A performance evaluation of local descriptors

K. Mikolajczyk C. Schmid

INRIA Rhône-Alpes, GRAVIR-CNRS
655, av. de l'Europe, 38330 Montbonnot, France
Krystian.Mikolajczyk,Cordelia.Schmid@inrialpes.fr

Abstract

In this paper we compare the performance of interest point descriptors. 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 point detector. The descriptors should be distinctive and at the same time robust to changes in viewing conditions as well as to errors of the point detector. Our evaluation uses as criterion detection rate with respect to false positive rate and is carried out for different image transformations. We compare SIFT descriptors [11], steerable filters [5], differential invariants [10], complex filters [17], moment invariants [21] and cross-correlation for different types of interest points [8, 11, 13, 14]. In this evaluation, we observe that the ranking of the descriptors does not depend on the point detector and that SIFT descriptors perform best. Steerable filters come second ; they can be considered a good choice given the low dimensionality.

1. Introduction

Local photometric descriptors computed at interest points have proved to be very successful in applications such as matching and recognition [11, 13, 16, 18]. 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 construct invariant "image regions" which are then used as support regions to compute invariant descriptors. For example, Mikolajczyk and Schmid [14] have developed affine invariant interest points with associated affine invariant regions. Tuytelaars and Van Gool [20] construct two types of affine invariant regions, one based on the combination of interest points and edges and the other based on image intensities. Lowe [11] proposes scale-invariant regions based on local extrema in scale-space built with difference-of-Gaussian (DoG) filters. Given invariant regions, the remaining questions are which is the most appropriate descriptor to characterize these regions, and does the choice of the descriptor depend on the region detector. These questions will be addressed in our paper.

To evaluate local descriptors we use the ROC (receiver operating characteristics) of the detection rate for a query image with respect to the false positive rate in a database of images. The evaluation is carried out for different descriptors, different interest point detectors and in the presence of different image transformations.

1.1. Related work

Performance evaluation has gained more and more importance in computer vision [3]. In the context of matching and recognition several authors have evaluated interest point detectors [7, 14, 19]. 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. The only previous work which evaluates the performance of point descriptors is by Carneiro and Jepson [2]. 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.

Local descriptors also called filters have been evaluated in the context of texture classification [15]. However, the results can not be directly transposed to point descriptors and there are point descriptors which have not been used for texture classification.

1.2. Overview

In section 2 we present a state of the art on local descriptors. Section 3 gives the implementation details for the detectors and descriptors used in our comparison as well as the experimental conditions. In section 4 we present the experimental results. In the last section we discuss the results.

2. Descriptors

Many different techniques for describing local image regions have been developed. The simplest descriptor is a vector of image pixels. The cross-correlation measure can then be used to compute a similarity score between two regions. However, the high dimensionality of such a description increases the computational complexity of recognition.

Therefore, this technique is mainly used for finding point-to-point correspondences between two images. The point neighborhood can be sub-sampled to reduce the dimension.

Distribution based descriptors. A simple descriptor is the distribution of the pixel intensities which can be represented by a histogram. A more expressive representation was introduced by Johnson and Hebert [9] in the context of 3D object recognition. Their representation (spin image) is generated using a histogram of the relative position of neighborhood points to the interest point in 3D space.

Non-parametric transformations. An approach, interesting for its robustness to illumination changes, was developed by Zabih and Woodfill [22]. It relies on local transforms based on non-parametric statistics, which use the information about ordering and reciprocal relations between the data, rather than the data values themselves. A small region is described by ordered binary relations of the intensities at neighboring points.

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 [6] overcomes these problems but a large number of Gabor filters is required to capture small changes in frequency and orientation, that is the description is high dimensional. Gabor filters and wavelets [12] 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 [10]. Florack et al. [4] derived differential invariants, which combine components of the *local jet* to obtain rotation invariance. Freeman and Adelson [5] 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.

Baumberg [1] and Schaffalitzky and Zisserman [17] 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 3.2). These filters differ from the Gaussian derivatives by a linear coordinates change in filter response space.

Other techniques. Lowe [11] proposed a descriptor in which a point neighborhood is represented with multiple images. These images are orientation planes representing a number of gradient orientations. Each image contains only the gradients corresponding to one orientation. Each orientation plane is blurred and re-sampled to allow for small shifts in positions of the gradients. This description provides robustness against localization errors and small geometric distortions.

Generalized moment invariants have been introduced by Van Gool et al. [21] to describe the multi-spectral nature of the data. The invariants combine central moments defined by: $M^a_{pq} = \int\int_\Omega x^p y^q [I(x, y)]^a dxdy$ with order $p + q$ and degree a. These moments are independent and can be easily computed for any order and degree. The moments characterize the shape and the intensity distribution in a limited region Ω.

3. Experimental setup

In the following we first describe the interest point detectors used in our comparison and the normalization of the associated regions, on which the descriptors are computed. We then give implementation details for the evaluated descriptors. Finally we discuss the evaluation criteria and the image data used in the tests.

3.1. Support regions

Point detectors. The point detectors determine the regions which are used to compute the descriptors. In this evaluation we have used four interest point detectors:

Harris points [8] are invariant to rotation. The support region is a fixed size neighborhood centered at the interest point.

Harris-Laplace points [13] are invariant to rotation and scale changes. The points are detected by a scale adapted Harris function and selected in scale-space by the Laplacian operator. The selected scale determines the size of the support region.

DoG points [11] are invariant to rotation and scale changes. The points are local scale-space maxima of the difference-of-Gaussians. The selected scale determines the size of the support region.

Harris-Affine points [14] are invariant to affine image transformations. Localization and scale are estimated in a similar way as in the Harris-Laplace detector and the affine invariant neighborhood is determined by the eigenvalues of the second moment matrix.

The code of the authors has been used for Harris-Laplace, DoG and Harris-Affine detectors. The thresholds used for each detector were constant for all the experiments. Given an image, the number of detected points is approximately equal for these detectors, on average 300 points. These are obtained with thresholds empirically chosen for each detector.

Region normalization. All regions are mapped to a circular region of constant radius to obtain scale and affine invariance for each descriptor. The normalized region diameter is 45 (2*22+1) pixels. The radius of the point neighborhood is 5 times bigger than the scale at which the interest point is detected. Point neighborhoods which are larger than the normalized region, are smoothed before the size normalization. The parameter σ of the smoothing Gaussian kernel is given by the ratio detected/normalized region size. Differential invariants and the complex filters are invariant to rotation. To obtain rotation invariance for the other descriptors the regions are rotated in the direction of the average gradient orientation, which is computed within a small point neighborhood. The geometric normalization of the image patch uses bilinear interpolation.

We evaluate two approaches to compensate for affine illumination changes of the pixel intensities $(aI(\mathbf{x}) + b)$. We can either normalize the image patch or compute invariant descriptors. If not stated otherwise, we normalize the image patch by the mean and the standard deviation of the pixel intensities within the point neighborhood: $I'(\mathbf{x}) = (I(\mathbf{x}) - \texttt{mean}(I))/\texttt{stdev(I)}$. Illumination invariants can be computed for derivative-based descriptors. 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 an appropriate power.

3.2. Descriptors

In the following we present the implementation details for the descriptors used in our experimental evaluation. We use five different region descriptors: SIFT [11], steerable filters [5], differential invariants [10], complex filters [17] and moment invariants [21]. In order to compare the performance of the descriptors to simple cross-correlation of image patches we include a descriptor based on sampled pixel values.

SIFT descriptors are computed on image patches with the code provided by Lowe [11]. He uses 8 orientation planes. For each orientation the gradient image is sampled over a 4x4 grid of locations. The descriptor is of dimension 128. The description vector is divided by the square root of the sum of squared components to obtain illumination invariance.

Steerable filters and *differential invariants* are computed with Gaussian derivatives. Changing the orientation of derivatives as proposed in [5] gives equivalent results to computing the local jet on rotated image patches. We use the second approach and apply Gaussian kernels with $\sigma = 7$ in the image patch of size 45. The derivatives are computed up to 4th order, that is the descriptor has dimension 13 (1+3+4+5). The differential invariants are computed up to 3rd order (dimension 8). The gradient magnitude is the first component of both descriptors.

Complex filters are derived from the following equation $K_{mn}(x, y) = (x+iy)^m(x-iy)^n G(x, y)$. Our experiments have shown that the complex filters provide better results if they are not weighted by the Gaussian function $G(x, y)$. This effect was also observed by Schaffalitzky. The kernels are computed for a unit disk of radius 1 and sampled at 45x45 locations. The code of the authors [17] has been used for generating the kernels. We use 15 filters defined by $m + n \leq 6$ (swapping m and n just gives complex conjugate filters); $m = n = 0$ gives the average intensity of the region. Rotation influences 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 descriptor is 10-dimensional (without M_{00}^{a}).

Cross correlation. To obtain this descriptor the point neighborhood is smoothed and uniformly sampled. To limit the descriptor dimension we sample at 13x13 pixel locations. The similarity between two descriptors is measured with cross-correlation.

Distance measure. The similarity between descriptors is computed with the Mahalanobis distance except for SIFT and cross correlation. We estimate one covariance matrix for each combination of descriptor/detector; the same matrix is used for all experiments. The matrices are estimated on a data set different from the test data. We use images of planar scenes which are viewed under all the transformations for which we evaluate the descriptors. The homography is used to establish point-to-point correspondences. We then compute the average over these individual point-based covariance matrices. The SIFT descriptors are compared with the Euclidean distance as proposed in [11].

3.3. Performance evaluation

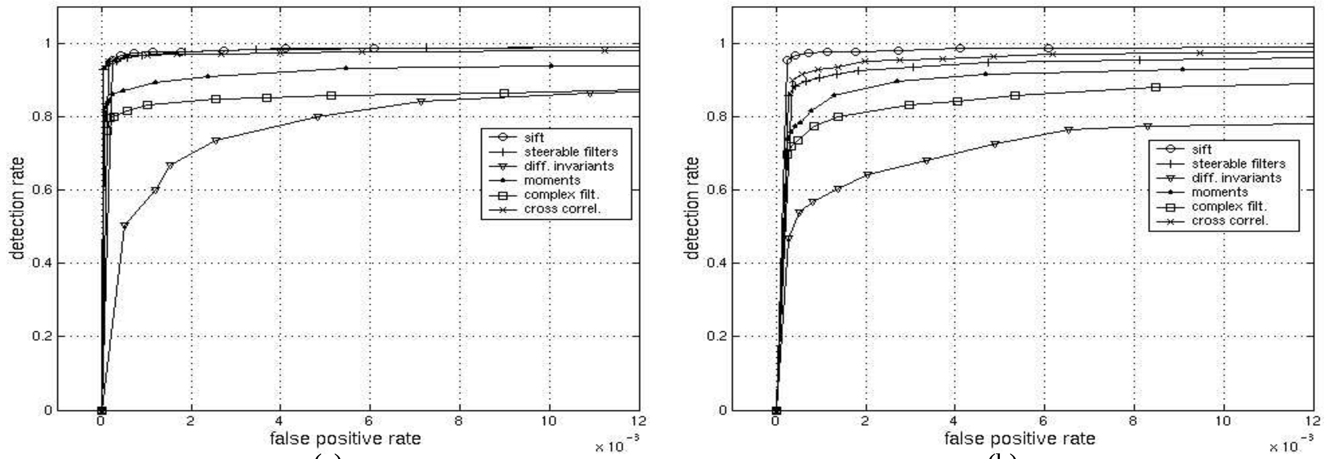
Evaluation criterion. We use a criterion similar to the one proposed in [2]. It is based on Receiver Operating Characteristics (ROC) of detection rate versus false positive rate. Two points $\mathbf{a}$ and $\mathbf{b}$ are similar if the distance between their descriptors is below an arbitrary threshold $d_M(D_{\mathbf{a}} - D_{\mathbf{b}}) < t$. The value of t is varied to obtain the ROC curves.

Given two images representing the same scene the detection rate is the number of correctly matched points with respect to the number of possible matches:

$$p_{correct} = \frac{\#\, correct\, matches}{\#\, possible\, matches}$$

To verify the correct matches we used the criterion proposed in [14]. A match is correct if the error in relative location is less than 3 pixels $||\mathbf{a} - H\mathbf{b}|| < 3$ and the error in image area covered by two corresponding point neighborhoods is less than 30% of the region union. The point location and region area are verified with an independently estimated homography H. The number of possible matches are determined by the same criteria.

The false positive rate is the probability of a false match in a database of descriptors. Each descriptor of the query image is compared with each descriptor of the database and we count the number of false matches. The probability of false positives is the total number of false matches with respect to the product of the number of database points and



(a) (b)

Figure 1: Evaluation for an image rotation of $45°$. (a) Descriptors computed for Harris points. (b) Descriptors computed for Harris-Laplace points.

the number of image points:

$$p_{false} = \frac{\#\, false\ matches}{(\#\, database\ points)(\#\, query\ image\ points)}$$

The results are displayed for a false positive rate up to 0.012, that is each point from the query image matches at most with 1.2% of points in the database. The threshold is usually set below this value otherwise the number of false matches is too high to provide reliable scene recognition.

Data set. We evaluate the descriptors on real image pairs with different geometric and photometric transformations, that is image rotation, scale changes, affine transformations and illumination changes. These transformations have been introduced by rotating the camera, varying the zoom and changing the viewpoint angle. We vary the illumination by changing the brightness and the position of the light source. We use planar scenes such that the homography can be used to verify the correct matches. For each type of transformation we use 3 image pairs, one is the query image and the other one is a part of the database. We then compute an average detection and false positive rate for these three images. To evaluate the false positive rate, that reflects the distinctiveness of the descriptors, we use a database of 1000 images. The images are extracted from 3 hours of a video, which includes movies, sport events and news reports. Similar images are mostly excluded by taking one image per 300 frames. There are about 300 000 points in the database. There is one database of descriptors for each combination of detector/descriptor. The test images are displayed in figure 5 and are available on the Internet [1].

4. Experimental results

In this section we present and discuss the experimental results of the evaluation. The performance is compared for image rotation, scale changes, affine transformations and illumination changes.

4.1. Rotation

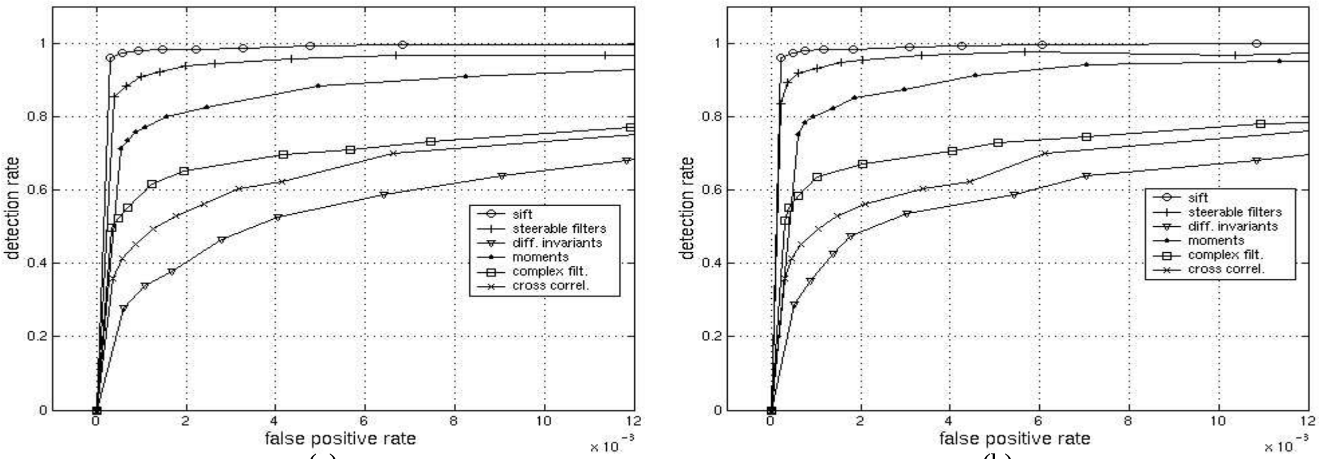
To evaluate the performance for image rotation we used images with a rotation angle of approximatively 45 degrees which represents the most difficult case. In figure 1(a) we compare the descriptors computed for standard Harris points. For these points image patches are fixed to a size of 21x21 pixels and $\sigma = 3.3$ for Gaussian derivatives. We can see that SIFT, steerable filters and cross correlation obtain the best results. The detection rates are lower for scale invariant Harris-Laplace points (cf. figure 1(b)). However, the ranking of the detectors remains the same. The best results are obtained by the SIFT descriptor, followed by cross-correlation and steerable filters. Note that for a 0.9 probability of correct detection, the probability of false match is about 4 times lower for steerable filters than for moment invariants.

There are three principal factors that influence the descriptors: the error in scale estimation, in point localization and in estimating the 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 1 that these errors have less impact on descriptors than the scale error which occurs in the case of Harris-Laplace. An error is introduced if the selected scales are not the same, which can happened due to noise.

4.2. Scale changes

In this section we evaluate the descriptors on images with combined rotation and scale changes. The scale changes are approximately of a factor 2.5 and the image rotation is of about 45 degrees. Figure 2(a) shows the performance of descriptors computed for Harris-Laplace points. We can see

[1] http://www.inrialpes.fr/movi/Mikolajczyk/Database



(a)

(b)

Figure 2: Evaluation for a scale change of a factor 2.5 combined with an image rotation of 45°. (a) Descriptors computed for Harris-Laplace points. (b) Descriptors computed for DoG points.

that the probability of false matches is lower for SIFT and steerable filters than for the other descriptors. Differential invariants obtain a significantly lower detection rate. We can observe that the performance of all descriptors is worse compared to rotation alone. Harris-Laplace interest points are detected for an arbitrary chosen range of scales with a 1.2 scale interval. All descriptors are influenced by the inaccuracy in scale estimation if the pre-selected scales do not match exactly with the real scale change between images. The performance of the cross-correlation drops more significantly than that of the other descriptors. Scale changes combined with rotation significantly deteriorate the results for this technique, as the correlation is very sensitive to the errors in region normalization.

Figure 2(b) shows the results for descriptors computed on regions detected with the DoG detector. The ranking of descriptors is the same as in the case of Harris-Laplace, but the results are slightly better for DoG points. This can be explained by the higher accuracy of the DoG detector used in this comparison. The DoG detector improves the point location and scale estimation by fitting a 3D quadratic function through the DoG function values around the location.

An interesting observation has been made when comparing descriptors computed on different size of a point neighborhood. The performance of descriptors computed for DoG points is much lower than for Harris points if the neighborhood radius is only 3 times the detection scale instead of 5. The DoG points unlike the Harris points are mainly blob-like structures and the signal changes are low in the center of the regions. The descriptor is more stable and distinctive if we capture more signal variations by increasing the size of the region. Harris points are detected at locations with significant signal changes therefore a larger neighborhood size has less influence on the descriptor performance.

4.3. Affine transformations

In this section we evaluate the performance if the viewpoint of the camera is changed by 60 degrees. This introduces a perspective transformation which can be locally approximated by an affine transformation. There are also some scale and brightness changes in the test images. To eliminate the effects of the affine transformation, we use the Harris-Affine detector which extracts affine-invariant regions. The descriptors are computed on point neighborhoods normalized with the locally estimated affine transformations. The performance of all descriptors (cf. figure 3) is lower than for other image transformations, i.e. scale changes and rotation. SIFT descriptors are more robust than the other ones. Note that SIFT descriptors computed on Harris-Laplace regions perform worse than any of the other descriptors (see `HL sift` in figure 3), as these regions and therefore the descriptors are only scale and not affine invariant. Steerable filters come second, but they perform significantly worse than SIFT descriptors.

4.4. Illumination changes

Figure 4 shows the results in the presence of illumination changes which have been obtained by changing the brightness as well as the position of the light source. The descriptors are computed for Harris-Laplace points. Figure 4(a) compares two approaches to obtain affine illumination invariance for differential descriptors: (i) based on region normalization (“steerable filers” and “diff. invariant” used in all our comparisons), (ii) based on the invariance of the descriptors (“inv. steer. filt.” and “inv. diff. inv.”), see section 3.1 for details. We observe that the descriptors computed on normalized regions are significantly better. Theoretically the two methods are equivalent. However, the product of the derivatives of the differential invariants gives rise to noise caused by scale and location errors as

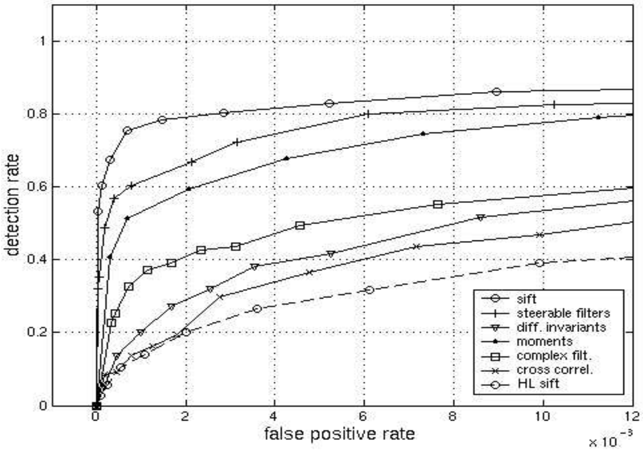


Figure 3: Evaluation for a viewpoint change of the camera of 60°. Descriptors computed for Harris-Affine points. `HL sift` is the SIFT descriptor computed for Harris-Laplace points.

well as non-affine illumination changes. The importance of affine illumination invariance is shown by the comparison to un-normalized descriptors (computed on un-normalized regions). These descriptors obtain worse results.

In figure 4(b) the standard descriptors are compared in the presence of illumination change. Note that it shows the results only for the detection rate higher than 0.6. SIFT descriptors are normalized by the technique proposed in [11], all other descriptors are computed on normalized image patches. We observe how the descriptors perform in the presence of small brightness changes which remain after the patch normalization. All descriptors obtain very good results except the differential invariants. Note that steerable filters perform better than SIFT descriptors. This is probably due to the normalization procedure used for SIFT which might be worth further investigations. We can also see that the photometric image transformations have less influence on descriptors compared to the geometric changes (cf. figure 2 and 3).

Discussion and Conclusions

In this paper we have presented an experimental evaluation of interest point descriptors on images with real geometric and photometric transformations. The goal was to compare descriptors computed on regions extracted with recently proposed detection techniques which are invariant to scale and affine changes. In all tests, except for light changes, SIFT descriptors obtain better results than the other descriptors. This shows the robustness and the distinctive character of the region-based SIFT descriptor. The second best descriptors are the steerable filters computed on image patches normalized to affine photometric and geometric transformations. It can be considered as a good choice given the low dimensionality of this descriptor.

The cross correlation measure gives unstable results. The performance depends on the accuracy of interest point and region detection, which decreases for significant geometric transformations. The differential invariants give significantly worse results that the steerable filters, which is surprising as they are based on the same basic components (Gaussian derivatives). The multiplication of derivatives necessary to obtain the rotation invariance increases the instability of the descriptors.

Regions detected by DoG are mainly blob-like structures. There are no significant signal changes in the center of the blob and therefore the Gaussian filter-based descriptors perform better on larger point neighborhoods.

Obviously, the comparison presented here is not exhaustive and it would be interesting to include more descriptors for example non-parametric descriptors, spin-images and Gabor filters. 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 very sensitive to the region parameters as well as localization 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.

It would be of interest to evaluate the impact of different sources of error which can occur in the estimation of region parameters. Performance of the operators under controlled synthetic image degradation will be a useful and valuable additional dimension of the work.

Acknowledgments

This work was supported by the European FET-open project VIBES and the European project LAVA (IST-2001-34405). We are grateful to David Lowe, Frederik Schaffalitzky and Andrew Zisserman for providing the code for their detectors/descriptors and for useful suggestions.

References

[1] A. Baumberg. Reliable feature matching across widely separated views. In *CVPR*, pp. 774–781, 2000.

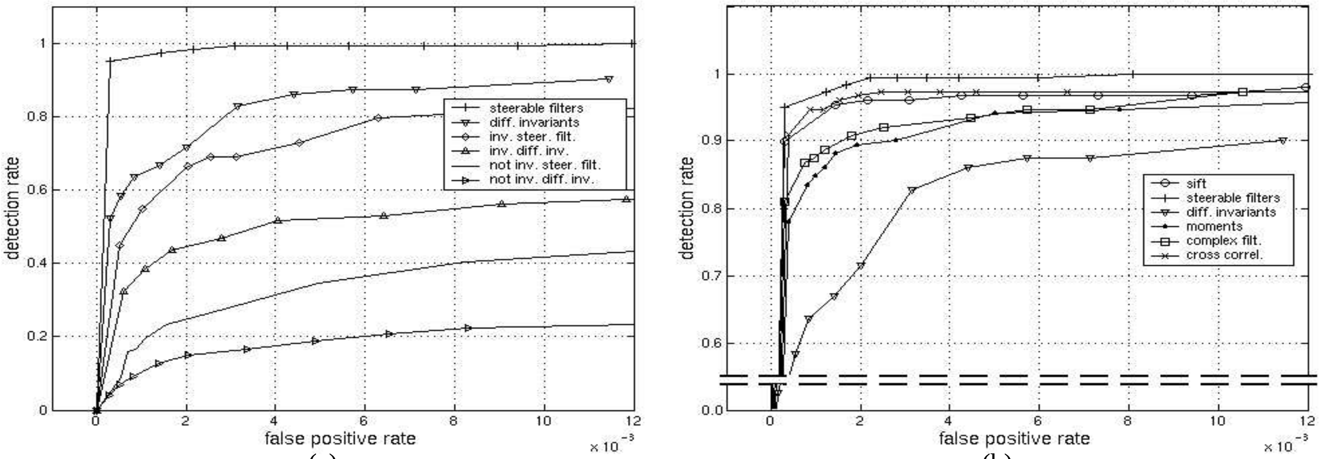
[2] G. Carneiro and A. D. Jepson. Phase-based local features. In *ECCV*, volume I, pp. 282–296, 2002.

[3] H. I. Christensen and P. J. Phillips, *Empirical Evaluation Methods in Computer Vision*, volume 50 of *Series in MPAI*. World Scientific Publishing Co., 2002.

[4] L. Florack, B. ter Haar Romeny, J. Koenderink, and M. Viergever. General intensity transformations and second order invariants. In *SCIA*, pp. 338–345, 1991.

[5] W. Freeman and E. Adelson. The design and use of steerable filters. *PAMI*, 13(9):891–906, 1991.

[6] D. Gabor. Theory of communication. *Journal I.E.E.*, 3(93):429–457, 1946.



(a) (b)

Figure 4: Evaluation for illumination changes. The descriptors are computed for Harris-Laplace points. (a) Illumination invariance of differential descriptors. “Steerable filters” and “diff. invariants” are the standard descriptors computed on the intensity normalized patches. “Inv. steer. filt.” and “inv. diff. inv.” are the illumination invariants and “not inv. steer. filt.” and “not inv. diff. inv.” are not normalized. (b) Descriptors computed on illumination normalized regions.

[7] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d’intérêt pour la couleur. In *RFIA*, pp. 257–266, 2000.

[8] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pp. 147–151, 1988.

[9] A. Johnson and M. Hebert. Object recognition by matching oriented points. In *CVPR*, pp. 684–689, 1997.

[10] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pp. 1150–1157, 1999.

[12] J. K. M. Vetterli. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[13] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, pp. 525–531, 2001.

[14] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV*, pp. 128–142, 2002.

[15] T. Randen and J. H. Husoy. Filtering for texture classification : A comparative study. *PAMI*, 21(4):291–310, 1999.

[16] F. Schaffalitzky and A. Zisserman. Automated scene matching in movies. In *Challenge of Image and Video Retrieval*, pp. 186–197, 2002.

[17] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *ECCV*, pp. 414–431, 2002.

[18] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–534, 1997.

[19] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, 2000.

[20] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, pp. 412–425, 2000.

[21] L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *ECCV*, pp. 642–651, 1996.

[22] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondance. In *ECCV*, pp. 151–158, 1994.

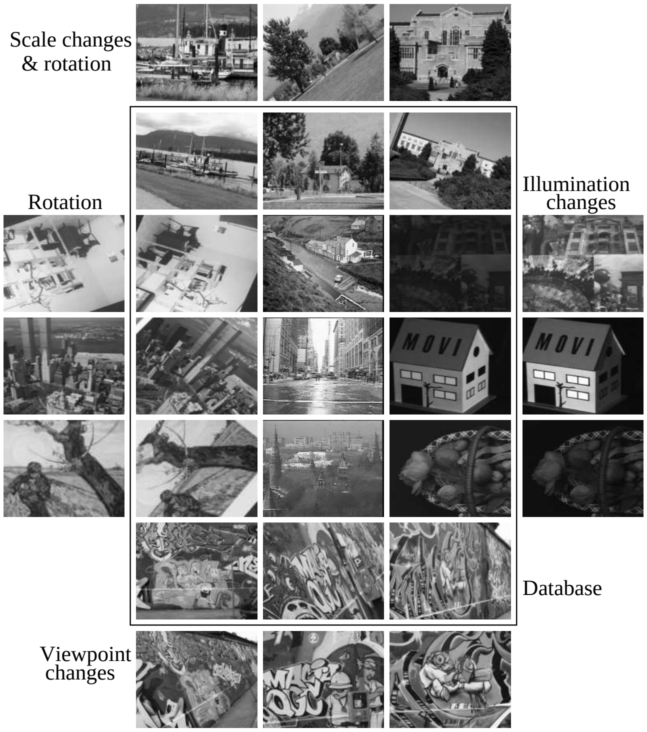


Figure 5: Test images.