A Zernike Moment Phase-Based Descriptor for Local Image Representation and Matching

Zen Chen and Shu-Kuo Sun

***Abstract*—A local image descriptor robust to the common photometric transformations (blur, illumination, noise, and JPEG compression) and geometric transformations (rotation, scaling, translation, and viewpoint) is crucial to many image understanding and computer vision applications. In this paper, the representation and matching power of region descriptors are to be evaluated. A common set of elliptical interest regions is used to evaluate the performance. The elliptical regions are further normalized to be circular with a fixed size. The normalized circular regions will become affine invariant up to a rotational ambiguity. Here, a new distinctive image descriptor to represent the normalized region is proposed, which primarily comprises the Zernike moment (ZM) phase information. An accurate and robust estimation of the rotation angle between a pair of normalized regions is then described and used to measure the similarity between two matching regions. The discriminative power of the new ZM phase descriptor is compared with five major existing region descriptors (SIFT, GLOH, PCA-SIFT, complex moments, and steerable filters) based on the precision-recall criterion. The experimental results, involving more than 15 million region pairs, indicate the proposed ZM phase descriptor has, generally speaking, the best performance under the common photometric and geometric transformations. Both quantitative and qualitative analyses on the descriptor performances are given to account for the performance discrepancy. First, the key factor for its striking performance is due to the fact that the ZM phase has accurate estimation accuracy of the rotation angle between two matching regions. Second, the feature dimensionality and feature orthogonality also affect the descriptor performance. Third, the ZM phase is more robust under the nonuniform image intensity fluctuation. Finally, a time complexity analysis is provided.**

***Index Terms*—Geometric and photometric transformations, image representation and matching, performance evaluation, phase and magnitude components, precision and recall, region descriptors, Zernike moments (ZM).**

I. Introduction

LOCAL features robust to common photometric transformations (blur, illumination, noise, and JPEG compression) and geometric transformations (rotation, scale, translation, and viewpoint) are crucial to most image understanding and computer vision applications including image matching, camera calibration, texture classification, and image retrieval, etc. [1]–[5]. The processing of local features involves three tasks: feature detection, feature description, and feature matching. The local features belong to an interest point (keypoint) or an interest region. Since a single image point carries little information, an interest point must be associated with its surrounding image patch. From this image patch, a second moment matrix of image intensities reveals the characteristic structure of the local image region. The keypoint detectors such as Harris corner detector [6] and the SIFT detector [7], which is based on the difference of Gaussians (DOG), utilize a circular window to search for a possible location of a keypoint. However, the image content in the circular window is not robust to affine deformations. Recently, a number of local feature detectors using a local elliptical window have been investigated. Matas *et al.* [5] presented a maximally stable extremal region (MSER) detector. Tuytelaars and Van Gool [8] developed an edge-based region (EBR) detector as well as an image-based (IBR) region detector. Mikolajczyk and Schmid [9] proposed Harris-Affine and Hessian-Affine detectors. The performances of the existing region detectors were evaluated [11], indicating MSER detector and Hessian-Affine detector are the two best.

After the regions of interest are detected, a region descriptor is needed for region representation. In the descriptor construction, the detected ellipse-shaped region is first normalized to a circular patch of a fixed size (typically, 41 × 41 pixels). The normalized circular patch can be shown to be affine invariant up to a rotational ambiguity [10], [33]. A good feature descriptor should have a great discriminative power. Five major types of existing descriptors are to be briefly reviewed in the next section to explore their capability for image representation.

After the region descriptor is determined, a matching function is defined to measure the similarity between regions extracted from different images of the same scene. The merits of various region detectors, coupled with their own region descriptors, are often judged based on the ROC (receiver operating characteristic) curve or the PR (precision-recall) curve.

In this paper, a new descriptor, called the Zernike moment phase-based descriptor (or ZM phase in short), is proposed. The phase information of a signal is more informative than the magnitude information during signal reconstruction, as demonstrated by Oppenheim [34]. The robustness of local phase information for measuring image velocity and binocular disparity was studied in [35], [36]. Recently, outputs of complex-valued steerable filter quadrature pairs are taken as the separate feature elements for the design of a local image descriptor [37], [38], instead of combining the magnitudes of the quadrature pair into a single feature element, as done in [12]. They empirically showed that their individual local descriptors have better performance

Manuscript received October 16, 2008; revised August 12, 2009. First published September 22, 2009; current version published December 16, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sabine Susstrunk.

The authors are with the Department of Computer Science, National Chiao Tung University, 1001 University Road, Hsinchu, Taiwan, R.O.C. (e-mail: zchen@cs.nctu.edu.tw; sksun@csie.nctu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2032890

than the gradient-based SIFT descriptor or differential invariants under the affine geometric deformation and lighting variation. However, the feature vector containing the separate steerable filter quadrature pair outputs is not an orthogonal vector itself. If the orthogonal descriptor is used instead, the features are uncorrelated and more informative. So we shall seek a genuine orthogonal feature vector to derive a novel local descriptor with a higher descriptive power.

The discriminative power of the new ZM phase descriptor is compared with five other major region descriptors based on the precision-recall criterion using the set of test images given in [12] plus some new images. To match the region pairs, a new matching function based on the ZM phase information is defined. For performance evaluation, important system parameters are taken into consideration, which include 1) region scene types, 2) region descriptor types, 3) region detector types, 4) region overlap error, and 5) transformation types. The experimental results, involving more than 15 million region pairs, indicate the proposed ZM phase has the best overall performance. Both quantitative and qualitative analyses on the descriptor performances are provided to account for the performance discrepancy.

Our main contributions include the following.

1) Design a new region descriptor and a new matching function based mainly on Zernike moment (ZM) phase information.
2) Propose an accurate estimation of the rotation angle between two matching regions.
3) Show the proposed ZM phase descriptor has the better overall performance compared to the five popular descriptors.

The paper is organized as follows. Section II reviews the five major types of region descriptors. Section III introduces the Zernike moment (ZM) transformation and the ZM basis filters. Section IV proposes the ZM phase descriptor along with a matching function, and discusses the discriminative powers of the ZM magnitude components and the ZM phase components. In Section V, the discriminative power of the new descriptor is compared with five existing region descriptors based on the precision-recall criterion, while taking important system parameters into consideration. In Section VI, quantitative and qualitative analyses on the descriptors are provided to account for the descriptor performance discrepancy. The conclusion is given in the last section.

II. Review of the Major Region Descriptors

Here, a brief introduction of five major classes of the existing descriptors is given to explore their strengths and weakness in order to compare them with the proposed ZM phase-based descriptor. Excellent reviews on the existing descriptors can be found in [12] and [13].

1) Filter-based Descriptors:

This class of descriptors includes steerable filters [14] and Gabor filters [15]. The steerable filter descriptor uses quadrature pairs of derivatives of Gaussian and their Hilbert transforms to synthesize any filter of a given frequency with arbitrary phase. On the other hand, the Gabor transform uses a number of Gabor filters tuned to various frequencies and orientations to represent the image patterns. Both the steerable filter and the Gabor filter descriptors need to seek a dominant orientation for image rotation alignment. If the reference and transformed descriptor feature vectors are not aligned well, their matching score will be poor. Besides, these descriptors are not totally orthogonal and their feature vector dimensions are generally low, so their discriminative powers are limited.

2) Moment-based descriptors:

The first class of moment-based descriptors is the geometric (or regular) moments. The $(p+q)$ order moment of an intensity or gradient image $f(x,y)$ is defined as follows:

$$m_{pq}=\sum_x\sum_y x^p y^q f(x,y), \quad p,q=0,1,2\ldots$$

Based on the geometric moments, a set of moment invariants can be derived from the nonlinear combinations of geometric moments to achieve affine invariance [16], [32]. The main problem with the geometric moments is that it is difficult to derive a sufficient number of invariants to describe complex shapes. Moreover, the higher-order moments are more sensitive to image noise than the lower-order moments. Therefore, the geometric moment invariants are usually suitable only for describing simple images [17].

The second class of moment-based descriptors is the complex moments of the form $K_{mn}(x,y)=\sum_x\sum_y(x+iy)^m(x-iy)^n f(x,y)$, where $f(x,y)$ is an image intensity function [18], [19]. Any rotation of the image changes the phases of the complex moments, but not the magnitudes. That is, the magnitudes of the filter responses are rotational invariant. There are 16 filters, defined by $m+n\le 6$ and $n\le m$, available for image patch description. This low-dimensional rotational invariant descriptor generally has a poor discriminative performance [12].

3) Distribution-based descriptors:

This class of descriptors includes SIFT [7], GLOH [12], PCA-SIFT [22], spin images and RIFT descriptors [3]. They use the distributions of the image content to represent the features of the image region.

The SIFT descriptor is represented by a 3-D histogram of gradient locations and orientations. The histogram of the gradient orientations is quantized in 8 bins and the region is partitioned into a 4 × 4 location grid, resulting in a feature vector of dimension 128. Although the gradient histogram provides stability against deformations of the image pattern, the grid partition of the measurement region has the boundary effect problem. Gaussian smoothing and tri-linear interpolation can be called to alleviate this problem. More importantly, SIFT requires an accurate dominant (gradient) orientation for image rotation alignment.

The PCA-SIFT descriptor is a dimension-reduced version of SIFT (dimension reduced from 3042 to 36 or lower) based on an eigenspace obtained by applying PCA to a collection of 21,000 image patches. On the other hand, the GLOH descriptor is also an extension of the SIFT descriptor. Instead of sampling gradient orientations in a rectangular grid, GLOH is defined in a log-polar location grid with 17 location bins. These location bins, together with 16 gradient orientation bins, form a feature vector of dimension 272. With PCA, the feature dimension is reduced to 128 based on a training data set of 47,000 image patches.

The SIFT and its variants depend on a dominant orientation of the normalized patch to achieve the rotation invariance. However, according to the experience of Lazebnik *et al.* reported in [3], the dominant orientation estimation tends to be unreliable, especially for normalized Laplacian regions in which strong edges at the center are often not available.

4) Derivative-based descriptors:

This type of descriptors uses local derivatives, called “local jets”, to construct the differential invariants that are rotationally invariant [23]. Schmid and Mohr [2] derive a set of differential invariants in terms of polynomials of local derivatives up to the third order for image retrieval. The derivative-based descriptors face with some problems: (a) the dimension of the rotationally invariant differential invariants is generally low [12], and (b) the differential invariants are often sensitive to image blur or image noise if smoothing operation is not used beforehand. (The steerable filters can be also classified as a derivative-based descriptor.)

5) Others:

Besides the above basic descriptor types, there are other extended descriptors including (i) color-based descriptors [21] which utilizes the color information for feature representation, (ii) textons [3], which are based on the responses of a texture image to a filter bank, can categorize the large-scaled texture images. In this paper, only the basic descriptors of the first four classes are concerned.

III. Fundamentals of Zernike Moments

Zernike moments (ZMs) have been used in object recognition and image analysis regardless of variations in position, size and orientation [20], [24]–[28]. Basically, the Zernike moments are the extension of the geometric moments by replacing the conventional transform kernel $x^m y^n$ with orthogonal Zernike polynomials. The relationships between the Zernike moments and geometric moments can be established [39]. The ZM coefficients are the outputs of the expansion of an image function into a complete orthogonal set of complex basis functions $\{V_{nm}(\rho,\theta)\}$. Teh and Chin [20] show that among many moment-based shape descriptors, Zernike moment magnitude components are rotationally invariant and most suitable for shape description.

The Zernike basis function $V_{nm}(\rho,\theta)$ with order n and repetition m is defined over a unit circle in the polar coordinates as follows:

$$V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{jm\theta} \text{ for } \rho \le 1 \tag{1}$$

where $\{R_{nm}(\rho)\}$ is a radial polynomial in the form of

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s}. \tag{2}$$

Here, n is a non-negative integer and m is an integer satisfying the conditions: $n-|m|$ is even and $|m| \le n$.

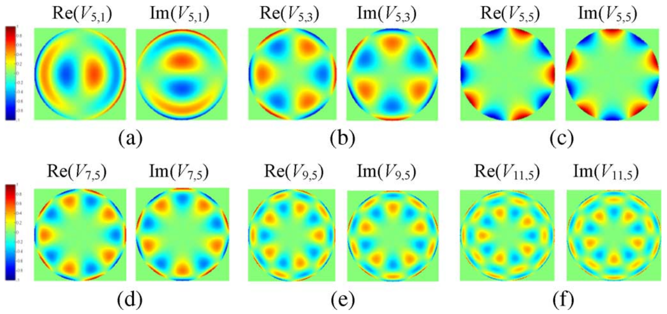


Fig. 1. Plots of the real part and imaginary part of $V_{nm}(\rho,\theta)$ for a fixed n: (a)$V_{5,1}$, (b) $V_{5,3}$, (c) $V_{5,5}$; and for a fixed $m(=5)$: (d) $V_{7,5}$, (e) $V_{9,5}$, and (f)$V_{11,5}$.

The set of basis functions $\{V_{nm}(\rho,\theta)\}$ is orthogonal, i.e.,

$$\int_0^{2\pi}\int_0^1 V^*_{nm}(\rho,\theta)V_{pq}(\rho,\theta)\rho d\rho d\theta = \frac{\pi}{n+1}\delta_{np}\delta_{mq}$$

$$\text{with } \delta_{ab} = \begin{cases} 1, & a=b \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The 2-D ZMs for a continuous image function $f(\rho,\theta)$ are represented by

$$\begin{aligned} Z_{nm} &= \frac{n+1}{\pi}\int_0^{2\pi}\int_0^1 f(\rho,\theta)V^*_{nm}(\rho,\theta)\rho d\rho d\theta \\ &= \frac{n+1}{\pi}\int_0^{2\pi} \mathrm{e}^{-jm\theta}\int_0^1 f(\rho,\theta)R_{nm}(\rho)\rho d\rho d\theta. \end{aligned} \tag{4}$$

For a digital image function, the 2-D ZMs are given as

$$Z_{nm} = \frac{n+1}{\pi}\sum\sum_{(\rho,\ \theta)\in\text{unit disk}} f(\rho,\theta)V^*_{nm}(\rho,\theta). \tag{5}$$

The Zernike moments can be viewed as the responses of the image function $f(\rho,\theta)$ to a set of quadrature-pair filters $\{V_{nm}(\rho,\theta)\}$. To this end, Fig. 1 depicts some examples of $V_{nm}(\rho,\theta)$. Notice that the real and imaginary functions of each basis function $V_{nm}(\rho,\theta)$ are out of phase by $\pi/2$; namely, they form quadrature pairs of filters. In addition, repetition m indicates m sector cycles of the function values along the azimuth angle θ, while n and m jointly specify a different number of annular patterns of the function.

IV. Design of a Zernike Moment Phase-Based Descriptor

We shall use the ZM phase information to design a novel region descriptor. Let the Zernike moments be sorted by m and n in order. The total number of ZM moments of the same repetition m is equal to $\lfloor (N-m)/2 \rfloor + 1$. Table I gives the sorted list of the 42 complex ZM moments for the case where the maximum order N and maximum repetition M are both equal to 12.

The sorted Zernike moments form a feature vector $\vec{P}$ as follows:

$$\vec{P} = \left[|Z_{11}|e^{j\varphi_{11}}, |Z_{31}|e^{j\varphi_{31}}, \cdots\cdots, |Z_{NM}|e^{j\varphi_{NM}}\right]^T \tag{6}$$

TABLE I
List of ZMs Sorted by m and n in Sequence for the Case Where $(n, m) = (12, 12)$

m	Moments	No.	m	Moments	No.
1	$Z_{11}, Z_{31}, Z_{51}, Z_{71}, Z_{91}, Z_{11,1}$	*6*	**7**	$Z_{77}, Z_{97}, Z_{11,7}$	*3*
2	$Z_{22}, Z_{42}, Z_{62}, Z_{82}, Z_{10,2}, Z_{12,2}$	*6*	**8**	$Z_{88}, Z_{10,8}, Z_{12,8}$	*3*
3	$Z_{33}, Z_{53}, Z_{73}, Z_{93}, Z_{11,3}$	*5*	**9**	$Z_{99}, Z_{11,9}$	*2*
4	$Z_{44}, Z_{64}, Z_{84}, Z_{10,4}, Z_{12,4}$	*5*	**10**	$Z_{10,10}, Z_{12,10}$	*2*
5	$Z_{55}, Z_{75}, Z_{95}, Z_{11,5}$	*4*	**11**	$Z_{11,11}$	*1*
6	$Z_{66}, Z_{86}, Z_{10,6}, Z_{12,6}$	*4*	**12**	$Z_{12,12}$	*1*

where $|Z_{nm}|$ is the ZM magnitude, and φ_{nm} is the ZM phase. Here the Zernike moments $|Z_{nm}|e^{j\varphi_{nm}}$ with $m = 0$ are not included, since they provide no information regarding the image matching. Zernike moments with $m < 0$ are not included either, since they can be inferred through $Z_{n,-m} = Z_{nm}^{*}$.

A. Image Description Power of the ZM Magnitude Components and the ZM Phase Components

Let the Zernike moments of a reference image and its rotated version be Z_{nm}^{ref}, Z_{nm}^{rot}, respectively. Then it is well known that [24], [28]

$$Z_{nm}^{\mathrm{rot}} = Z_{nm}^{\mathrm{ref}} e^{-jm\alpha} \tag{7}$$

where $\alpha \in [0, 2\pi]$ is the rotation angle. Therefore, the magnitudes of Zernike moments of the two images are the same, i.e., $|Z_{nm}^{\mathrm{ref}}| = |Z_{nm}^{\mathrm{rot}}|$, but their phase difference (or phase shift) is given by

$$\Omega_{nm} \equiv \arg\left(\frac{Z_{nm}^{\mathrm{rot}}}{Z_{nm}^{\mathrm{ref}}}\right) = m\alpha, \quad 0 < \Omega_{nm} \leq 2m\pi, \text{ or} \tag{8}$$

$$\begin{aligned}\Phi_{nm} &= \left(\varphi_{nm}^{\mathrm{ref}} - \varphi_{nm}^{\mathrm{rot}}\right) \bmod (2\pi) \\ &= (m\alpha) \bmod (2\pi), \quad 0 < \Phi_{nm} \leq 2\pi.\end{aligned} \tag{9}$$

In the following, under a mixture of rotation, inversion, and flipping operations, the Zernike moments of a reference image can be shown to be rotationally invariant in terms of the magnitudes, but not the phases.

Let a rotated-and-inverted (the inverted is in terms of gray values) image version of the reference image $f^{\mathrm{ref}}(\rho, \theta)$ be given by $f^{\mathrm{rot-inv}}(\rho, \theta) = 255 - f^{\mathrm{ref}}(\rho, \theta + \alpha)$. It can readily be shown that the two magnitudes are equal ($|Z_{nm}^{\mathrm{ref}}| = |Z_{nm}^{\mathrm{rot-inv}}|$) and their phase difference is given by

$$\begin{aligned}\Phi_{nm} &= \left[\varphi_{nm}^{\mathrm{ref}} - \varphi_{nm}^{\mathrm{rot-inv}}\right] \bmod (2\pi) \\ &= \left[\varphi_{nm}^{\mathrm{ref}} - \left(\varphi_{nm}^{\mathrm{ref}} - m\alpha + \pi\right)\right] \bmod (2\pi).\end{aligned} \tag{10}$$

Next, let a rotated-and-mirrored version of the reference image $f^{\mathrm{ref}}(\rho, \theta)$ be given by $f^{\mathrm{rot-mirror}}(\rho, \theta) = f^{\mathrm{ref}}(\rho, \pi - (\theta + \alpha))$. Then it can be shown that their magnitudes are also equal: $|Z_{nm}^{\mathrm{rot-mirror}}| = |Z_{nm}^{\mathrm{ref}}|$ and their phase difference is given by

$$\begin{aligned}\Phi_{nm} &= \left(\varphi_{nm}^{\mathrm{ref}} - \varphi_{nm}^{\mathrm{rot-mirror}}\right) \bmod (2\pi) \\ &= \left[2\varphi_{nm}^{\mathrm{ref}} - m(\pi - \alpha)\right] \bmod (2\pi).\end{aligned} \tag{11}$$

B. Zernike Moment Phase Descriptor and Its Similarity Measure

From above, it can be seen that the phase information of Zernike moments is more informative than the magnitude information in terms of the discriminative power. Therefore, a new image region descriptor is proposed which is mainly based on the phase components of the feature vector, while the magnitude components are used only as the weighting factors.

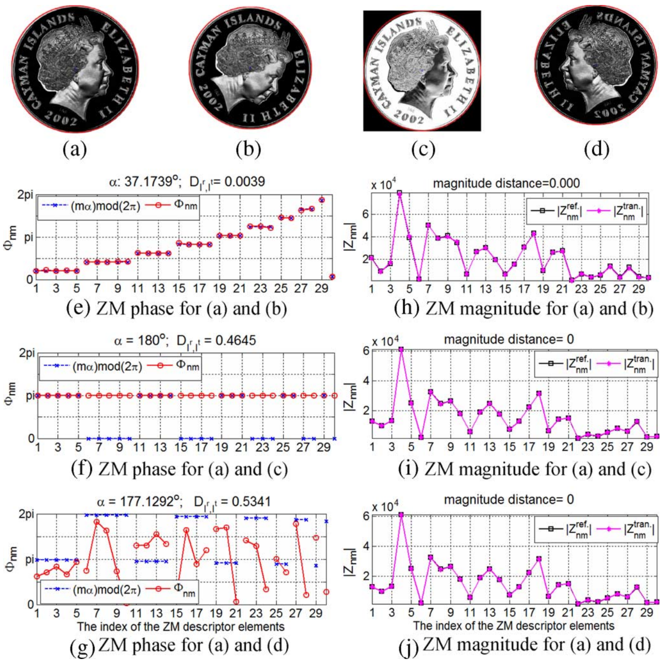
Let $I^{r}(x, y)$ and $I^{t}(x, y)$ as the reference and transformed image regions with their respective ZM feature vectors $\vec{P}_r = \{|Z_{nm}^{r}|e^{i\varphi_{nm}^{r}}\}$ and $\vec{P}_t = \{|Z_{nm}^{t}|e^{i\varphi_{nm}^{t}}\}$. Here the transformed image can be either a rotated version of the reference image or a different image. If there exists a rotation angle $\hat{\alpha}$ between $I^{r}(x, y)$ and $I^{t}(x, y)$, then $|\Phi_{nm} - (m\hat{\alpha}) \mathrm{mod}(2\pi)|$, which denotes the absolute phase difference between the two image regions after the rotation alignment, is equal to 0; otherwise, $|\Phi_{nm} - (m\hat{\alpha}) \mathrm{mod}(2\pi)|$ is a nonzero value in the interval $(0, 2\pi)$ and $\hat{\alpha}$ is simply a putative estimate of a nonexistent rotation angle. To derive a reliable estimate using all available phase differences Φ_{nm}, we define a weighted and normalized phase difference to check the existence of a rotation angle $\hat{\alpha}$ as follows [see (12), shown at the bottom of the page], where $\Phi_{nm} = (\varphi_{nm}^{r} - \varphi_{nm}^{t}) \mathrm{mod}(2\pi)$, $\hat{\alpha}$ is the estimated rotation angle to be described later, and w_{nm} is a normalized weighting factor of the form

$$w_{nm} = \frac{|Z_{nm}^{r}| + |Z_{nm}^{t}|}{\sum\limits_{n,m} \left(|Z_{nm}^{r}| + |Z_{nm}^{t}|\right)} \tag{13}$$

such that the phase components associated with small magnitudes are weighted less. The weighted and normalized phase difference $D_{I^{r}, I^{t}}$ lies in the interval [0, 1] and is dimensionless since it is derived from ratios of angles.

Fig. 2(a)–(d) shows a reference coin image and its three variants: a rotated one (with a rotation angle 37.22°), an inverted one, and a mirrored one, as described above. Image matching between the reference and each variant based on either the phase components or the magnitude components of Zernike moments are shown in Fig. 2(e)–(j), where the ZM order (n, m) ranges from (1, 1) to (10, 10). The estimated values of $(m\hat{\alpha}) \mathrm{mod}(2\pi)$ are colored in blue and are connected for components with the same m values. The actual phase differences Φ_{nm} are shown in the red color. On the other hand, the ZM magnitude components

$$D_{I^{r}, I^{t}} = \sum_{m} \sum_{n} w_{nm} \frac{\min\left\{\left|\Phi_{nm} - (m\hat{\alpha}) \mathrm{mod}(2\pi)\right|, 2\pi - \left|\Phi_{nm} - (m\hat{\alpha}) \mathrm{mod}(2\pi)\right|\right\}}{\pi} \tag{12}$$



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

(e) ZM phase for (a) and (b)

(h) ZM magnitude for (a) and (b)

(f) ZM phase for (a) and (c)

(i) ZM magnitude for (a) and (c)

(g) ZM phase for (a) and (d)

(j) ZM magnitude for (a) and (d)

Fig. 2. (a) Reference coin image. (b) A rotated variant of the reference coin image (with a rotation angle 37.22°). (c) An inverted variant. (d) A mirrored variant. (e)–(g) The diagrams of the ZM phase differences (h)–(j) The diagrams of the ZM magnitude components.

for each pair of images are colored in purple. Notice that the magnitude component diagrams are the same for all the three pairs, but the phase component diagrams are different. Therefore, the phase components have a better discriminative power than the magnitude components.

C. Estimation of the Rotation Angle From a Rotated Image

In [29], Kim and Kim represented the rotation angle between an original image and its rotated image through the use of the Zernike moment phase shift as

$$\Omega_{nm}=(\varphi_{nm}+2k_1\pi)-(\varphi^r_{nm}+2k_2\pi)=\Phi_{nm}+2\pi k_{nm}=m\alpha. \tag{14}$$

They then proposed a probabilistic model $P(\hat{\alpha}) = \sum_m\sum_n \xi_{nm}P(\hat{\alpha}|n,m)$ to estimate the rotation angle α where ξ_{nm} is the weighting factor proportional to the ZM magnitude $|Z_{nm}|$. For each possible solution $\hat{\alpha}_{nm} = (\Phi_{nm}/m) + (2\pi/m)k_{nm}$, they used a probability density function $P(\hat{\alpha}|n,m) = (1/m)\sum_{k_{nm}=0}^{m-1}\delta\{\hat{\alpha} - ((\Phi_{nm}/m)+(2\pi/m)k_{nm})\} * G(\hat{\alpha},\sigma_{nm})$, a convolution of an impulse train with a scaled Gaussian kernel, to estimate α. Notice that the estimation is done in the discrete angle steps. In order to be accurate, the estimation step size must be as small as possible. Let the estimation step size be 0.01°. For the case where $(N,M) = (10,10)$, there are 30 generated Zernlike moments $\{Z_{nm}\}$. From each fixed Zernike moment Z_{nm}, an estimator of the rotation angle is given by $\hat{\alpha}_{nm} = (\Phi_{nm}/m) + (2\pi/m)k_{nm}$. There are 30 such estimators. To find the common solution to the rotation angle α using these 30 estimators, a common histogram with a bin size of 360 × 100 (assuming the estimation step size is 0.01°) is used to tabulate the possible rotation angle produced by each of the 30 estimators. Therefore, the total number of histogram bin values computed is 360 × 100 × 30 (=1,080,000), which is rather large. In addition, the method may face the ambiguity in multiple peaks in the histogram constructed.

Here a new estimation method of the rotation angle $\hat{\alpha}$ is proposed, which is implemented in the continuous angle space rather than in the discrete space. The basic idea behind the proposed method for estimating the rotation angle $\hat{\alpha}_m$ is to avoid the m ambiguities in the value of k_{nm}. Instead, the rotation angle $\hat{\alpha}$ can be found from the phase difference using any two adjacent Φ_{nm} and $\Phi_{n,m-1}$, $m \neq 0$, through

$$\begin{aligned}\alpha &= m\alpha-(m-1)\alpha\\ &=(\Phi_{nm}+2\pi k_{nm})-(\Phi_{n,m-1}+2\pi k_{n,m-1})\\ &=(\Phi_{nm}-\Phi_{n,m-1})\bmod 2\pi, \quad m\neq 0.\end{aligned} \tag{15}$$

Since $m = 1,2,\ldots,M$, $n = 1,2,\ldots,N$, there are $\sum_{m=1}^{M}(\lfloor (N-m)/2\rfloor + 1)$ ways to compute the rotation angle $\hat{\alpha}$. A more robust estimation is to weight the estimated angles by the individual magnitude $|Z_{nm}|$.

An iterative computation of the rotation angle $\hat{\alpha}$ using all available Zernike moments sorted by m is given below:

The ZM phase-based rotation angle estimation algorithm

```
Initialization: $\hat{\alpha}_0 = 0$ and $c_0 = 0$
For $m = 1,2,\ldots,M$
    For $n = m, m+2,\ldots,m+2\lfloor (N-m)/2\rfloor$
        $\delta_{nm} = [(\Phi_{nm}-(m-1)\hat{\alpha}_{m-1}] \bmod 2\pi$
        $w_{nm} = \frac{|Z^r_{nm}|+|Z^t_{nm}|}{2}$
    End
        $s_m = \sum_{k=0}^{\lfloor \frac{N-m}{2}\rfloor} \frac{w_{m+2k,m}}{m}$
        $\delta_m = \frac{1}{s_m}\sum_{k=0}^{\lfloor \frac{N-m}{2}\rfloor} \frac{w_{m+2k,m}}{m}\delta_{m+2k,m}$
        $\hat{\alpha}_m = \frac{1}{c_{m-1}+s_m}(c_{m-1}\hat{\alpha}_{m-1}+s_m\delta_m)$
        $c_m = c_{m-1}+s_m$
End
$\hat{\alpha} = \hat{\alpha}_M$
```

V. Experimental Results for Performance Evaluation

We will examine the system performance with respect to important system parameters including 1) region scene types, 2) region descriptor types, 3) region detector types, 4) region overlap error, and 5) transformation types. The region scene types under consideration are the structured and textured scenes. The test images available at the website [30], plus some new images, are used in the experiments. The transformation types considered

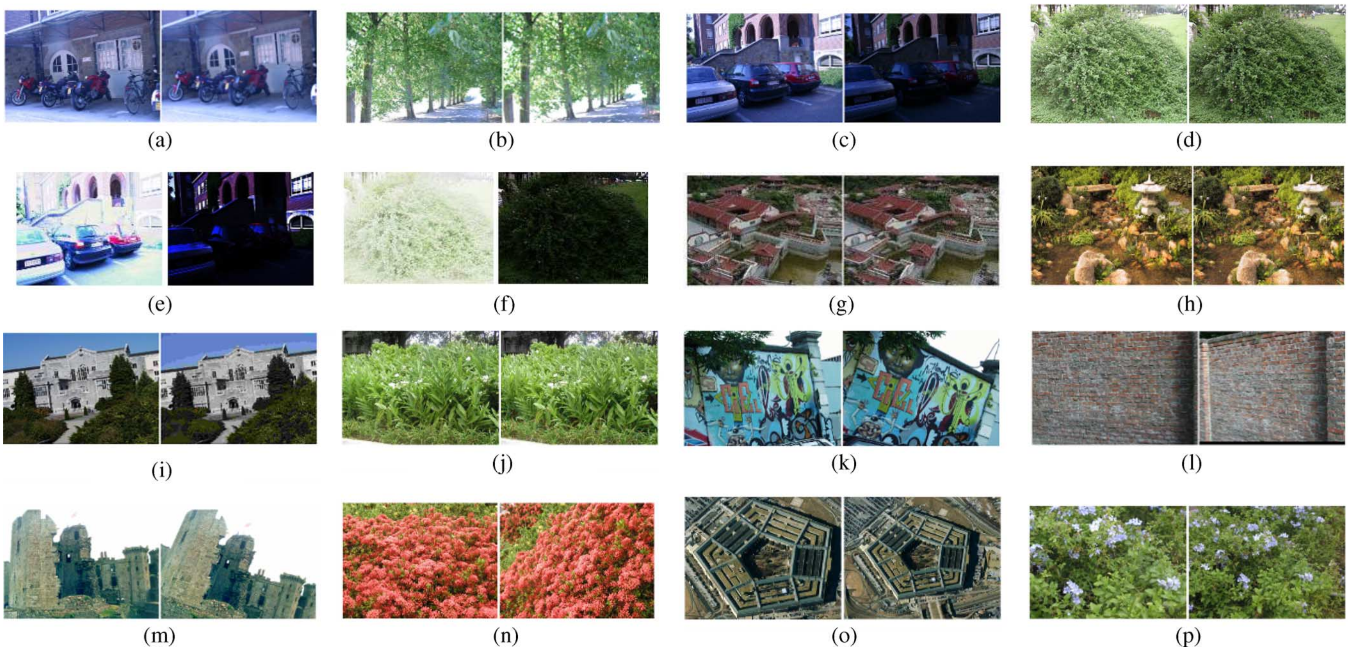


Fig. 3. Representative test image pairs taken from the textured and structured scenes under a specified photometric or geometric transformation. (a) Bikes (blur), (b) tree (blur), (c) Leuven (lighting), (d) bush 1 (lighting), (e) Leuven (nonlinear lighting), (f) bush 1 (nonlinear lighting), (g) Chinese compound (noise), (h) Japanese garden (noise), (i) UBC (JPEG), (j) garden (JPEG), (k) graffiti (viewpoint), (l) wall brick (viewpoint), (m) castle (rotation), (n) flower (rotation), (o) Pentagon (scaling), (p) bush 2 (scaling).

here contain the common photometric transformations (blur, illumination, noise, and JPEG compression) and geometric transformations (rotation, scaling, translation, and viewpoint). Fig. 3 shows the representative test image pairs taken for the textured and structured scenes.

In regard to the region descriptor types we include the proposed ZM phase and five popular descriptors: SIFT, GLOH, PCA-SIFT, steerable filters, and complex moments. In the beginning of the experiment, we need to choose a region detector in order to extract the regions of interest from the given image. Here, we decide to choose either MSER detector or Hessian-affine detector. Once the region detector type is decided, the program codes available at the website [30] are used to obtain (a) regions of interest, (b) the dominant orientation in a region image, and (c) the descriptor feature vectors of SIFT, GLOH, PCA-SIFT, steerable filter and complex moment for each region of interest. Then we run our program codes to generate our ZM phase descriptor, and to calculate the similarity measures and generate the precision-recall curves to evaluate the descriptor performances, as done in [12]. Totally, there are eight types of transformations, two types of scenes, and at least four image pairs for each transformation. On the average, one image pair generates 250,000 $(= 500 \times 500)$ region pairs for matching. All together the experiments involve more than 15 million region pairs.

Table II lists the typical feature vector dimensions of the six descriptors used in the experiments. Later, a discussion on the feature dimensionality will be provided.

A. Performance Evaluation Criteria-PR Curve

For region matching, the extracted regions of the reference and transformed images are examined for (a) their distance measure and (b) their spatial overlap error under the applied transformation. There are three strategies for region matching proposed in [12]: (a) the threshold-based matching, (b) the nearest-neighbor-based matching, and (c) two-nearest-neighbor-based matching. Although these three matching methods are functionally different, their ranking results of the performances of the various descriptors are virtually the same; the first one is generally recommended [12], [38]. Therefore, we adopt the threshold-based matching strategy in which the distance measure between a region pair is compared to a given distance threshold, D_t.

TABLE II
TYPICAL FEATURE VECTOR DIMENSIONS OF THE SIX DESCRIPTORS

Descriptor	SIFT	GLOH	ZM phase	PCA-SIFT	Complex moments	Steerable filters
Feature dimension	128	128	42	36	15	14

On the other hand, the region overlap error is represented by the overlap ratio between the region intersection area and the region union area under the known planar homography [12], [31], that is, $O_e = 1 - (A \cap H^T BH)/(A \cup H^T BH)$, where A and B are the two matching regions and H is the given homograph between the two region patches. A region pair is called a match if it passes the region similarity test, namely, the distance measure between the image pair does not exceed the distance threshold D_t; otherwise, no match is found. A match is said to be correct, if the region pair also passes the region overlap test given by $O_e < O_t$ for a given overlap error threshold O_t. A match is said to be false, if the pair fails the region overlap test. Sometimes, with a tight overlap error threshold, say $O_t = 0.1$, even though the two regions pass the region similarity test, but they fail the region overlap test due to $O_t < O_e < 1$. It seems not very fair to

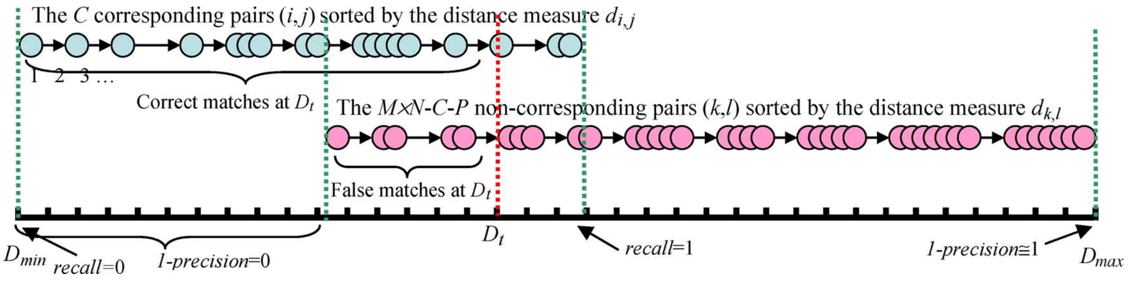

The C corresponding pairs (i,j) sorted by the distance measure $d_{i,j}$
1 2 3 …
Correct matches at D_t
The $M\times N$-C-P non-corresponding pairs (k,l) sorted by the distance measure $d_{k,l}$
False matches at D_t
D_{min}
recall=0
1-precision=0
D_t
recall=1
1-precision≅1
D_{max}

Fig. 4. PR curve generation process with a varying distance threshold D_t.

call such a pair a false match when compared to a typical false match whose region overlap error O_e is equal to 1; namely, the two regions do not intersect and are, therefore, not related at all. Hereafter, a matching pair with a region overlap error in between such that $O_t < O_e < 1$ is considered as a "don't care" pair. In other words, the new definition of a false match is a match that passes the region similarity test and its region overlap error O_e must be equal to 1.

It is important to realize a fixed distance threshold cannot be used to evaluate the descriptor performances. Instead, a precision-recall (PR) curve, created by varying the distance threshold, must be used.

Recall is the ratio of the number of correct matches to the number of corresponding region pairs satisfying the region overlap test: $O_e < O_t$

$$\text{recall} = \frac{\#\ \text{correct matches}}{\#\ \text{correspondences}}. \tag{16}$$

Precision is the ratio of the number of correct matches to the total number of correct and false matches

$$1 - \text{precision} = \frac{\#\ \text{false matches}}{\#\ \text{correct matches} + \#\ \text{false matches}}. \tag{17}$$

Fig. 4 depicts a PR curve generation process. Assume there are M, N regions detected in the reference and transformed images, respectively. The regions in the two images form $M \times N$ matching region pairs. Among these $M \times N$ pairs let the number of corresponding region pairs, which are each with a region overlap error O_e smaller than the specified bound O_t, be C. Also, let the number of the "don't care" pairs be P. Now sort the C corresponding pairs and the $M \times N - C - P$ noncorresponding pairs, respectively, by their distance measures $d_{i,j}$ in an ascending order. The range of distance measures for the set of C corresponding pairs generally overlaps with that of the set of noncorresponding pairs. Start to increase the distance threshold D_t from the minimum value $D_{\min}$ to the maximum value $D_{\max}$, as shown in Fig. 4. The recall value is initially equal to zero, so is the value of (1-precision). As D_t passes over $D_{\min}$, more and more correct matches occur and the recall value is increasing, while the (1-precision) value remains 0 since there have been no false matches so far. When D_t reaches the minimum distance measure of the noncorresponding region pairs, false matching pairs begin to appear and the value of 1-precision is increasing from 0. Notice that the recall is always monotonically increasing and reaches 1 when the distance threshold is equal to the maximum distance measure of the C corresponding region pairs. At the end, when the distance threshold is equal to $D_{\max}$, the (1-precision) value approaches 1. Be aware that the (1-precision) value is monotonically increasing when D_t is sufficiently large, but it may decrease at the early stage, if the relative growth rate of false matches is smaller than that of the correct matches.

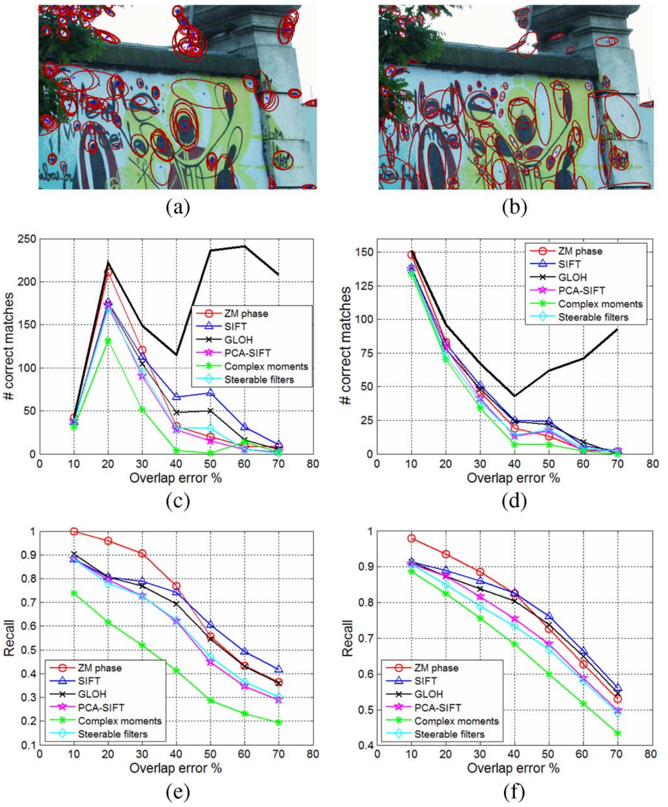


Fig. 5. Evaluation for different overlap errors. (a), (b) Detected Hessian-affine regions and MSER regions under viewpoint change for structured graffti scene. (c), (d) The number of correct matches versus the overlap error. Also, the top black line shows the number of region correspondences detected. (e), (f) Recall versus the overlap error.

B. Effects of Region Detector Types and Region Overlap Error

As mentioned above, the best two region detectors, MSER and Hessian-affine, are reported in [11]. We shall present the evaluation results for these two detectors side by side.

Fig. 5 shows the region detection results and the two curves about the relation between recall and region overlap error and that between the number of correct matches and region overlap error using Hessian-affine regions and MSER regions, respectively. The images used are a pair of graffiti structure scene.

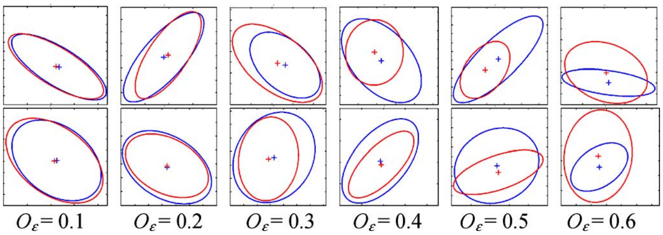


Fig. 6. Examples of the detected region pairs with different overlap errors O_ε ranging from 0.1 to 0.6. The ellipses indicate the region boundary with blue color and red color for reference region A and the transformed region given by H^TBH, respectively. The cross symbols show the key point positions.

There are around 400 regions extracted by either detector. The number of correct matches and the number of correspondences for each overlap error are computed for a single section of overlap errors ranging from the previous one to the current one. For instance, the score for 20% is computed for the overlap error interval from 10% to 20%. Also, the recall values are calculated, by keeping the precision at 0.5, as done in [12]. We observe that the top black line, which shows the number of region correspondences dictated by the given overlap error bound O_e, bounces back at overlap error 40%. This is due to a natural increase in the region correspondences at the given higher region overlap error bound, resulting in "one-to-many" or "many-to-one" overlapped region pairs extracted from the reference and sensed scenes. Usually these new corresponding region pairs are less similar in comparison to those at a smaller overlap error bound, causing a drop in the number of new correct matches. On the other hand, for a small overlap error bound the correspondences are mostly the "one-to-one" overlapped region pairs.

We observe that the proposed ZM phase descriptor has a higher recall versus region overlap error curve than other descriptors for the region overlap error in the interval [0.1, 0.4] for both sets of Hessian-affine and MSER regions. The portion of curve is less meaningful when O_t gets larger. This is because when O_t gets larger, the corresponding regions are less similar, as indicated in Fig. 6. As mentioned above, when the overlap error bound increases over 0.4, the intersection area of these new corresponding region pairs becomes smaller, resulting in the drop of the number of correct matches and the decrease in the recall value under a fixed precision level (0.5 in this case). At a large overlap error bound the Zernike phase maintains the same tight control on the similarity matching of the new corresponding pairs based on the orthogonal moment features, so the increase in the new correct matches is rather small. On the other hand, SIFT and GLOH have less stringent control on the similarity measure based on the 8-gradient orientation bin tabulation on the 4 × 4 location grid, so there are more new correct matches when the overlap error bound increases.

We should not bother considering the corresponding region pairs associated with a large overlap error bound, since many belong to "one-to-many" or "many-to-one" correspondences. The inclusion of these less similar pairs or outliers will result in the erroneous estimations in the later stages such as in the estimations of homography, fundamental matrix and epipolar geometry, etc. Therefore, we set the O_t value to 0.3 rather than 0.5 used in [12].

From now on, only MSER regions will be considered in the later experiments, since the descriptor performance characteristics are similar for MSER and Hessian-affine regions.

C. Transformation Types

Since the elliptical region is already normalized into a circular image, the normalized region is affine invariant. Nevertheless, the normalized region is not necessarily invariant to rotation. Thus, for most of the descriptors including SIFT, SIFT variants and the steerable filters, the image rotation problem must be solved first by finding a dominant gradient orientation. Similarly, the circular image intensity normalization has made the region descriptor robust to intensity scaling and offset, but not to image blur, image noise, image compression, and the illumination change.

In image registration the two images can be taken by a single camera or different cameras, and the images can be taken during a short period or on different days. These shooting scenarios determine the type of image transformation encountered. For instance, if the two images are shot by different cameras or at different periods, the photometric conditions of the two shootings will be different, not to mention the possible viewpoint change. In general, a geometric transformation is accompanied by some sort of photometric change due to differences in the camera setting and the surface reflection angles.

1) Robustness Under Photometric Transformations: To focus on the effects of photometric transformations, we try to avoid the effect of a geometric transformation by setting the region overlap error threshold O_t to a small value (0.2 $\sim$ 0.3). Overall speaking, the ZM phase obtains the best performance results for all textured scenes under all type of photometric transformations and for the structured scenes under image blur and nonlinear lighting. The performances of the ZM phase, SIFT, GLOH and PCA-SIFT are comparable for the structured scenes under affine lighting change, image noise and JPEG when the value of 1-precision is very small. The analysis on these performance results will be given later.

a) Image blur: The performance is measured under image blur introduced by changing the camera focus setting. Fig. 7(a) and (b) shows the respective PR curves for the bike structured scene [see Fig. 3(a)] and the tree textured scene [see Fig. 3(b)]. The performance ranking indicates that the best descriptor is ZM phase for both the structured and textured scenes considered. On the other hand, SIFT performs better than its variants, GLOH and PCA-SIFT, for the textured scene, while its variant performs better for the structured scene, as reported in [12]. The last ranking position is the complex moments. This is because its low-dimensional feature vector (15 in this case) and its exclusive use of the moment magnitudes without the phase information.

b) Illumination Change:

(i) Affine Lighting Change

To evaluate the descriptor performances under illumination changes, a collection of images has been taken by changing the camera iris settings. Fig. 7(c) and (d) shows the PR curves for the Leuven structured scene and the bush 1 textured scene shown in Fig. 3(c) and (d), respectively. The best three descriptors in order are ZM phase, SIFT, and GLOH for the bush 1 textured

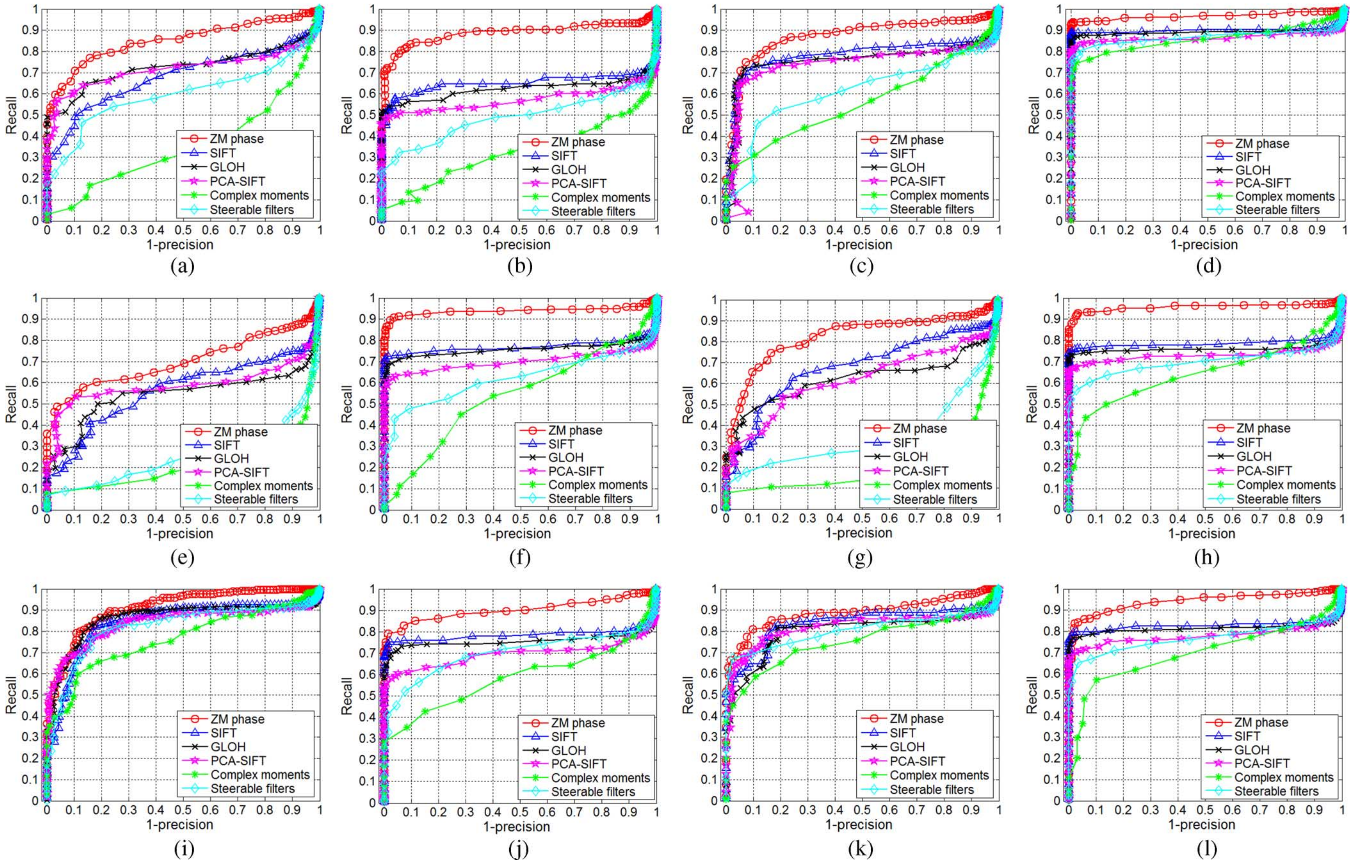


Fig. 7. PR curves for performance evaluations under different specified photometric transformations, all with an overlap error threshold $O_t = 0.3$. (a) Image blur, (b) image blur (textured), (c) lighting (structured), (d) lighting (textured), (e) overexposure (structured), (f) overexposure (textured), (g) underexposure (structured), (h) underexposure (textured), (i) image noise (structured), (j) image noise (textured), (k) JPEG (structured), (l) JPEG (textured).

scene and the situation remains the same for the structured scene except when the value of 1-precision is less than 0.03.

(ii) Nonlinear Lighting Change

The nonlinear lighting is quite common in practice. Fig. 7(e)–(h) shows the PR curves under the overexposure and underexposure lighting for the Leuven structured scene and the bush 1 textured scene shown in Fig. 3(e) and (f). In comparison with the PR curves in Fig. 7(c) and (d), it can be seen that the performances of the SIFT-based descriptors become significantly worse. To the contrary, the performance results of the ZM phase change insignificantly, especially in the case of the textured scene. This will be explained later.

c) Image noise: The performances are evaluated by adding a different amount of Gaussian noise to the images. Fig. 7(i) and (j) shows the PR curve for a Chinese compound structured scene [see Fig. 3(g)] and a Japanese garden textured scene [see Fig. 3(h)], respectively. The ZM phase has the best overall result among all the descriptors for the textured scene and is comparable to the SIFT-based descriptors for the structured scene.

d) JPEG Compression

Fig. 7(k) and (l) depicts the PR curves under JPEG compression for the UBC structured scene shown in Fig. 3(i) and the garden textured scene shown in Fig. 3(j), respectively. The qualities of the compressed images range from 10% to 30% of the original one. The performance ranking is similar to that under the noise attack.

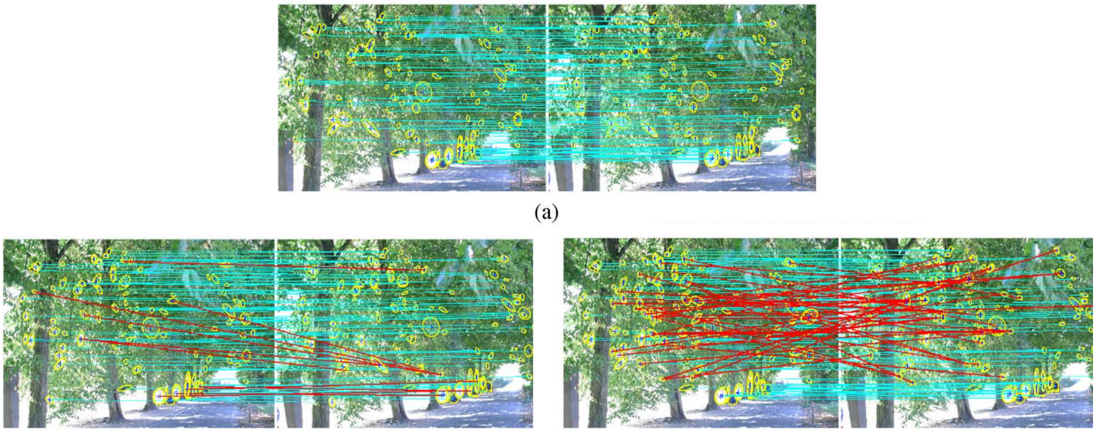
To show the performance discrepancies between the top best three descriptors (ZM phase, GLOH and SIFT) under image blur, Table III shows the matching statistics for the bike structured scene and the tree textured scene with a region overlap error of 0.3 and a recall value of 0.6. Fig. 8 depicts the correct and false region matches for the tree textured scene, when using ZM phase, GLOH and SIFT, respectively. There are 0, 11, and 42 false matches (shown by red lines) for ZM phase, SIFT and GLOH, respectively. All these descriptors have 112 correct matches (shown by green lines).

2) Robustness Under Geometric Transformations: To focus on the effects of geometric transformations, we try intentionally not to change the photometric conditions. As shall be seen, under all geometric transformations, the ZM phase performs best for all textured scenes, but is comparable to the SIFT-based descriptors for the structured scenes when the value of 1-precision is less than 0.05.

a) Viewpoint change: We used six images of the textured and structured scenes taken under a viewing angle ranging from 10 to 50 degrees. Fig. 9(a) and (b) gives the PR curves for graffiti structured scenes [see Fig. 3(k)] and the brick textured scenes [see Fig. 3(l)], respectively. The ranking of the four best descriptors remain unchanged for the specified viewing angle range [10°, 50°]. The ZM phase descriptor clearly overpowers the five other descriptors for the textured scene, but not so for the structured scene.

TABLE III
Matching Statistics for the Bike Structured Scene and Tree Textured Scene, All With $O_t = 0.3$ and $\mathrm{recall} = 0.6$

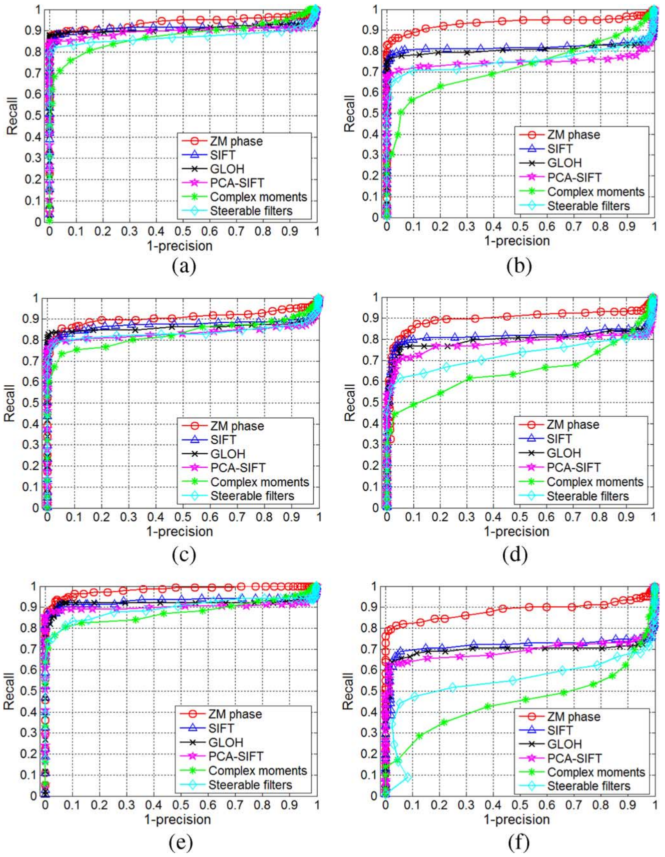
Scene	# MSER regions		# correspondences	ZM phase			SIFT			GLOH		
	Left Image	Right image		Threshold D_t	# correct	# false	Threshold D_t	# correct	# false	Threshold D_t	# correct	# false
Structured (bikes)	449	387	161	0.167	97	4	0.183	96	35	1600	96	14
Textured (tree)	631	531	186	0.179	112	0	0.220	112	11	1543	112	42


(a)

(b)

(c)

Fig. 8. Correct matches (in green) and false matches (in red) obtained by the descriptors, respectively, all with $\mathrm{recall} = 0.6$ and $O_t = 0.3$. (a) By ZM phase, (b) by SIFT, (c) by GLOH.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9. PR curves under geometric transformation, all with $O_t = 0.3$. (a) Viewpoint (structured), (b) viewpoint (textured), (c) rotation (structured), (d) rotation (textured), (e) scaling (structured), (f) scaling (textured).

b) Rotation change: The images considered are taken by rotating the camera axis from 30° to 45°. The descriptors for the castle structured scene [Fig. 3(m)] and the flower textured scene [Fig. 3(n)] under image rotation are evaluated. Fig. 9(c) and (d) shows the PR curves for the scenes, respectively. The ranking of the top three descriptors remains the same throughout the range of rotation angle and it is similar to the case of viewpoint change.

c) Scale change: Fig. 9(e) and (f) shows the performance measures for the descriptors under the scale change using the Pentagon structured scene [Fig. 3(o)] and bush 2 textured scene [Fig. 3(p)], respectively. The scaling factor is close to 2. The performance rankings are similar to the above two cases of geometric transformations.

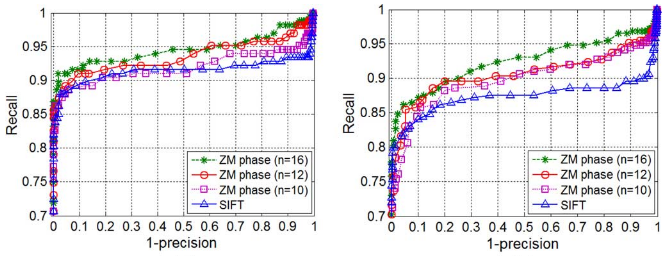
D. Feature Dimensionality

To extend the SIFT descriptor, both GLOH and PCA-SIFT increase the feature size and then apply PCA to reduce the feature dimensionality. The features of these descriptors are originally correlated and become orthogonal after the application of PCA. However, their optimal dimensions are determined by the training images in the database.

The utilization of Zernike moments up to a higher order generally leads to a more accurate estimate of the region rotation angle and a better image representation power. Fig. 10 depicts the PR curves for two structured scenes under two different attacks where the ZM descriptor uses moments of order N up to 10, 12, and 16, respectively. The corresponding feature dimensions are 30, 42, and 72. It can be seen that the descriptor performance becomes better as the feature dimension gets increased. The selection of order $N = 12$ is a tradeoff between the computational complexity and the descriptor performance.

TABLE IV
ROTATION ANGLE ESTIMATION ERRORS FOR ALL CORRESPONING REGION PAIRS SPECIFIED BY $O_t = 0.3$

Transform type	Scene type	method	<5°		< 10°		< 20°		< 30°	
			mean	coverage	mean	coverage	mean	coverage	mean	coverage
blur	Textured (tree)	ZM	1.801°	86.022%	2.333°	97.312%	2.502°	98.925%	2.502°	98.925%
		SIFT	2.484°	39.785%	4.134°	59.140%	6.090°	72.581%	7.787°	80.108%
	Structured (Bikes)	ZM	1.663°	92.547%	2.027°	98.758%	2.162°	100%	2.162°	100%
		SIFT	2.147°	51.553%	3.605°	72.671%	4.543°	80.124%	5.080°	82.609%
affine lighting	Textured (bush 1)	ZM	0.764°	96.755%	0.947°	99.705%	0.979°	100%	0.979°	100%
		SIFT	1.747°	68.437%	2.724°	84.366%	3.366°	89.676%	3.426°	89.971%
	Structured (Leuven)	ZM	1.506°	93.662%	1.641°	98.775%	2.115°	100%	2.115°	100%
		SIFT	1.726°	62.676%	2.631°	78.873%	3.313°	83.803%	4.071°	86.620%
non-linear lighting (underexp.)	Textured (bush 1)	ZM	1.099°	94.561%	1.284°	97.908%	1.363°	98.745%	1.363°	98.745%
		SIFT	2.002°	53.556%	3.115°	69.874%	4.327°	79.498%	4.532°	80.335%
	Structured (Leuven)	ZM	1.220°	93.662%	1.481°	95.592%	1.584°	99.296%	1.715°	100%
		SIFT	1.906°	63.380%	2.944°	80.282%	3.308°	83.803%	3.463°	84.507%
noise	Textured (Japan garden)	ZM	1.564°	92.857%	1.838°	98.352%	1.893°	98.901%	1.893°	98.901%
		SIFT	2.423°	42.857%	4.071°	65.385%	5.859°	80.121%	6.765°	83.516%
	Structured (Compound)	ZM	1.349°	93.293%	1.666°	99.085%	1.763°	100%	1.763°	100%
		SIFT	1.781°	69.207%	2.814°	85.671%	3.377°	90.244%	3.377°	90.244%
JPEG	Textured (garden)	ZM	1.318°	93.817%	1.654°	100%	1.654°	100%	1.654°	100%
		SIFT	2.107°	50.269%	3.722°	73.387%	4.948°	83.871%	5.272°	85.215%
	Structured (UBC)	ZM	1.112°	93.158%	1.326°	96.842%	1.552°	98.947%	1.552°	98.947%
		SIFT	1.852°	68.947%	2.724°	82.632%	3.162°	86.316%	3.419°	87.368%
Rotation	Textured (flower)	ZM	1.310°	97.692%	1.370°	99.231%	1.370°	99.231%	1.370°	99.231%
		SIFT	2.346°	54.483%	3.910°	81.379%	4.662°	89.655%	4.973°	91.034%
	Structured (castle)	ZM	1.061°	98.755%	1.117°	100%	1.117°	100%	1.117°	100%
		SIFT	1.777°	74.274%	2.544°	87.552%	2.963°	91.286%	2.963°	91.286%
Scaling	Textured (bush 2)	ZM	1.414°	92.623%	1.625°	97.541%	1.840°	99.180%	1.840°	99.180%
		SIFT	2.222°	53.279%	3.519°	70.492%	4.340°	76.230%	5.390°	80.328%
	Structured (Pentagon)	ZM	0.913°	98.551%	0.999°	100%	0.999°	100%	0.999°	100%
		SIFT	1.356°	78.261%	2.154°	90.58%	2.529°	93.478%	2.529°	93.478%



(a) (b)

Fig. 10. PR curves for ZM phase with the maximum order $N = 10$, 12, and 16, together with the associated PR curves of SIFT for two structured scenes under two different attacks, all with $O_t = 0.3$. (a) Graffiti scene (viewpoint change), (b) castle scenes (rotation change).

VI. ANALYSIS ON PERFORMANCE EVALUTION RESULTS

Since the complex moments and the steerable filters never rank in the first position throughout the experiments, they will be excluded for further consideration. In addition, the SIFT, GLOH, and PCA-SIFT have similar performance results under all the transformations reported. In the following, it is sufficient to compare the performances of SIFT and ZM phase.

A. Rotation Angle Error Statistics and Its Effect on the Descriptor Performance

The descriptor performance discrepancy can be attributed to the accuracy of the rotation angle estimation by the descriptors. The dominant orientation of the SIFT descriptor relies on the peak detection in the 36-bin histogram of the gradient directions obtained from the region image, while the ZM phase descriptor computes the image rotation angle via the weighted and normalized phase difference. Table IV breaks down the estimated rotation angle errors $(\varepsilon_{\text{angle}})$ under the categories of 5, 10, 20, and 30 degrees for both textured scenes and structured scenes under all transformations except viewpoint change. The rotation angle errors are evaluated by computing the estimated rotation angle for all normalized corresponding region pairs, and further compare them with respect to the actual angle. The actual angle can be obtained by the ground truth homographies given from [30], which are essentially similarity transforms. The rotation angle error statistics are not available under the viewpoint change, since the associated rotation angle between two regions under viewpoint change is not fixed.

From Table IV the average rotation angle errors of the ZM phase is less than those of SIFT for both structured scenes and textured scenes when $\varepsilon_{\text{angle}} < 30°$. More importantly, the coverage percentage for ZM phase is more than 86% while SIFT only has 40% to 78% coverage when $\varepsilon_{\text{angle}} < 5°$. The coverage percentage is computed as the ratio between the number of region pairs, with rotation angle estimation error $(\varepsilon_{\text{angle}})$ less than a specific value ($\varepsilon_t = 5°$, 10°, 20°, or 30° in Table IV), and the total number of correspondence

$$\text{coverage percentage} = \frac{\#\ \text{corresponding pairs with}\ \varepsilon_{\text{angle}} < \varepsilon_t}{\#\ \text{correspondences}}.$$

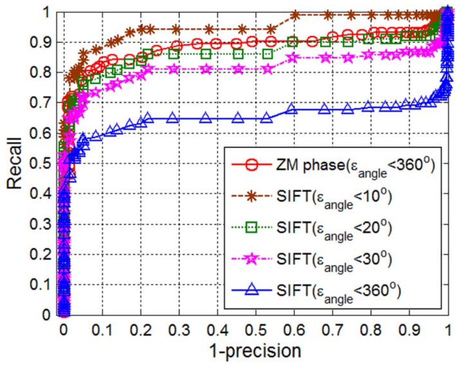


Fig. 11. PR curves for the tree textured scene under image blur with the removal of regions with rotation angle error not exceeding a specified level of 10°, 20°, 30°, and 360°, respectively.

The large rotation angle errors of SIFT are due to the big error caused by ambiguity in the multiple dominant orientation peaks. This is the main reason why the SIFT performance becomes poor.

Lowe [7] suggested solving the multiple dominant orientation problem by creating multiple keypoints at the same location, each with one of the multiple dominant orientations (In this case there is no clear rule for counting the multiple keypoints as correct or false matches when generating the PR curves). In Fig. 11, the PR curves for the flower textured scene under image blur is plotted where region pairs with rotation angle error no less than 10°, 20°, 30°, and 360° are removed, respectively. The ZM phase performs better than SIFT for rotation angle errors not exceeding 20°, 30°, and 360°, but not for the case of rotation angle errors $< 10°$, where SIFT does not face the multiple dominant orientation problem, as described previously.

B. *Effects of Feature Dimensionality and Feature Orthogonality on the Descriptor Performance*

Generally speaking, the high-dimensional feature vector contains more descriptive information at the expense of memory space. For example, PCA-SIFT and GLOH start with a feature dimension of 3042 and 272, respectively. However, the components of these feature vectors are correlated and partially redundant. By the application of PCA, a subset of eigenvectors associated with the larger eigenvalues can be extracted and the projection of the original feature vector to the sub-eigenspace reduces the original dimension down to 128 or even smaller. The dimensionality reduction can be determined based on the percentage of the sum of eigenvalues retained.

We know the ZM phase applies a set of orthogonal ZM moments to design the feature vector such that the feature components are mutually independent and more informative. With the same dimensionality (or the same memory space) the set of orthogonal features generally results in a better descriptive power to distinguish the different image patterns embedded in the textured scenes. However, when the image patterns in the scenes are highly similar, it requires a higher feature dimensionality in order to reflect the subtle pattern difference, as indicated previously in Fig. 10.

C. *Effect of Image Intensity Fluctuation on the Descriptor Performance*

Finally, we give a rule of thumb or a simplified explanation why the ZM phase descriptor performs better than other existing descriptors under nonuniform image intensity fluctuation, since the exact analysis varies with the underlying image and, therefore, is rather complicated. First of all, the transformed image is obtained from the reference image in accordance with a given photometric or geometric transform, so their image pattern structures are correlated. After the affine intensity normalization, their image intensity distributions become closer and tangled. Next, the phase difference of the ZM phase descriptor is computed as

$$\begin{aligned}\Delta\varphi_{nm} &= \varphi_{nm}^{\mathrm{tran}} - \varphi_{nm}^{\mathrm{ref}}\\ &= \tan^{-1}\left(\frac{\mathrm{Im}\left(Z_{nm}^{\mathrm{tran}}\right)}{\mathrm{Re}\left(Z_{nm}^{\mathrm{tran}}\right)}\right) - \tan^{-1}\left(\frac{\mathrm{Im}\left(Z_{nm}^{\mathrm{ref}}\right)}{\mathrm{Re}\left(Z_{nm}^{\mathrm{ref}}\right)}\right)\end{aligned}$$

where $\mathrm{Im}(Z_{nm}^{\mathrm{tran}})/\mathrm{Re}(Z_{nm}^{\mathrm{tran}}) = (\mathrm{Im}(Z_{nm}^{\mathrm{ref}}) + \Delta\mathrm{Im}(Z_{nm}))/(\mathrm{Re}(Z_{nm}^{\mathrm{ref}}) + \Delta\mathrm{Re}(Z_{nm}))$ with $\Delta\mathrm{Re}(Z_{nm})$ and $\Delta\mathrm{Im}(Z_{nm})$ being the real and imaginary ZM components of the difference image between the reference and transformed images. Since the image structures of the transformed and reference images are similar, so it is likely that the phase angles of the reference and transformed images are in phase (i.e., no phase difference after the image rotation alignment), especially when their ZM magnitudes are both large. The weighted sum of the absolute phase differences is, therefore, close to zero. On the other hand, the probability that the reference and transformed images are out of a phase (a significant phase difference) is small. Consequently, most of the ZM moment counterparts of the mage pair support the single majority of the estimated rotation angle, even though there is some fluctuation in the ZM magnitudes. This leads to the accurate rotation angle estimation when using the ZM phase.

On the other hand, the SIFT-based methods utilize the gradient information. The local gradient angles in the transformed image remain considerably unchanged (except under image blur which causes the gradient angles damaged), but their gradient magnitudes change somewhat nonuniformly. Besides, there are generally several different gradient angles found in an image especially for the textured image. (This may not be the case for structured scenes with a distinguished edge orientation.) Therefore, the 36-bin orientation histogram will contain multiple candidates on the histogram ballot. When gradient magnitudes change nonuniformly, the vote counting result of the multiple candidates will change. This leads to a change of the dominant orientation in the transformed image. It, in turn, triggers further nonlinear changes in the 128-dimensional SIFT feature vector, regardless of the unit length feature vector renormalization at the process end. This is why the performance of the SIFT-based methods generally degrades under a given transformation especially for the textured scenes. We shall use an example to justify our above reasoning.

Fig. 12 shows the result for the performance comparison between ZM phase and SIFT under nonlinear lighting change (a power-law (gamma) transform with $\mathrm{gamma} = 3$). Fig. 12(a)

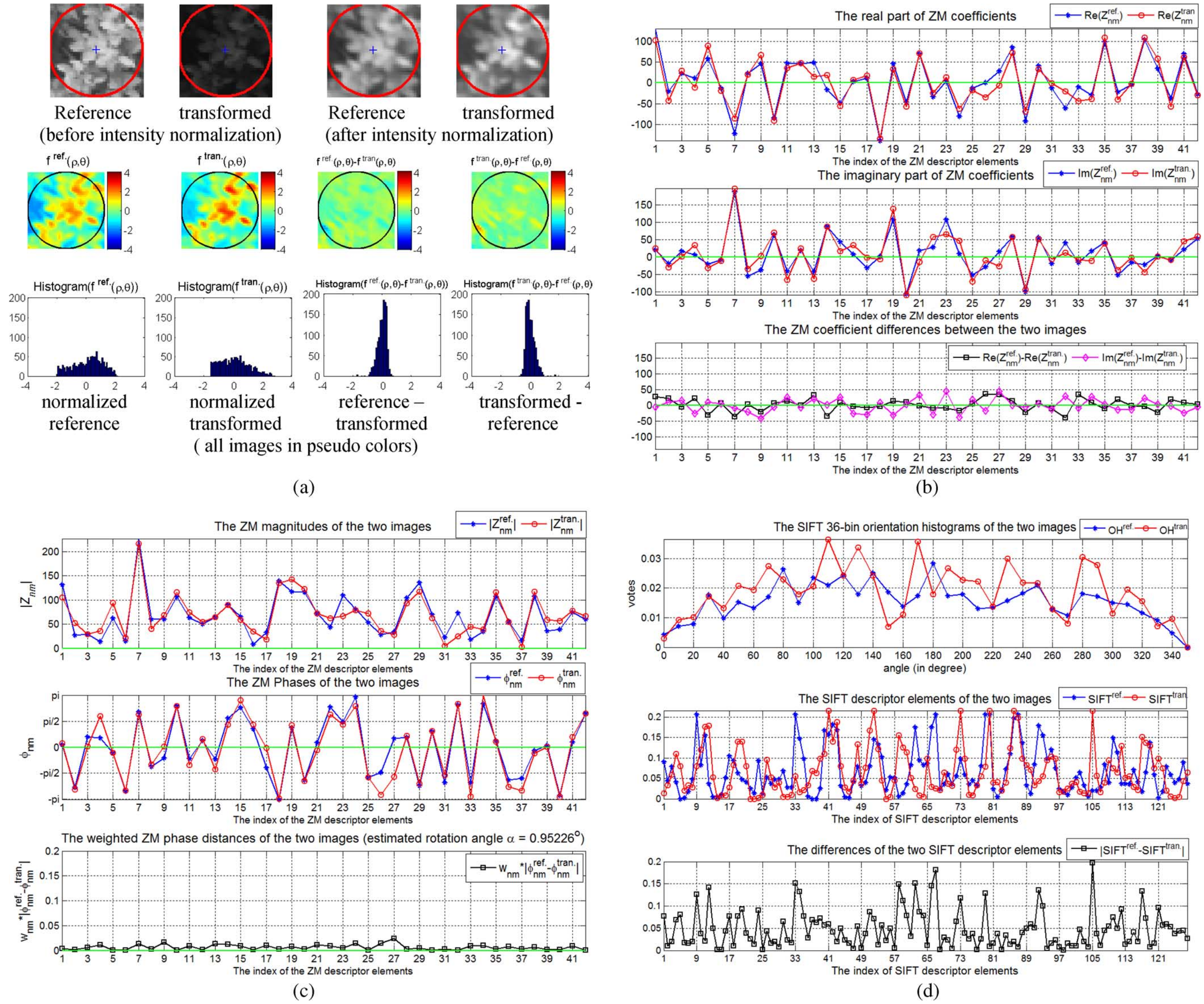


Fig. 12. Performance comparison of ZM phase and SIFT under nonlinear lighting change. The detected ellipse-shaped regions are normalized to a circular patch through the affine normalization process beforehand.

shows the region pair before and after affine intensity normalization in the gray color or in the pseudo color for better visualization, along with their difference images and difference intensity histograms. We can observe that the image structure of the transformed and difference images looks similar to that of the reference image. This leads to the fact that despite a few parts, the real and imaginary parts of the ZM moments for region pair are nearly identical, as indicated in Fig. 12(b). Therefore, the majority of the weighted phase differences are nearly zero, as shown in Fig. 12(c). On the other hand, the nonuniform intensity fluctuation causes the dominant orientation histogram and the 128-dimensional SIFT feature vectors to change nonuniformly, resulting in an expected greater dissimilarity between the two images shown in Fig. 12(d).

In summary, noise, lighting change, compression, and blurring belong to the photometric transformation type which causes the image intensities to vary. On the other hand, viewpoint change, scaling and rotation belong to the geometric transformation type which first relocates the positions of the image points, and then requires some sort of intensity interpolation to compute the image intensities at the new image points; the new image intensities contain some nonuniform fluctuation (except the rotation transformation which generally causes a very minor intensity fluctuation). We can apply the above-mentioned reasoning to conclude the ZM phase descriptor is generally more robust than the SIFT-based methods under these transformations, especially for the textured scenes which generally contain the complex edge orientation information.

D. Time Complexity

The computation time for evaluating the descriptor performance consists of the region extraction time, the descriptor feature vector construction time and the region matching time. Because all descriptors use the same set of regions of interest detected, so their region extraction times are the same. As for the feature vector construction time, the numbers of multiplications and additions required to compute Zernike moments up to order N for a $q \times q$ image patch are both of order $O(N^2q^2)$ [40]. However, this calculation can be accelerated by using the symmetrical properties of Zernike basis functions [41], or achieve in real time performance by using special hardware accumulation grid architecture [42]. As for the region matching including the rotation angle estimation, the numbers of multiplications and additions required by the ZM phase descriptor are both of order $O(N^2)$. Theoretically speaking, the SIFT-based descriptor has a shorter region matching time per region pair, compared to the ZM phase descriptor. However, if desired, we can first use the ZM moment magnitude components, which are known rotationally invariant, to compute the distance between two given feature vectors. Only when the magnitude-based distance passes the condition checking, the ZM phase descriptor needs further calculation of the weighted and normalized phase difference to check if there exists a rotation angle between two matching regions.

VII. Conclusions

In this paper, a new region descriptor called the ZM phase is presented, which is robust to common photometric and geometric transformations. A method for an accurate and robust estimation of the rotation angle between two matching regions, implemented in the continuous angle domain without the need of specifying a discrete angle histogram bin resolution, is described. Then a measure for image similarity matching is expressed by a weighted and normalized phase difference. The proposed descriptor is compared with five popular descriptors, SIFT, PCA-SIFT, GLOH, steerable filter, and complex moments, based on the precision-recall criterion with respect to a number of important system parameters. There are more than 15 million region pairs analyzed. The results show that the proposed ZM phase has the leading performance under all photometric and geometric transformations for all textured scenes. As for the structured scenes, the ZM phase has the best performances under image blur and nonlinear lighting, but is comparable to the SIFT-based descriptors under other transformations when the values of 1-precision are small. The analyses on the performance evaluation results are given to account for the performance discrepancy. First, the descriptor performance depends on the estimation accuracy of the rotation angle between two matching regions. Table IV shows the rotation angle estimation error of the ZM phase is better when compared to SIFT. Second, the feature dimensionality and feature orthogonality also affect the descriptor performance. Third, the ZM phase is more robust than SIFT-based descriptors under the nonuniform image intensity fluctuation.

Further investigation on the incorporation of the proposed descriptor into various applications such as textured image classification and image retrieval is currently underway. Hopefully, the ZM phase descriptor can be better understood and improved.

References

[1] L. Van Gool, T. Moons, and D. Ungureanu, "Affine/photometric invariants for planar intensity patterns," in *Proc. 4th Eur. Conf. Computer Vision*, 1996, vol. II, pp. 642–651.

[2] C. Schmid and R. Mohr, "Local grayvalue invariants for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 530–535, May 1997.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[4] T. Ueshiba and F. Tomita, "Plane-based calibration algorithm for multi-camera systems via factorization of homography matrices," in *Proc. Int. Conf. Computer Vision*, 2003, vol. 2, pp. 966–973.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, pp. 761–767, 2004.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[8] T. Tuytelaars and L. Van Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.

[9] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[10] T. Lindeberg and J. Garding, "Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure," *Image Vis. Comput.*, vol. 15, no. 6, pp. 415–434, 1997.

[11] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, and J. Matas, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1/2, pp. 43–72, 2005.

[12] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[13] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, pp. 1771–1787, 2008.

[14] W. Freeman and E. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[15] T. S. Lee, "Image representation using Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–97, Oct. 1996.

[16] J. Heikkilä, "Pattern matching with affine moment descriptors," *Pattern Recognit.*, vol. 37, pp. 1825–1834, 2004.

[17] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognit.*, vol. 27, pp. 1–19, 2004.

[18] S. Paschalakis and P. Lee, "Pattern recognition in grey level images using moment based invariant features," in *Proc. Int. Conf. Image Processing and Its Applications*, 1999, vol. 1, pp. 245–249.

[19] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets," in *Proc. 7th Eur. Conf. Computer Vision*, 2002, pp. 414–431.

[20] C.-H. Teh and R. T. Chin, "On image analysis by the methods of moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 4, pp. 496–513, Apr. 1988.

[21] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[22] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 506–513.

[23] L. M. J. Florack, J. Koenderink, B. M. T. H. Romeny, J. J. Koenderink, and M. A. Viergever, "General intensity transformations and differential invariants," *J. Math. Imag. Vis.*, vol. 4, no. 2, pp. 171–187, 1994.

[24] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.

[25] W. Y. Kim and Y. S. Kim, "A region-based shape descriptor using Zernike moments," *Signal Process.: Image Commun.*, vol. 16, pp. 95–102, 2000.

[26] S. K. Hwang, M. Billinghurst, and W. Y. Kim, "Local descriptor by Zernike moments for real-time keypoint matching," in *Proc. IEEE Congr. Image Signal Process.*, 2008, pp. 781–785.

[27] Y. Xin, M. Pawlak, and S. Liao, "Accurate computation of Zernike moments in polar coordinates," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 581–587, Feb. 2007.

[28] H. Lin, J. Si, and G. P. Abousleman, "Orthogonal rotation-invariant moments for digital image processing," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 272–282, Mar. 2007.

[29] W. Y. Kim and Y. S. Kim, "Robust rotation angle estimator," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 768–773, Aug. 1999.

[30] [Online]. Available: http://www.robots.ox.ac.uk/~vgg/research/affine/

[31] Z. Chen and H. L. Chou, "A novel 3D planar object reconstruction from multiple uncalibrated images using the plane-induced homographies," *Pattern Recognit. Lett.*, vol. 25, no. 12, pp. 1399–1410, 2004.

[32] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, pp. 179–187, 1962.

[33] A. Baumberg, "Reliable feature matching across widely separated views," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2000, pp. 774–781.

[34] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–550, 1981.

[35] D. J. Fleet and A. D. Jepson, "Stability of phase information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 12, pp. 1253–1268, Dec. 1993.

[36] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77–104, 1990.

[37] G. Carneiro and A. D. Jepson, "Multi-scale phase-based local features," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2003, pp. I/736–I/743.

[38] S. Winder and M. Brown, "Learning local image descriptors," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[39] M. R. Teague, "Image analysis via the general theory of moments," *J. Opt. Soc. Amer.*, vol. 70, no. 8, pp. 1468–1478, 1980.

[40] G. Amayeh, A. Erol, G. Bebis, and M. Nicolescu, "Accurate and efficient computation of high order Zernike moments," in *Proc. 1st Int. Symp. Vision and Computation*, 2005, pp. 462–469.

[41] S. K. Hwang and W. Y. Kim, "A novel approach to the fast computation of Zernike moments," *Pattern Recognit.*, vol. 39, no. 11, pp. 2065–2076, Nov. 2006.

[42] L. Kotoulas and I. Andreadis, "Real-time computation of Zernike moments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 801–809, 2005.

Zen Chen is a Professor of computer science at National Chiao Tung University, Hsin-Chu, Taiwan, R.O.C. He received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN.

After graduating from Purdue University, he was with Burroughs Corporation, Detroit, MI, and joined National Chiao Tung University, Taiwan, afterward. He spent a year (1981–1982) at the Lawrence Berkeley Laboratory, University of California, Berkeley, as a visiting scientist and half a year as a visiting professor (August 1989 to January 1990) at the Center for Automation Research, University of Maryland, College Park. His research interests include computer vision, pattern recognition, virtual reality, as well as parallel algorithms and architectures.

Dr. Chen is a Fellow of the International Association for Pattern Recognition (IAPR). He was the founding President of the Chinese Society of Image Processing and Pattern Recognition in Taiwan and a member of the society of IAPR. He received the Outstanding Engineering Professor Award from the Chinese Institute of Engineers and the Outstanding Research Awards from the National Science Council of Taiwan.

Shu-Kuo Sun received the M.S. degree in electrical engineering from the Cheng Chung Institute of Technology, Tao-yuan, Taiwan, R.O.C., in 1993, and the Ph.D. degree in computer science from National Chiao Tung University, Hsin-Chu, Taiwan, in 2009.

He is currently a Postdoctoral Fellow at the Department of Computer Science, National Chiao Tung University. His research interests include feature detection, region descriptor, image registration, remote sensing, and pattern recognition.