing Teacher's College, Anqing, China, in 1982. From 1984 to 1994, he acted first as a lecturer then as an associate professor in Anqing Teacher's College. From 1995 to 2000, he acted as a professor in Anhui University, Hefei, China. Since 2000, he has been with the Institute of Automation, Chinese Academy of Sciences, where he is now a professor. His research interests are in Computer Vision, which include 3D Reconstruction, Active Vision, and Image Based Modeling and Rendering.

Zhanyi Hu received his B.S. degree in Automation from the North China University of Technology, Beijing, China, in 1985, and the Ph.D. degree in Computer Vision from the University of Liege, Belgium, in Jan. 1993. Since 1993, he has been with the Institute of Automation, Chinese Academy of Sciences, where he is now a professor. His research interests are in Robot Vision, which include Camera Calibration and 3D Reconstruction, Vision Guided Robot Navigation. He was the local chair of ICCV'2005, an area chair of ACCV'2009, and is a PC chair of ACCV'2012.