

A performance evaluation of local descriptors

Krystian Mikolajczyk and Cordelia Schmid

Dept. of Engineering Science

University of Oxford

Oxford, OX1 3PJ

United Kingdom

km@robots.ox.ac.uk

INRIA Rhône-Alpes

655, av. de l'Europe

38330 Montbonnot

France

schmid@inrialpes.fr

Abstract

In this paper we compare the performance of descriptors computed for local interest regions, as for example extracted by the Harris-Affine detector [32]. Many different descriptors have been proposed in the literature. However, it is unclear which descriptors are more appropriate and how their performance depends on the interest region detector. The descriptors should be distinctive and at the same time robust to changes in viewing conditions as well as to errors of the detector. Our evaluation uses as criterion recall with respect to precision and is carried out for different image transformations. We compare shape context [3], steerable filters [12], PCA-SIFT [19], differential invariants [20], spin images [21], SIFT [26], complex filters [37], moment invariants [43], and cross-correlation for different types of interest regions. We also propose an extension of the SIFT descriptor, and show that it outperforms the original method. Furthermore, we observe that the ranking of the descriptors is mostly independent of the interest region detector and that the SIFT based descriptors perform best. Moments and steerable filters show the best performance among the low dimensional descriptors.

Index Terms

Local descriptors, interest points, interest regions, invariance, matching, recognition.

I. INTRODUCTION

Local photometric descriptors computed for interest regions have proved to be very successful in applications such as wide baseline matching [37, 42], object recognition [10, 25], texture

Corresponding author is K. Mikolajczyk, km@robots.ox.ac.uk.

recognition [21], image retrieval [29, 38], robot localization [40], video data mining [41], building panoramas [4], and recognition of object categories [8, 9, 22, 35]. They are distinctive, robust to occlusion and do not require segmentation. Recent work has concentrated on making these descriptors invariant to image transformations. The idea is to detect image regions covariant to a class of transformations, which are then used as support regions to compute invariant descriptors.

Given invariant region detectors, the remaining questions are which is the most appropriate descriptor to characterize the regions, and does the choice of the descriptor depend on the region detector. There is a large number of possible descriptors and associated distance measures which emphasize different image properties like pixel intensities, color, texture, edges etc. In this work we focus on descriptors computed on gray-value images.

The evaluation of the descriptors is performed in the context of matching and recognition of the same scene or object observed under different viewing conditions. We have selected a number of descriptors, which have previously shown a good performance in such a context and compare them using the same evaluation scenario and the same test data. The evaluation criterion is recall-precision, i.e. the number of correct and false matches between two images. Another possible evaluation criterion is the ROC (Receiver Operating Characteristics) in the context of image retrieval from databases [6, 31]. The detection rate is equivalent to recall but the false positive rate is computed for a database of images instead of a single image pair. It is therefore difficult to predict the actual number of false matches for a pair of similar images.

Local features were also successfully used for object category recognition and classification. The comparison of descriptors in this context requires a different evaluation setup. However, it is unclear how to select a representative set of images for an object category and how to prepare the ground truth, since there is no linear transformation relating images within a category. A possible solution is to select manually a few corresponding points and apply loose constraints to verify correct matches, as proposed in [18].

In this paper the comparison is carried out for different descriptors, different interest regions and for different matching approaches. Compared to our previous work [31], this paper performs a more exhaustive evaluation and introduces a new descriptor. Several descriptors and detectors have been added to the comparison and the data set contains a larger variety of scenes types and transformations. We have modified the evaluation criterion and now use recall-precision for image pairs. The ranking of the top descriptors is the same as in the ROC based evaluation [31].

Furthermore, our new descriptor, gradient location and orientation histogram (*GLOH*), which is an extension of the SIFT descriptor, is shown to outperform SIFT as well as the other descriptors.

A. Related work

Performance evaluation has gained more and more importance in computer vision [7]. In the context of matching and recognition several authors have evaluated interest point detectors [14, 30, 33, 39]. The performance is measured by the repeatability rate, that is the percentage of points simultaneously present in two images. The higher the repeatability rate between two images, the more points can potentially be matched and the better are the matching and recognition results.

Very little work has been done on the evaluation of local descriptors in the context of matching and recognition. Carneiro and Jepson [6] evaluate the performance of point descriptors using ROC (Receiver Operating Characteristics). They show that their phase-based descriptor performs better than differential invariants. In their comparison interest points are detected by the Harris detector and the image transformations are generated artificially. Recently, Ke and Sukthankar [19] have developed a descriptor similar to the SIFT descriptor. It applies Principal Components Analysis (PCA) to the normalized image gradient patch and performs better than the SIFT descriptor on artificially generated data. The criterion recall-precision and image pairs were used to compare the descriptors.

Local descriptors (also called filters) have also been evaluated in the context of texture classification. Randen and Husoy [36] compare different filters for one texture classification algorithm. The filters evaluated in this paper are Laws masks, Gabor filters, wavelet transforms, DCT, eigenfilters, linear predictors and optimized finite impulse response filters. No single approach is identified as best. The classification error depends on the texture type and the dimensionality of the descriptors. Gabor filters were in most cases outperformed by the other filters. Varma and Zisserman [44] also compared different filters for texture classification and showed that MRF perform better than Gaussian based filter banks. Lazebnik et al. [21] propose a new invariant descriptor called “spin image” and compare it with Gabor filters in the context of texture classification. They show that the region-based spin image outperforms the point-based Gabor filter. However, the texture descriptors and the results for texture classification cannot be directly transposed to region descriptors. The regions often contain a single structure without repeated patterns, and the statistical dependency frequently explored in texture descriptors cannot

be used in this context.

B. Overview

In section II we present a state of the art on local descriptors. Section III describes the implementation details for the detectors and descriptors used in our comparison as well as our evaluation criterion and the data set. In section IV we present the experimental results. Finally, we discuss the results.

II. DESCRIPTORS

Many different techniques for describing local image regions have been developed. The simplest descriptor is a vector of image pixels. Cross-correlation can then be used to compute a similarity score between two descriptors. However, the high dimensionality of such a description results in a high computational complexity for recognition. Therefore, this technique is mainly used for finding correspondences between two images. Note that the region can be sub-sampled to reduce the dimension. Recently, Ke and Sukthankar [19] proposed to use the image gradient patch and to apply PCA to reduce the size of the descriptor.

Distribution based descriptors. These techniques use histograms to represent different characteristics of appearance or shape. A simple descriptor is the distribution of the pixel intensities represented by a histogram. A more expressive representation was introduced by Johnson and Hebert [17] for 3D object recognition in the context of range data. Their representation (spin image) is a histogram of the relative positions in the neighborhood of a 3D interest point. This descriptor was recently adapted to images [21]. The two dimensions of the histogram are distance from the center point and the intensity value.

Zabih and Woodfill [45] have developed an approach robust to illumination changes. It relies on histograms of ordering and reciprocal relations between pixel intensities which are more robust than raw pixel intensities. The binary relations between intensities of several neighboring pixels are encoded by binary strings and a distribution of all possible combinations is represented by histograms. This descriptor is suitable for texture representation but a large number of dimensions is required to build a reliable descriptor [34].

Lowe [25] proposed a scale invariant feature transform (SIFT), which combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The

descriptor is represented by a 3D histogram of gradient locations and orientations, see figure 1 for illustration. The contribution to the location and orientation bins is weighted by the gradient magnitude. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and small errors in the region detection. Geometric histogram [1] and shape context [3] implement the same idea and are very similar to the SIFT descriptor. Both methods compute a 3D histogram of location and orientation for edge points where all the edge points have equal contribution in the histogram. These descriptors were successfully used, for example, for shape recognition of drawings for which edges are reliable features.

Spatial-frequency techniques. Many techniques describe the frequency content of an image. The Fourier transform decomposes the image content into the basis functions. However, in this representation the spatial relations between points are not explicit and the basis functions are infinite, therefore difficult to adapt to a local approach. The Gabor transform [13] overcomes these problems, but a large number of Gabor filters is required to capture small changes in frequency and orientation. Gabor filters and wavelets [27] are frequently explored in the context of texture classification.

Differential descriptors. A set of image derivatives computed up to a given order approximates a point neighborhood. The properties of local derivatives (*local jet*) were investigated by Koenderink [20]. Florack et al. [11] derived differential invariants, which combine components of the *local jet* to obtain rotation invariance. Freeman and Adelson [12] developed steerable filters, which steer derivatives in a particular direction given the components of the *local jet*. Steering derivatives in the direction of the gradient makes them invariant to rotation. A stable estimation of the derivatives is obtained by convolution with Gaussian derivatives. Figure 2(a) shows Gaussian derivatives up to order 4.

Baumberg [2] and Schaffalitzky and Zisserman [37] proposed to use complex filters derived from the family $K(x, y, \theta) = f(x, y) \exp(i\theta)$, where θ is the orientation. For the function $f(x, y)$ Baumberg uses Gaussian derivatives and Schaffalitzky and Zisserman apply a polynomial (cf. section III-B and figure 2(b)). These filters differ from the Gaussian derivatives by a linear coordinates change in filter response space.

Other techniques. Generalized moment invariants have been introduced by Van Gool et al. [43] to describe the multi-spectral nature of the image data. The invariants combine central moments defined by $M_{pq}^a = \int \int_{\Omega} x^p y^q [I(x, y)]^a dx dy$ with order $p + q$ and degree a . The moments char-

acterize shape and intensity distribution in a region Ω . They are independent and can be easily computed for any order and degree. However, the moments of high order and degree are sensitive to small geometric and photometric distortions. Computing the invariants reduces the number of dimensions. These descriptors are therefore more suitable for color images where the invariants can be computed for each color channel and between the channels.

III. EXPERIMENTAL SETUP

In the following we first describe the region detectors used in our comparison and the region normalization necessary for computing the descriptors. We then give implementation details for the evaluated descriptors. Finally, we discuss the evaluation criterion and the image data used in the tests.

A. Support regions

Region detectors use different image measurements and are either scale or affine invariant. Lindeberg [23] has developed a scale-invariant “blob” detector, where a “blob” is defined by a maximum of the normalized Laplacian in scale-space. Lowe [25] approximates the Laplacian with difference-of-Gaussian (DoG) filters and also detects local extrema in scale-space. Lindeberg and Gårding [24] make the blob detector affine-invariant using an *affine adaptation* process based on the second moment matrix. Mikolajczyk and Schmid [29, 30] use a multi-scale version of the Harris interest point detector to localize interest points in space and then employ Lindeberg’s scheme for scale selection and affine adaptation. A similar idea was explored by Baumberg [2] as well as Schaffalitzky and Zisserman [37]. Tuytelaars and Van Gool [42] construct two types of affine-invariant regions, one based on a combination of interest points and edges and the other one based on image intensities. Matas et al. [28] introduced Maximally Stable Extremal Regions extracted with a watershed like segmentation algorithm. Kadir et al. [18] measure the entropy of pixel intensity histograms computed for elliptical regions to find local maxima in affine transformation space. A comparison of state-of the art affine region detectors can be found in [33].

1) *Region detectors*: The detectors provide the regions which are used to compute the descriptors. If not stated otherwise the detection scale determines the size of the region. In this evaluation we have used five detectors :

Harris points [15] are invariant to rotation. The support region is a fixed size neighborhood of 41×41 pixels centered at the interest point.

Harris-Laplace regions [29] are invariant to rotation and scale changes. The points are detected by the scale-adapted Harris function and selected in scale-space by the Laplacian-of-Gaussian operator. Harris-Laplace detects corner-like structures.

Hessian-Laplace regions [25, 32] are invariant to rotation and scale changes. Points are localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian. This detector is similar to the DoG approach [26], which localizes points at local scale-space maxima of the difference-of-Gaussian. Both approaches detect the same blob-like structures. However, Hessian-Laplace obtains a higher localization accuracy in scale-space, as DoG also responds to edges and detection is unstable in this case. The scale selection accuracy is also higher than in the case of the Harris-Laplace detector. Laplacian scale selection acts as a matched filter and works better on blob-like structures than on corners since the shape of the Laplacian kernel fits to the blobs. The accuracy of the detectors affects the descriptor performance.

Harris-Affine regions [32] are invariant to affine image transformations. Localization and scale are estimated by the Harris-Laplace detector. The affine neighborhood is determined by the *affine adaptation* process based on the second moment matrix.

Hessian-Affine regions [32, 33] are invariant to affine image transformations. Localization and scale are estimated by the Hessian-Laplace detector and the affine neighborhood is determined by the *affine adaptation* process.

Note that Harris-Affine differs from Harris-Laplace by the affine adaptation, which is applied to Harris-Laplace regions. In this comparison we use the same regions except that for Harris-Laplace the region shape is circular. The same holds for the Hessian based detector. Thus the number of regions is the same for affine and scale invariant detectors. Implementation details for these detectors as well as default thresholds are described in [32]. The number of detected regions varies from 200 to 3000 per image depending on the content.

2) *Region normalization*: The detectors provide circular or elliptic regions of different size, which depends on the detection scale. Given a detected region it is possible to change its size or shape by scale or affine covariant construction. Thus, we can modify the set of pixels which contribute to the descriptor computation. Typically, larger regions contain more signal

variations. Hessian-Affine and Hessian-Laplace detect mainly blob-like structures for which the signal variations lie on the blob boundaries. To include these signal changes into the description, the measurement region is 3 times larger than the detected region. This factor is used for all scale and affine detectors. All the regions are mapped to a circular region of constant radius to obtain scale and affine invariance. The size of the normalized region should not be too small in order to represent the local structure at a sufficient resolution. In all experiments this size is arbitrarily set to 41 pixels. A similar patch size was used in [19]. Regions which are larger than the normalized size, are smoothed before the size normalization. The parameter σ of the smoothing Gaussian kernel is given by the ratio measurement/normalized region size. Spin images, differential invariants and complex filters are invariant to rotation. To obtain rotation invariance for the other descriptors the normalized regions are rotated in the direction of the dominant gradient orientation, which is computed in a small neighborhood of the region center. To estimate the dominant orientation we build a histogram of gradient angles weighted by the gradient magnitude and select the orientation corresponding to the largest histogram bin, as suggested in [25].

Illumination changes can be modeled by an affine transformation $aI(\mathbf{x}) + b$ of the pixel intensities. To compensate for such affine illumination changes the image patch is normalized with mean and standard deviation of the pixel intensities within the region. The regions, which are used for descriptor evaluation, are normalized with this method if not stated otherwise. Derivative-based descriptors (steerable filters, differential invariants) can also be normalized by computing illumination invariants. The offset b is eliminated by the differentiation operation. The invariance to linear scaling with factor a is obtained by dividing the higher order derivatives by the gradient magnitude raised to the appropriate power. A similar normalization is possible for moments and complex filters, but has not been implemented here.

B. Descriptors

In the following we present the implementation details for the descriptors used in our experimental evaluation. We use ten different descriptors: SIFT [25], gradient location and orientation histogram (GLOH), shape context [3], PCA-SIFT [19], spin images [21], steerable filters [12], differential invariants [20], complex filters [37], moment invariants [43], and cross-correlation of sampled pixel values. Gradient location and orientation histogram (*GLOH*) is a new descriptor

which extends SIFT by changing the location grid and using PCA to reduce the size.

SIFT descriptors are computed for normalized image patches with the code provided by Lowe [25]. A descriptor is a 3D histogram of gradient location and orientation, where location is quantized into a 4x4 location grid and the gradient angle is quantized into 8 orientations. The resulting descriptor is of dimension 128. Figure 1 illustrates the approach. Each orientation plane represents the gradient magnitude corresponding to a given orientation. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

Gradient location-orientation histogram (GLOH) is an extension of the SIFT descriptor designed to increase its robustness and distinctiveness. We compute the SIFT descriptor for a log-polar location grid with 3 bins in radial direction (the radius set to 6, 11 and 15) and 8 in angular direction (cf. figure 1(e)), which results 17 location bins. Note that the central bin is not divided in angular directions. The gradient orientations are quantized in 16 bins. This gives a 272 bin histogram. The size of this descriptor is reduced with PCA. The covariance matrix for PCA is estimated on 47 000 image patches collected from various images (see section III-C.1). The 128 largest eigenvectors are used for description.

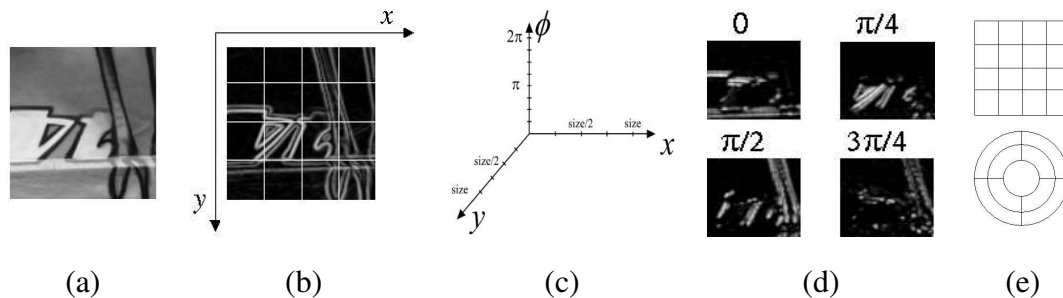


Fig. 1. SIFT descriptor. (a) Detected region. (b) Gradient image and location grid. (c) Dimensions of the histogram. (d) 4 of 8 orientation planes. (e) Cartesian and the log-polar location grids. The log-polar grid shows 9 location bins used in shape context (4 in angular direction).

Shape context is similar to the SIFT descriptor, but is based on edges. Shape context is a 3D histogram of edge point locations and orientations. Edges are extracted by the Canny [5] detector. Location is quantized into 9 bins of a log-polar coordinate system as displayed in figure 1(e) with the radius set to 6, 11 and 15 and orientation quantized into 4 bins (horizontal, vertical and two diagonals). We therefore obtain a 36 dimensional descriptor. In our experiments we weight a point contribution to the histogram with the gradient magnitude. This has shown to give better

results than using the same weight for all edge points, as proposed in [3]. Note that the original shape context was computed only for edge point locations and not for orientations.

PCA-SIFT descriptor is a vector of image gradients in x and y direction computed within the support region. The gradient region is sampled at 39×39 locations therefore the vector is of dimension 3042. The dimension is reduced to 36 with PCA.

Spin image is a histogram of quantized pixel locations and intensity values. The intensity of a normalized patch is quantized into 10 bins. A 10 bin normalized histogram is computed for each of 5 rings centered on the region. The dimension of the spin descriptor is 50.

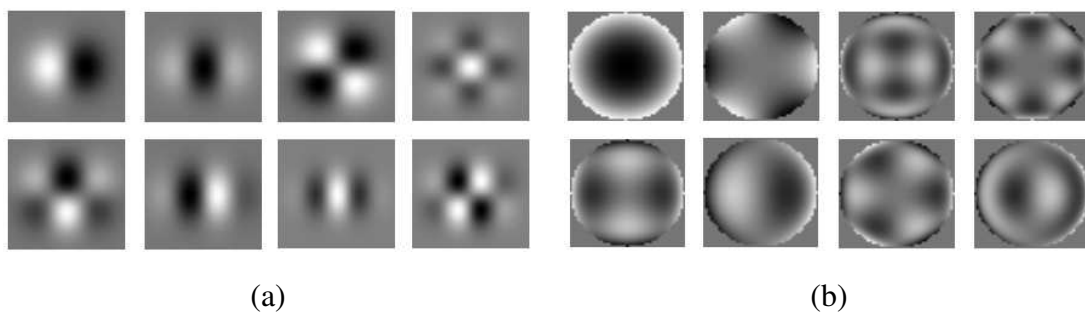


Fig. 2. Derivative based filters. (a) Gaussian derivatives up to 4th order. (b) Complex filters up to 6th order. Note that the displayed filters are not weighted by a Gaussian, for figure clarity.

Steerable filters and *differential invariants* use derivatives computed by convolution with Gaussian derivatives of $\sigma = 6.7$ for an image patch of size 41. Changing the orientation of derivatives as proposed in [12] gives equivalent results to computing the local jet on rotated image patches. We use the second approach. The derivatives are computed up to 4th order, that is the descriptor has dimension 14. Figure 2(a) shows 8 of 14 derivatives; the remaining derivatives are obtained by rotation by 90° . The differential invariants are computed up to 3rd order (dimension 8). We compare steerable filters and differential invariants computed up to the same order (cf. section IV-A.3).

Complex filters are derived from the following equation $K_{mn}(x, y) = (x + iy)^m (x - iy)^n G(x, y)$. The original implementation [37] has been used for generating the kernels. The kernels are computed for a unit disk of radius 1 and sampled at 41×41 locations. We use 15 filters defined by $m + n \leq 6$ (swapping m and n just gives complex conjugate filters); the response of the filters with $m = n = 0$ is the average intensity of the region. Figure 2(b) shows 8 of 15 filters.

Rotation changes the phase but not the magnitude of the response, therefore we use the modulus of each complex filter response.

Moment invariants are computed up to 2nd order and 2nd degree. The moments are computed for derivatives of an image patch with $M_{pq}^a = \frac{1}{xy} \sum_{x,y} x^p y^q [I_d(x,y)]^a$, where $p + q$ is the order, a is the degree and I_d is the image gradient in direction d . The derivatives are computed in x and y directions. This results in a 20-dimensional descriptor (2x10 without M_{00}^a). Note that originally moment invariants were computed on color images [43].

Cross correlation. To obtain this descriptor the region is smoothed and uniformly sampled. To limit the descriptor dimension we sample at 9x9 pixel locations. The similarity between two descriptors is measured with cross-correlation.

Distance measure. The similarity between descriptors is computed with the Mahalanobis distance for steerable filters, differential invariants, moment invariants and complex filters. We estimate one covariance matrix C for each combination of descriptor/detector; the same matrix is used for all experiments. The matrices are estimated on images different from the test data. We used 21 image sequences of planar scenes which are viewed under all the transformations for which we evaluate the descriptors. There are approximately 15 000 chains of corresponding regions with at least 3 regions per chain. An independently estimated homography is used to establish the chains of correspondences (cf. section III-C.1 for details on the homography estimation). We then compute the average over the individual covariance matrices of each chain. We also experimented with diagonal variance matrices and nearly identical results were obtained. The Euclidean distance is used to compare histogram based descriptors, that is SIFT, GLOH, PCA-SIFT, shape context and spin images. Note that the estimation of covariance matrices for descriptor normalization differs from the one used for PCA. For PCA, one covariance matrix is computed from approximately 47 000 descriptors.

C. Performance evaluation

1) *Data set:* We evaluate the descriptors on real images with different geometric and photometric transformations and for different scene types. Figure 3 shows example images of our data set¹ used for the evaluation. Six image transformations are evaluated: rotation (a) & (b); scale

¹The data set is available at <http://www.robots.ox.ac.uk/~vgg/research/affine>

change (c) & (d); viewpoint change (e) & (f); image blur (g) & (h); JPEG compression (i); and illumination (j). In the case of rotation, scale change, viewpoint change and blur, we use two different scene types. One scene type contains structured scenes, that is homogeneous regions with distinctive edge boundaries (e.g. graffiti, buildings) and the other contains repeated textures of different forms. This allows to analyze the influence of image transformation and scene type separately.

Image rotations are obtained by rotating the camera around its optical axis in the range of 30 and 45 degrees. Scale change and blur sequences are acquired by varying the camera zoom and focus respectively. The scale changes are in the range of 2-2.5. In the case of the viewpoint change sequences the camera position varies from a fronto-parallel view to one with significant foreshortening at approximately 50-60 degrees. The light changes are introduced by varying the camera aperture. The JPEG sequence is generated with a standard xv image browser with the image quality parameter set to 5%. The images are either of planar scenes or the camera position was fixed during acquisition. The images are therefore always related by a homography (plane projective transformation). The ground truth homographies are computed in two steps. First, an approximation of the homography is computed using manually selected correspondences. The transformed image is warped with this homography so that it is roughly aligned with the reference image. Second, a robust small baseline homography estimation algorithm is used to compute an accurate residual homography between the reference image and the warped image, with automatically detected and matched interest points [16]. The composition of the approximate and residual homography results in an accurate homography between the images.

In section IV we display the results for image pairs from figure 3. The transformation between these images is significant enough to introduce some noise in the detected regions. Yet, many correspondences are found and the matching results are stable. Typically, the descriptor performance is higher for small image transformations but the ranking remains the same. There are few corresponding regions for large transformations and the recall-precision curves are not smooth.

A data set different from the test data was used to estimate the covariance matrices for PCA and descriptor normalization. In both cases we have used 21 image sequences of different planar



Fig. 3. Data set. Examples of images used for the evaluation, (a)(b) Rotation, (c)(d) Zoom+rotation, (e)(f) Viewpoint change, (g)(h) Image blur, (i) JPEG compression, (j) Light change.

scenes which are viewed under all the transformations for which we evaluate the descriptors².

2) *Evaluation criterion:* We use a criterion similar to the one proposed in [19]. It is based on the number of correct matches and the number of false matches obtained for an image pair.

Two regions **A** and **B** are matched if the distance d between their descriptors $D_{\mathbf{A}}$ and $D_{\mathbf{B}}$ is below a threshold t . Each descriptor from the reference image is compared with each descriptor from the transformed one and we count the number of correct matches as well as the number of false matches. The value of t is varied to obtain the curves. The results are presented with *recall* versus *1-precision*. Recall is the number of correctly matched regions with respect to the number of corresponding regions between two images of the same scene:

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

The number of correct matches and correspondences is determined with the overlap error [30]. The overlap error measures how well the regions correspond under a transformation, here a homography. It is defined by the ratio of the intersection and union of the regions $\epsilon_S = 1 - (\mathbf{A} \cap \mathbf{H}^T \mathbf{B} \mathbf{H}) / (\mathbf{A} \cup \mathbf{H}^T \mathbf{B} \mathbf{H})$ where **A** and **B** are the regions and H is the homography between the images (cf. section III-C.1). Given the homography and the matrices defining the regions the error is computed numerically. Our approach counts the number of pixels in the union and the intersection of regions. Details can be found in [33]. We assume that a match is correct if the error in the image area covered by two corresponding regions is less than 50% of the region union, that is $\epsilon_S < 0.5$. The overlap is computed for the measurement regions which are used to compute the descriptors. Typically, there are very few regions with larger error that are correctly matched and these matches are not used to compute the recall. The number of correspondences (possible correct matches) are determined with the same criterion.

The number of false matches relative to the total number of matches is represented by 1-precision.

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

Given recall, 1-precision and the number of corresponding regions, the number of correct matches can be determined by $\# \text{ correspondences} \cdot recall$ and the number of false matches

²The data set is available at <http://www.robots.ox.ac.uk/~vgg/research/affine>

by $\#correspondences \cdot recall \cdot (1 - precision)/precision$. For example, there are 3708 corresponding regions between the images used to generate figure 4(a). For a point on the GLOH curve with recall of 0.3 and 1-precision of 0.6, the number of correct matches is $3708 \cdot 0.3 = 1112$, and the number of false matches is $3708 \cdot 0.3 \cdot 0.6 / (1 - 0.6) = 1668$. Note that recall and 1-precision are independent terms. Recall is computed with respect to the number of corresponding regions and 1-precision with respect to the total number of matches.

Before we start the evaluation we discuss the interpretation of figures and possible curve shapes. A perfect descriptor would give a recall equal to 1 for any precision. In practice, recall increases for an increasing distance threshold, as noise which is introduced by image transformations and region detection increases the distance between similar descriptors. Horizontal curves indicate that the recall is attained with a high precision and is limited by the specificity of the scene i.e. the detected structures are very similar to each other and the descriptor cannot distinguish them. Another possible reason for non-increasing recall is that the remaining corresponding regions are very different from each other (partial overlap close to 50%) and therefore the descriptors are different. A slowly increasing curve shows that the descriptor is affected by the image degradation (viewpoint change, blur, noise etc.). If curves corresponding to different descriptors are far apart and have different slopes, then the distinctiveness and robustness of the descriptors is different for a given image transformation or scene type.

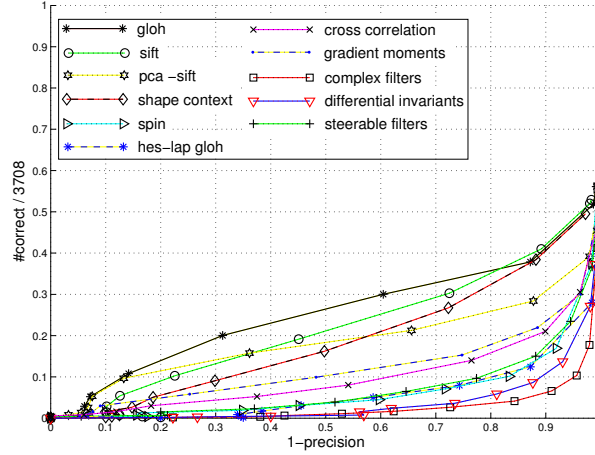
IV. EXPERIMENTAL RESULTS

In this section we present and discuss the experimental results of the evaluation. The performance is compared for affine transformations, scale changes, rotation, blur, jpeg compression and illumination changes. In the case of affine transformations we also examine different matching strategies, the influence of the overlap error and the dimension of the descriptor.

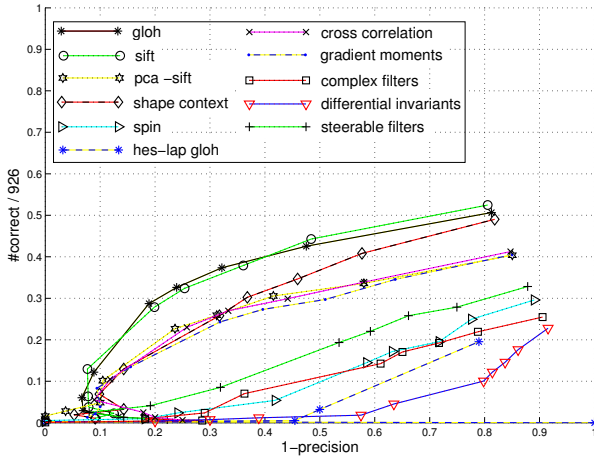
A. Affine transformations

In this section we evaluate the performance for viewpoint changes of approximately 50 degrees. This introduces a perspective transformation which can locally be approximated by an affine transformation. This is the most challenging transformation of the ones evaluated in this paper. Note that there are also some scale and brightness changes in the test images, see figure 3(e)(f). In the following we first examine different matching approaches. Second, we investigate the

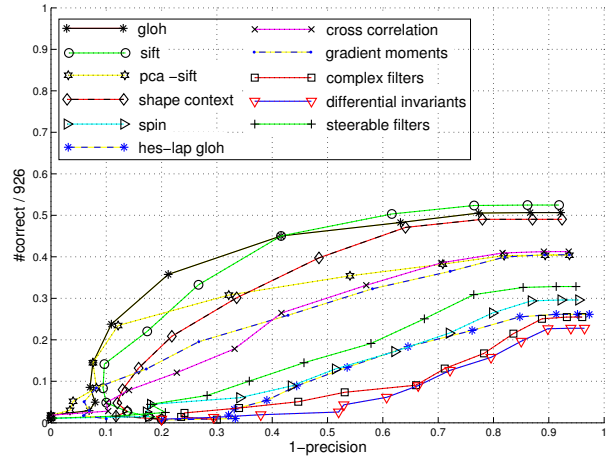
influence of the overlap error on the matching results. Third, we evaluate the performance for different descriptor dimensions. Fourth, we compare the descriptor performance for different region detectors and scene types.



(a)



(b)



(c)

Fig. 4. Comparison of different matching strategies. Descriptors computed on Hessian-Affine regions for images from figure 3(e). (a) Threshold based matching. (b) Nearest neighbor matching. (c) Nearest neighbor distance ratio matching. hes-lap gloh is the GLOH descriptor computed for Hessian-Laplace regions (cf. section IV-A.4).

1) *Matching strategies:* The definition of a match depends on the matching strategy. We compare three of them. In the case of threshold based matching two regions are matched if the distance between their descriptors is below a threshold. A descriptor can have several matches and several of them may be correct. In the case of nearest neighbor based matching two regions

\mathbf{A} and \mathbf{B} are matched if the descriptor $D_{\mathbf{B}}$ is the nearest neighbor to $D_{\mathbf{A}}$ and if the distance between them is below a threshold. With this approach a descriptor has only one match. The third matching strategy is similar to nearest neighbor matching except that the thresholding is applied to the distance ratio between the first and the second nearest neighbor. Thus, the regions are matched if $\|D_{\mathbf{A}} - D_{\mathbf{B}}\| / \|D_{\mathbf{A}} - D_{\mathbf{C}}\| < t$ where $D_{\mathbf{B}}$ is the first and $D_{\mathbf{C}}$ is the second nearest neighbor to $D_{\mathbf{A}}$. All matching strategies compare each descriptor of the reference image with each descriptor of the transformed image.

Figure 4(a)(b)(c) shows the results for the three matching strategies. The descriptors are computed on Hessian-Affine regions. The ranking of the descriptors is similar for all matching strategies. There are some small changes between nearest neighbor matching (NN) and matching based on the nearest neighbor distance ratio (NNDR). For low false positive rates in figures 4(a) and (b) PCA-SIFT obtains better scores than SIFT. In figure 4(c), which shows the results for NNDR, SIFT is significantly better than PCA-SIFT, whereas GLOH obtains a score similar to SIFT. Cross correlation and complex filters obtain slightly better scores than for threshold based and nearest neighbor matching. Moments perform as well as cross correlation and PCA-SIFT in the NNDR matching (cf. figure 4(c)).

The precision is higher for the nearest neighbor based matching (cf. figure 4(b) and (c)) than for the threshold based approach (cf. figure 4(a)). This is because the nearest neighbor is mostly correct, although the distance between similar descriptors varies significantly due to image transformations. Nearest neighbor matching selects only the best match and rejects all the others below the threshold therefore there are less false matches and the precision is high. Matching based on nearest neighbor distance ratio is similar but additionally penalizes the descriptors which have many similar matches, i.e. the distance to the nearest neighbor is comparable to the distances to other descriptors. This further improves the precision. The nearest neighbor based techniques can be used in the context of matching, however they are difficult to apply when descriptors are searched in a large database. The distance between descriptors is then the main similarity criterion. The results for distance threshold based matching reflect the distribution of the descriptors in the space, we therefore use this method for our experiments.

2) *Region overlap*: In this section we investigate the influence of the overlap error on the descriptor performance. Figure 5(a) displays recall with respect to overlap error. To measure the recall for different overlap errors we fix the distance threshold for each descriptor such that the

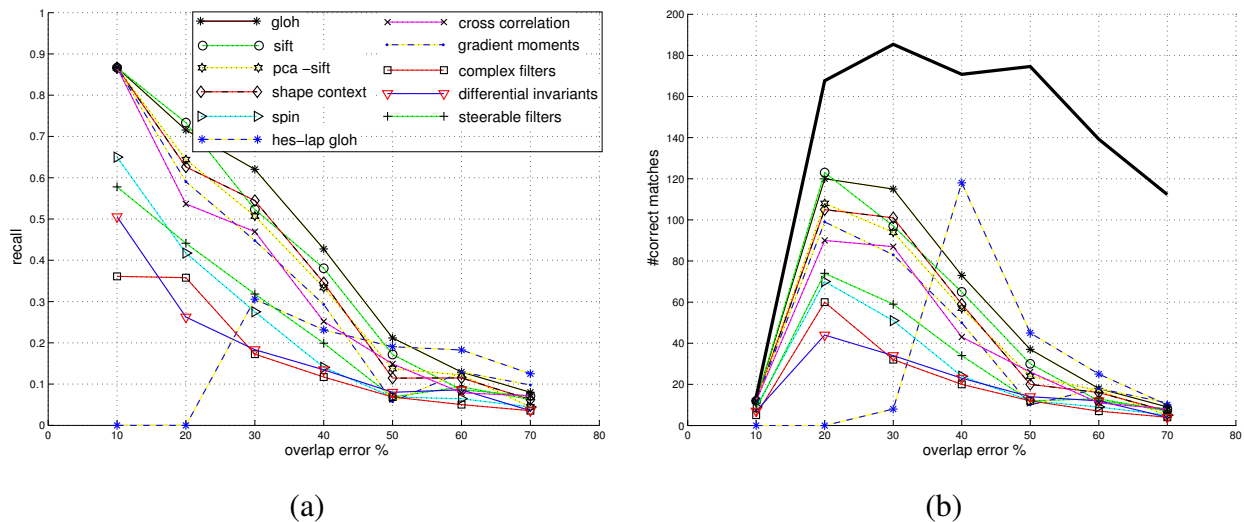


Fig. 5. Evaluation for different overlap errors. Test images are from figure 3(e) and descriptors are computed for Hessian-Affine regions. The descriptor thresholds are set to obtain precision=0.5. (a) Recall with respect to the overlap error. (b) Number of correct matches with respect to the overlap error. The bold line shows the number of Hessian-Affine correspondences.

precision is 0.5. Figure 5(b) shows the number of correct matches obtained for a false positive rate of 0.5 and for different overlap errors.

The number of correct matches as well as the number of correspondences is computed for a range of overlap errors, i.e. the score for 20% is computed for an overlap error larger than 10% and lower than 20%. As expected the recall decreases with increasing overlap error (cf. figure 5(a)). The ranking is similar to the previous results. We can observe that the recall for cross correlation drops faster than for other high dimensional descriptors, which indicates lower robustness of this descriptor to the region detector accuracy. We also show the recall for GLOH combined with scale invariant Hessian-Laplace detector (*hes-lap gloh*). The recall is zero up to an overlap error of 20 % as there are no corresponding regions for such small errors. The recall increases to 0.3 at 30% overlap and slowly decreases for larger errors. The recall for *hes-lap gloh* is slightly above the others because the large overlap error is mainly caused by size differences in the circular regions, unlike for affine regions where the error also comes from the affine deformations which significantly affect the descriptors.

Figure 5(b) shows the actual number of correct matches for different overlap errors. This figure also reflects the accuracy of the detector. The bold line shows the number of corresponding regions extracted with Hessian-Affine. There are few corresponding regions with an error below

10%, but nearly 90% of them are correctly matched with the SIFT based descriptors, PCA-SIFT, moments and cross correlation (cf. figure 5(a)). Most of the corresponding regions are located in the range of 10% and 60% overlap errors, whereas most of the correct matches are located in the range 10% to 40%. In the following experiments the number of correspondences is counted between 0% and 50% overlap error. We allow for 50% error because the regions with this overlap error can be matched if they are centered on the same structure, unlike the regions which are shifted and only partially overlapping. If the number of detected regions is high, the probability of an accidental overlap of two regions is also high, although they may be centered on different image structures. The large range of allowed overlap errors results in a large number of correspondences which also explains low recall.

3) *Dimensionality*: The derivatives-based descriptors and the complex filters can be computed up to an arbitrary order. Figure 6(a) displays the results for steerable filters computed up to 3rd and 4th order, differential invariants up to 2nd and 3rd order and complex filters up to 2nd and 6th order. This results in 5, 9 dimensions for differential invariants; 9, 14 dimensions for steerable filters; 9, 15 dimensions for complex filters. We used the test images from figure 3(e) and descriptors are computed for Hessian-Affine regions. Note that the vertical axes in figure 6 are scaled. The difference between steerable filters computed up to 3rd and up to 4th order is small but noticeable. This shows that the 3rd and 4th order derivatives are still distinctive. We can observe a similar behavior for different orders of differential invariants and complex filters. Steerable filters computed up to 3th order perform better than differential invariants computed up to the same order. The multiplication of derivatives necessary to obtain rotation invariance increases the instability.

Figure 6(b) shows the results for high dimensional, region-based descriptors (GLOH, PCA-SIFT and cross correlation). The GLOH descriptor is computed for 17 location bins and 16 orientations and the 128 largest eigenvectors are used ($g_{loh} = 128$). The performance is slightly lower if only 40 eigenvectors are used ($g_{loh} = 40$) and much lower for all 272 dimensions ($g_{loh} = 272$). A similar behavior is observed for PCA-SIFT and cross correlation. Cross correlation is evaluated for 36, 81, and 400 dimensions, i.e. 6x6, 9x9 and 20x20 samples and results are best for 81 dimensions (9x9). Figure 6(b) shows that the optimal number of dimensions in this experiment is 128 for GLOH, 36 for PCA-SIFT and 81 for cross correlation. In the following we use the number of dimensions which gave the best results here.

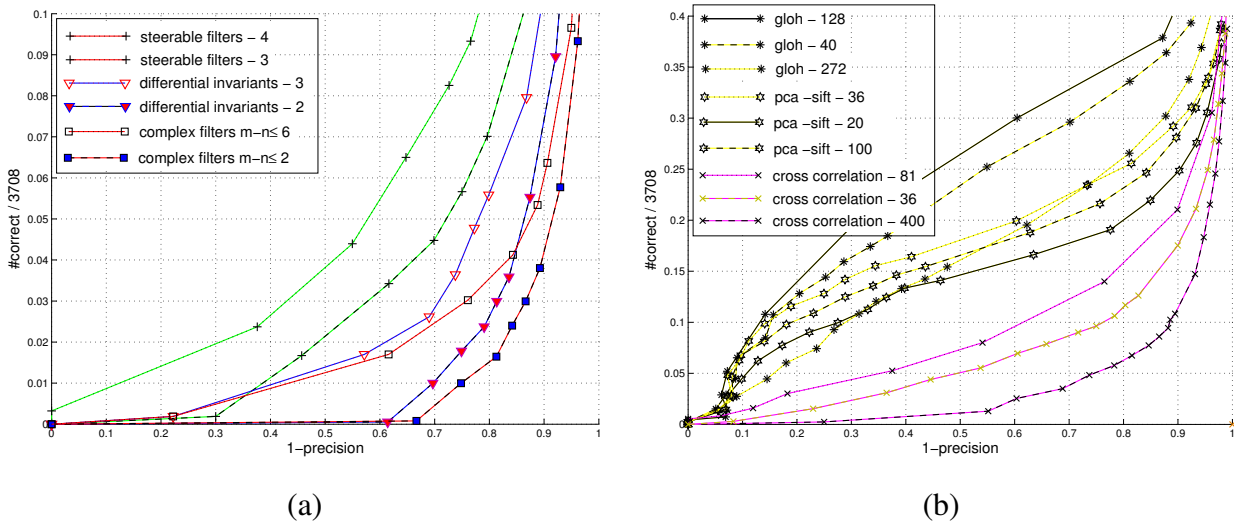


Fig. 6. Evaluation for different descriptor dimensions. Test images are from figure 3(e) and descriptors are computed for Hessian-Affine regions. (a) Low dimensional descriptors. (b) High dimensional, region-based descriptors.

Table I displays the sum of the first 10 eigenvalues and the sum of all eigenvalues for the descriptors. These eigenvalues result from PCA of descriptors normalized by their variance. The numbers given in table I correspond to the amount of variance captured by different descriptors, therefore to their distinctiveness. PCA-SIFT has the largest sum, followed by GLOH, SIFT and the other descriptors. Moments have the smallest value. This reflects the discriminative power of the descriptors, but the robustness is equally important. Therefore, the ranking of the descriptors can be different in other experiments.

4) *Region and scene types*: In this section we evaluate the descriptor performance for different affine region detectors and different scene types. Figures 7(a) and 7(b) show the results for the structured scene with Hessian-Affine and Harris-Affine regions, and figures 7(c) and (d) for the textured scene for Hessian-Affine and Harris-Affine regions respectively.

The recall is better for the textured scene (figure 7(c) and (d)) than for the structured one (figures 7(a) and (b)). The number of detected regions is significantly larger for the structured scene, which contains many corner-like structures. This leads to an accidental overlap between regions, therefore a high number of correspondences. This also means that the actual number of correct matches is larger for the structured scene. The textured scene contains similar motifs, however the regions capture sufficiently distinctive signal variations. The difference in performance of SIFT based descriptors and others is larger on the textured scene which indicates that a

Descriptor	$\sum_{i=1:10} eigenvalue(i)$	$\sum_i eigenvalue(i)$
PCA-SIFT	1.0839e+12	1.9743e+12
GLOH	1.4085e+11	2.8277e+11
SIFT	3.4210e+09	6.4541e+09
Shape context	3.3582e+09	7.1149e+09
Spin images	4.4791e+09	5.2355e+09
Cross correlation	1.0657e+09	1.4076e+09
Steerable filters	4.1529e+07	4.2909e+07
Differential invariants	2.5970e+07	2.6349e+07
Complex filters	1.6328e+07	1.8264e+07
Moments	1.3829e+07	1.8100e+07

TABLE I

DISTINCTIVENESS OF THE DESCRIPTORS. SUM OF THE FIRST 10 AND SUM OF ALL EIGENVALUES FOR DIFFERENT DESCRIPTORS.

large discriminative power is necessary to match them. Note that the GLOH descriptor performs best on the structured scene and SIFT obtains the best results for the textured images.

Descriptors computed for Harris-Affine regions (see figure 7(d)) give slightly worse results than those computed for Hessian-Affine regions (see figure 7(c)). This is observed for both, structured and textured scenes. The method for scale selection and for affine adaptation is the same for Harris and Hessian based regions. However, as mentioned in section III-A, the Laplacian based scale selection combined with the Hessian detector gives more accurate results.

Note that GLOH descriptors computed on scale invariant regions perform worse than many other descriptors (see `hes-lap gloh` and `har-lap gloh` in figure 7), as these regions and therefore the descriptors are only scale and not affine invariant.

B. Scale changes

In this section we evaluate the descriptors for combined image rotation and scale change. Scale changes lie in the range 2-2.5 and image rotations in the range $30^\circ - 45^\circ$. Figure 8(a) shows the performance of descriptors computed for Hessian-Laplace regions detected on a structured scene (see figure 3(c)), and figure 8(c) on a textured scene (see figure 3(d)). Harris-Laplace regions are used in figures 8(b)(d). We can observe that GLOH gives the best results on Hessian-Laplace

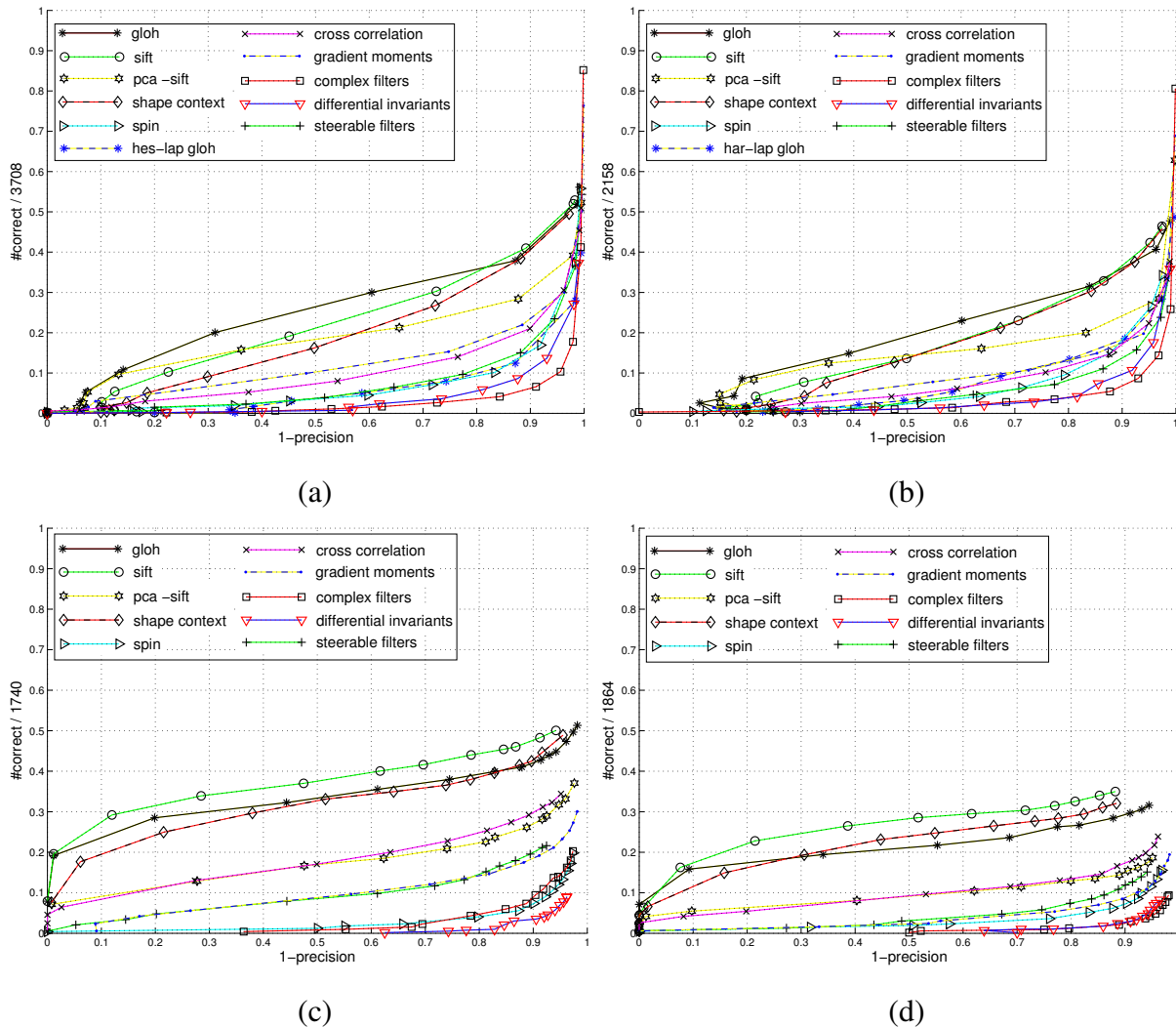


Fig. 7. Evaluation for a viewpoint changes of 40 – 60 degrees. (a) Results for a structured scene, cf. figure 3(e) with Hessian-Affine regions. (b) Results for a structured scene, cf. figure 3(e) with Harris-Affine regions. (c) Results for a textured scene, cf. figure 3(f), Hessian-Affine regions. (d) Results for a textured scene, cf. figure 3(f), Harris-Affine regions. *har-lap gloh* is the GLOH descriptor computed for Harris-Laplace regions. *hes-lap gloh* is the GLOH descriptor computed for Hessian-Laplace regions.

regions. In the case of Harris-Laplace SIFT and shape context obtain better results than GLOH if 1-precision is larger than 0.1. The ranking for other descriptors is similar.

We can observe that the performance of all descriptors is better than in the case of viewpoint changes. The regions are more accurate since there are less parameters to estimate. As in the case of viewpoint changes the results are better for the textured images. However, the number of corresponding regions is 5 times larger for Hessian-Laplace and 10 times for Harris-Laplace

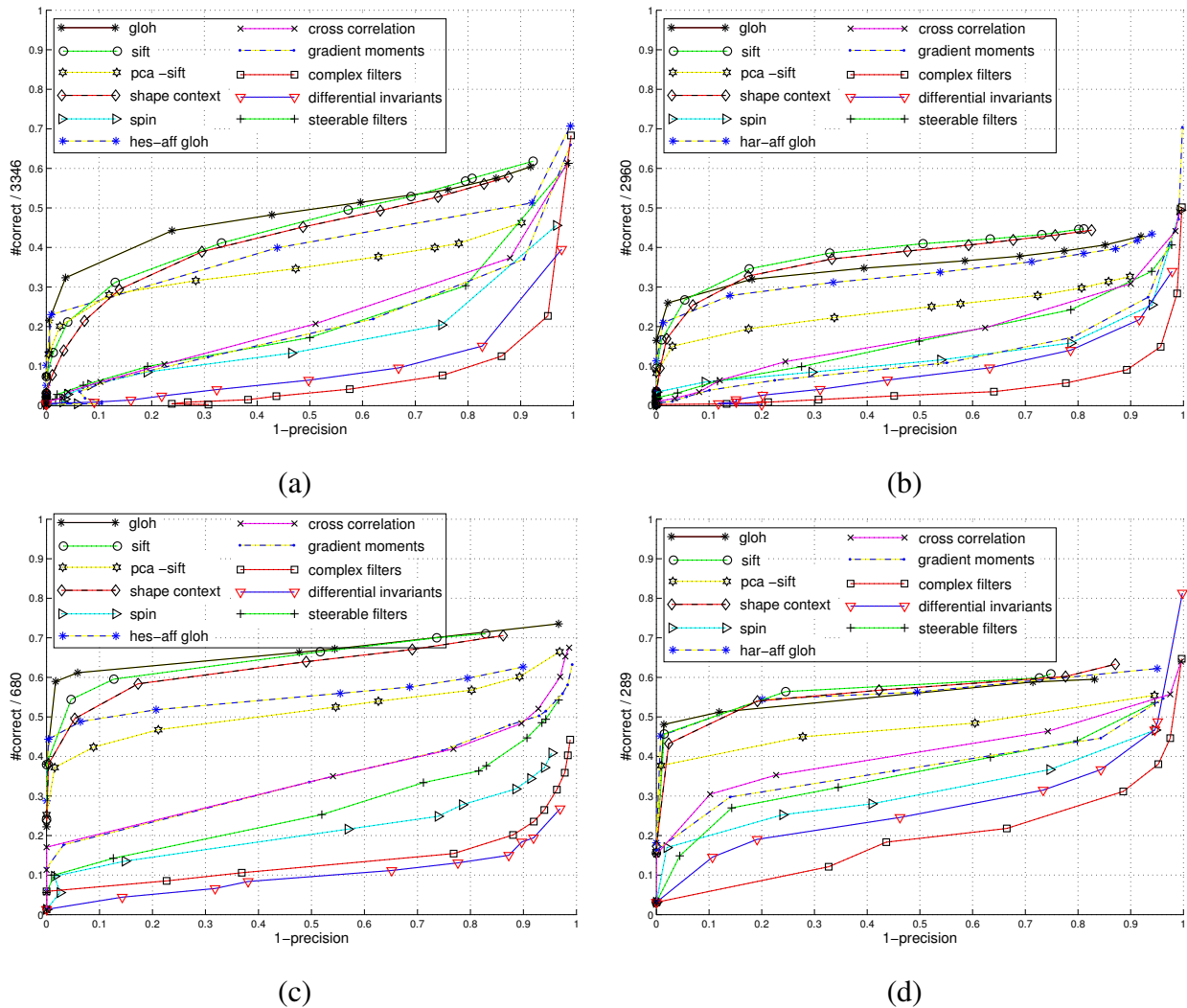


Fig. 8. Evaluation for scale changes of a factor 2-2.5 combined with an image rotation of $30^\circ - 45^\circ$. (a) Results for a structured scene, cf. figure 3(c) with Hessian-Laplace regions. (b) Results for a structured scene, cf. figure 3(c) with Harris-Laplace regions. (c) Results for a textured scene, cf. figure 3(d) with Hessian-Laplace regions. (d) Results for a textured scene, cf. figure 3(d) with Harris-Laplace regions. *hes-aff gloh* is the GLOH descriptor computed for Hessian-Affine regions and *har-aff gloh* is the GLOH descriptor computed for Harris-Affine regions.

on the structured scene than on the textured one.

GLOH descriptors computed on affine invariant regions detected by Harris-Affine (*har-aff gloh*) and Hessian-Affine (*hes-aff gloh*) obtain slightly lower scores than SIFT based descriptors computed on scale invariant regions, but they perform better than all the other descriptors, unlike on images with viewpoint changes (cf. figures 4 and 7). This is observed for both structured and textured scenes. This shows that affine invariant detectors can also be

used in the presence of scale changes if combined with an appropriate descriptor.

C. Image rotation

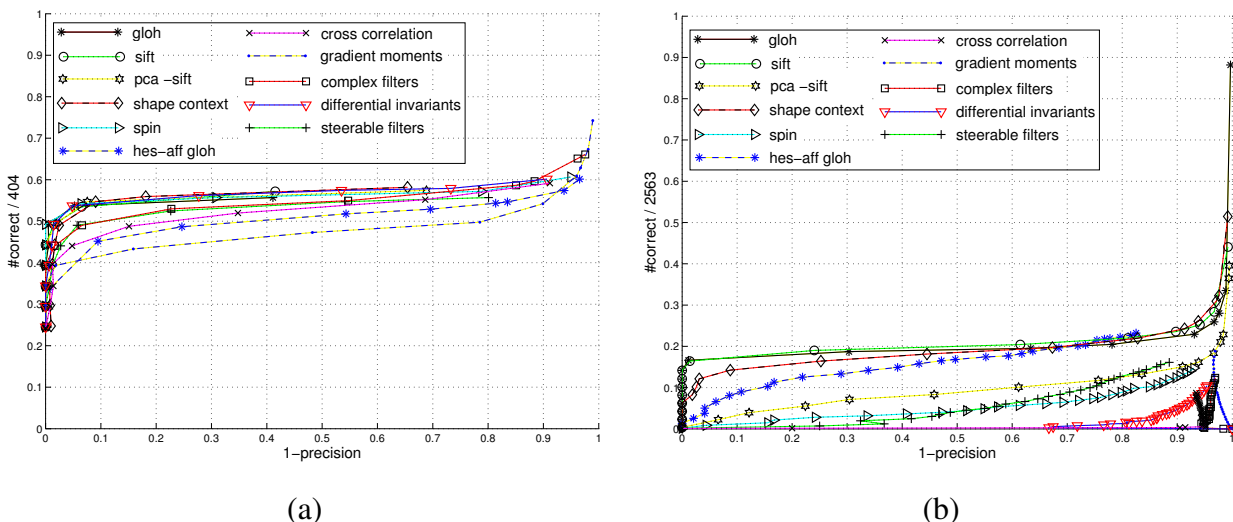


Fig. 9. Evaluation for an image rotation of $30^\circ - 45^\circ$. Descriptors computed for Harris points. *hes-aff gloh* - GLOH descriptor computed for Hessian-Affine regions. (a) Results for the structured images from figure 3(a). There are 1671 correspondences for Hessian-Affine. (b) Results for the textured images from figure 3(b). There are 1671 correspondences for Hessian-Affine.

To evaluate the performance for image rotation we used images with a rotation angle in the range between 30 and 45 degrees. This represents the most difficult case. In figure 9(a) we compare the descriptors computed for standard Harris points detected on a structured scene (cf. figure 3(a)). All curves are horizontal at similar recall values, i.e. all descriptors have a similar performance. Note that moments obtain a low score for this scene type. The applied transformation (rotation) does not affect the descriptors. The recall is below 1 because many correspondences are established accidentally. Harris detector finds many points close to each other and many support regions accidentally overlap due to the large size of the region (41 pixels).

To evaluate the influence of the detector errors we display the results for the GLOH descriptor computed on Hessian-Affine regions (*hes-aff gloh*). The performance is insignificantly lower than for descriptors computed of fixed size patches centered on Harris points. The number of correct matches is higher for the affine invariant detector. There are three types of error that

influence the descriptors computation: the region error, the localization error and the error of the estimated orientation angle. In the case of standard Harris the scale and therefore the patch size remains fixed. The only noise comes from the inaccuracy of the localization and from the angle estimation. We notice in figure 9(a) that these errors have less impact on descriptor performance than the region error which occurs in the case of Hessian-Affine. The error due to the orientation estimation is small since the rotation invariant descriptors do not perform better than the non-invariant ones.

Figure 9(b) presents the results for scanned text displayed in figure 3(b). The rank of the descriptors changes. GLOH, SIFT and shape context obtain the best results. Moments, differential invariants, cross correlation and complex filters fail on this example. The precision is low for all the descriptors. The descriptors do not capture small variations in texture which results in many false matches. GLOH descriptor computed on affine invariant regions (*hes-aff gloh*) performs well, i.e. lower than on Harris point, but better than most of the other descriptors.

D. Image blur

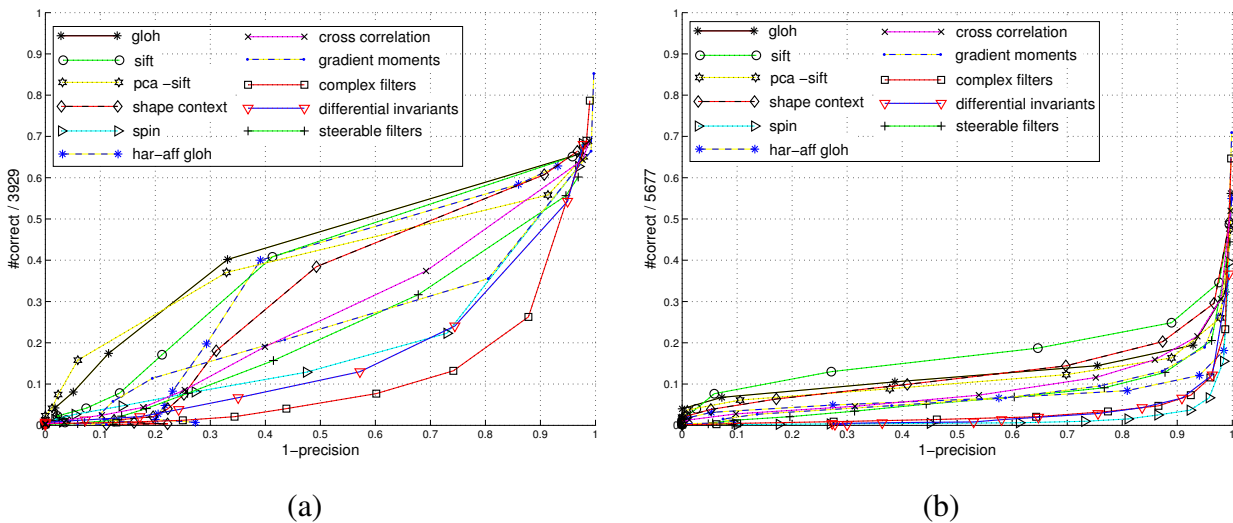


Fig. 10. Evaluation for blur. Descriptors are computed on Hessian-Affine regions. (a) Results for a structured scene, cf. figure 3(g) (*har-aff gloh* - Harris-Affine regions, 1125 correspondences). (b) Results for a textured scene, cf. figure 3(h) (*har-aff gloh* - Harris-Affine regions, 6197 correspondences).

In this section the performance is measured for images with a significant amount of blur. Blur was introduced by changing the camera focus. Figure 10(a) shows the results for the structured

scene and figure 10(b) for the textured scene. The images are displayed in figure 3(g) and (h), respectively. Results are presented for regions detected with Hessian-Affine. We also show the results for GLOH computed on Harris-Affine regions (`har-aff gloh`).

The results show that all descriptors are affected by this type of image degradation, although there are no geometric transformations in these images. The pixel intensities and the shape of local structures change in an unpredictable way and the descriptors are not robust to such deformations. It is difficult to model these deformations, therefore the comparisons on artificially generated data are frequently overly optimistic.

GLOH and PCA-SIFT give the highest scores. The performance of shape context, which is based on edges, decreases significantly compared to geometric changes (sections IV-A and IV-B). The edges disappear in the case of a strong blur.

GLOH computed on Harris-Affine regions obtains a significantly lower score than on Hessian-Affine regions. Blur has a larger influence on the performance of the Harris-Affine detector than on the performance of the Hessian-Affine detector. Similar observations were made in [33].

The results for the textured scene (cf. figure 10(b)) are even more influenced by blurring. The descriptors cannot distinguish the detected regions since blurring makes them nearly identical. SIFT gives the largest number of matches in this scene. Cross-correlation obtains the lowest score among the high dimensional descriptors but higher than low dimensional ones.

E. JPEG compression

In figure 11 we evaluate the influence of JPEG compression for a structured scene (cf. figure 3(i)). The quality of the transformed image is 5% of the reference one. Results are presented for regions detected with Hessian-Affine.

The performance of descriptors is better than in the case of blur (cf. section IV-D), but worse than in case of rotation and scale changes of structured scenes (cf. sections IV-C and IV-B). The performance gradually increases with decreasing precision for all descriptors, i.e. all descriptors are affected by JPEG artifacts. PCA-SIFT obtains the best score for a low false positive rate and SIFT for a false positive rate above 0.2. The results for GLOH lie in between those two descriptors.

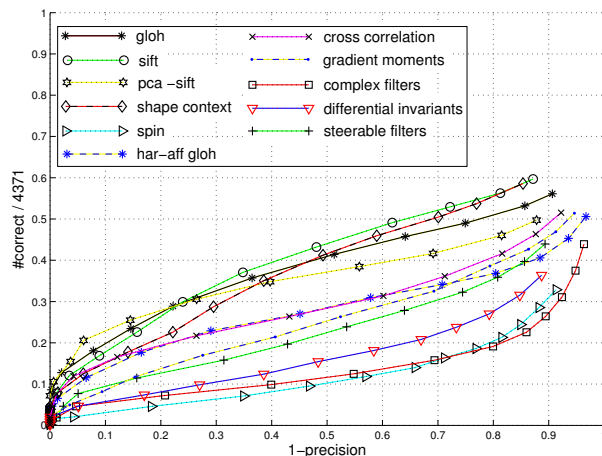


Fig. 11. Evaluation for JPEG compression, cf. figure 3(i). Descriptors are computed on Hessian-Affine regions (har-aff gloh - Harris-Affine regions, 4142 correspondences).

F. Illumination changes

Figure 12 shows the results for illumination changes which have been obtained by changing the camera settings. The image pair is displayed in figure 3(j). The descriptors are computed for Hessian-Affine regions. Figure 12(a) compares two approaches to obtain affine illumination invariance for differential descriptors: (i) based on region normalization (steerable filters and diff. invariant used in all our comparisons), (ii) based on the invariance of the descriptors (invariant steerable filters and invariant differential invariants), see section III-A for details. We observe that the descriptors computed on normalized regions are significantly better. Theoretically, the two methods are equivalent. However, the ratio of derivatives amplifies the noise due to region and location errors as well as non-affine illumination changes. The importance of affine illumination invariance is shown by the comparison with descriptors which are not intensity normalized (not invariant steerable filters, not invariant differential invariants). These descriptors obtain worse results. The score is not zero because these descriptors are based on derivatives which eliminate the constant factor from the intensity.

In figure 12(b) the standard descriptors are compared in the presence of illumination changes. All the descriptors are computed on normalized image patches. GLOH obtains the best matching score. The same descriptor computed on Harris-Affine regions obtains an equivalent score.

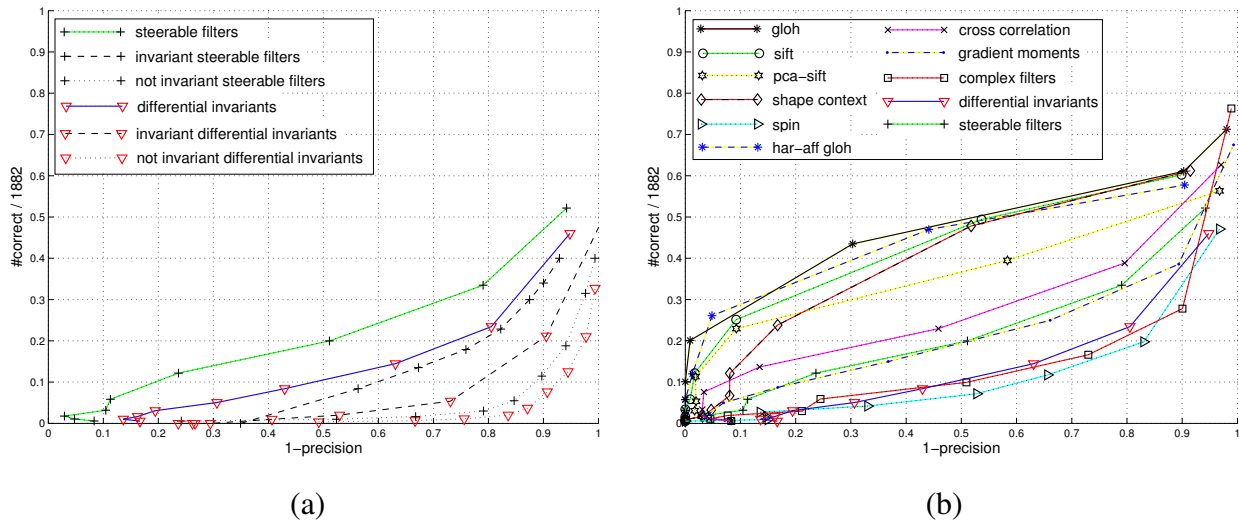


Fig. 12. Evaluation for illumination changes, cf. figure 3(j). The descriptors are computed for Hessian-Affine regions (har-aff gloh - Harris-Affine regions, 1120 correspondences). (a) Illumination invariance of differential descriptors. steerable filters and differential invariants are the standard descriptors computed on the intensity normalized regions. invariant steerable filters and invariant differential invariants are the illumination invariants and not invariant steerable filters and not invariant differential invariants are not intensity normalized. (b) Descriptors computed on illumination normalized regions.

G. Matching example

This section illustrates a matching example for images with a viewpoint change of more than 50° , see figure 13. Hessian-Affine detects 2511 and 2337 regions in the left and right image respectively. There are 747 correspondences identified by the overlap criterion defined in section III-C. For the 400 nearest neighbor matches obtained with the GLOH descriptor, 192 are correct (displayed in yellow in figure 13) and 208 are false (displayed in blue).

Table II presents recall, false positive rate and the number of correct matches obtained with different descriptors. These results are all based on a fixed number of 400 nearest neighbor matches. GLOH obtains the highest recall - 0.25, a slightly lower score is obtained by SIFT and shape context. Complex filters achieve the lowest score - 0.06. The number of correct matches vary from 192 to 44. There are approximately 4.4 times less correct matches for complex filters than for GLOH. This clearly shows the advantage of SIFT-based descriptors.



Fig. 13. Matching example. There are 400 nearest neighbor matches obtained with the GLOH descriptor on Hessian-Affine regions. There are 192 correct matches (yellow) and 208 false matches (blue).

Descriptor	recall	1-precision	#nearest neighbor correct matches
GLOH	0.25	0.52	192
SIFT	0.24	0.56	177
Shape context	0.22	0.59	166
PCA-SIFT	0.19	0.65	139
Moments	0.18	0.67	133
Cross correlation	0.15	0.72	113
Steerable filters	0.12	0.78	90
Spin images	0.09	0.84	64
Differential invariants	0.07	0.87	54
Complex filters	0.06	0.89	44

TABLE II

RECALL, 1-PRECISION AND NUMBER OF CORRECT MATCHES OBTAINED WITH DIFFERENT DESCRIPTORS FOR A FIXED NUMBER OF 400 NEAREST NEIGHBOR MATCHES ON THE IMAGE PAIR DISPLAYED IN FIGURE 13. THE REGIONS ARE DETECTED WITH HESSIAN-AFFINE.

V. DISCUSSION AND CONCLUSIONS

In this paper we have presented an experimental evaluation of interest region descriptors in the presence of real geometric and photometric transformations. The goal was to compare descriptors computed on regions extracted with recently proposed scale and affine-invariant detection techniques. Note that the evaluation was designed for matching and recognition of the same object or scene.

In most of the tests GLOH obtains the best results, closely followed by SIFT. This shows the robustness and the distinctive character of the region-based SIFT descriptor. Shape context also shows a high performance. However, for textured scenes or when edges are not reliable its score is lower.

The best low dimensional descriptors are gradient moments and steerable filters. They can be considered as an alternative when the high dimensionality of the histogram-based descriptors is an issue. Differential invariants give significantly worse results than steerable filters, which is surprising as they are based on the same basic components (Gaussian derivatives). The multiplication of derivatives necessary to obtain rotation invariance increases the instability.

Cross correlation gives unstable results. The performance depends on the accuracy of interest point and region detection, which decreases for significant geometric transformations. Cross correlation is more sensitive to these errors than other high dimensional descriptors.

Regions detected by Hessian-Laplace and Hessian-Affine are mainly blob-like structures. There are no significant signal changes in the center of the blob therefore descriptors perform better on larger neighborhoods. The results are slightly but systematically better on Hessian regions than on Harris regions due to their higher accuracy.

The ranking of the descriptors is similar for different matching strategies. We can observe that SIFT gives relatively better results if nearest neighbor distance ratio is used for thresholding. Note that the precision is higher for nearest neighbor based matching than for threshold based matching.

Obviously, the comparison presented here is not exhaustive and it would be interesting to include more scene categories. However, the comparison seems to indicate that robust region-based descriptors perform better than point-wise descriptors. Correlation is the simplest region-based descriptor. However, our comparison has shown that it is sensitive to region errors. It would be interesting to include correlation with patch alignment which corrects for these errors

and to measure the gain obtained by such an alignment. Of course this is very time consuming and should only be used for verification.

Similar experiments should be conducted for recognition and classification of object and scene categories. An evaluation of the descriptors in the context of texture classification and classification of similar local structures will be a useful and valuable addition to our work. This would probably imply clustering of local structures based on the descriptors and an evaluation of these clusters. It would be also interesting to compare the SIFT based descriptors in the evaluation framework proposed in [21, 44].

ACKNOWLEDGMENT

This research was supported by the European FET-open project VIBES and the European project LAVA (IST-2001-34405).

REFERENCES

- [1] A. Ashbrook, N. Thacker, P. Rockett, and C. Brown. Robust recognition of scaled shapes using pairwise geometric histograms. In D. Pycock, editor, *Proceedings of the sixth British Machine Vision Conference, Birmingham, UK*, pages 503–512, 1995.
- [2] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 774–781, 2000.
- [3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.
- [4] M. Brown and D. Lowe. Recognising panoramas. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*, pages 1218–1227, 2003.
- [5] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.
- [6] G. Carneiro and A. D. Jepson. Phase-based local features. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 282–296, 2002.
- [7] H. I. Christensen and P. J. Phillips, editors. *Empirical Evaluation Methods in Computer Vision*, volume 50 of *Series in Machine Perception and Artificial Intelligence*. World Scientific Publishing Co., 2002.
- [8] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*, pages 634–640, 2003.
- [9] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, pages 264–271, 2003.

- [10] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *Proceedings of the 8th European Conference on Computer Vision, Prague, Tcheque Republic*, pages 40–54, 2004.
- [11] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. General intensity transformations and second order invariants. In *Proceedings of the 7th Scandinavian Conference on Image Analysis, Aalborg, Denmark*, pages 338–345, 1991.
- [12] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.
- [13] D. Gabor. Theory of communication. *Journal I.E.E.*, 3(93):429 – 457, 1946.
- [14] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d’intérêt pour la couleur. In *12ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, pages 257–266, 2000.
- [15] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.
- [16] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
- [17] A. Johnson and M. Hebert. Object recognition by matching oriented points. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pages 684–689, 1997.
- [18] T. Kadir, M. Brady, and A. Zisserman. An affine invariant method for selecting salient regions in images. In *Proceedings of the 8th European Conference on Computer Vision, Prague, Tcheque Republic*, pages 345–457, 2004.
- [19] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Washington, USA*, pages 511–517, 2004.
- [20] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.
- [21] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, pages 319–324, 2003.
- [22] B. Leibe and B. Schiele. Interleaved Object Categorization and Segmentation. In *Proceedings of the 14th British Machine Vision Conference, Norwich, UK*, pages 759–768, 2003.
- [23] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
- [24] 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 and Vision Computing*, 15(6):415–434, 1997.
- [25] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*,

- 2(60):91–110, 2004.
- [26] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 1150–1157, 1999.
 - [27] J. K. M. Vetterli. *Wavelets and Subband Coding*. Prentice Hall, 1995.
 - [28] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the 13th British Machine Vision Conference, Cardiff, UK*, pages 384–393, 2002.
 - [29] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001.
 - [30] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 128–142, 2002.
 - [31] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, pages 257–264, 2003.
 - [32] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 1(60):63–86, 2004.
 - [33] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. Submitted to *International Journal of Computer Vision*.
 - [34] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
 - [35] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of the 8th European Conference on Computer Vision, Prague, Tcheque Republic*, pages 71–84, 2004.
 - [36] T. Randen and J. H. Husoy. Filtering for texture classification : A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, 1999.
 - [37] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 414–431, 2002.
 - [38] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534, 1997.
 - [39] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.
 - [40] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *International Conference on Intelligent Robots and Systems, IROS 2002, Lausanne, Switzerland*, pages 226–231, 2002.
 - [41] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*. pages 1470–1478, 2003.
 - [42] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 1(59):61–85, 2004.

- [43] L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *Proceedings of the 4th European Conference on Computer Vision, Cambridge, UK*, pages 642–651, 1996.
- [44] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, pages 477–484, 2003.
- [45] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, pages 151–158, 1994.