Indexing based on scale invariant interest points

Krystian Mikolajczyk* Cordelia Schmid

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

Abstract

This paper presents a new method for detecting scale invariant interest points. The method is based on two recent results on scale space: 1) Interest points can be adapted to scale and give repeatable results (geometrically stable). 2) Local extrema over scale of normalized derivatives indicate the presence of characteristic local structures. Our method first computes a multi-scale representation for the Harris interest point detector. We then select points at which a local measure (the Laplacian) is maximal over scales. This allows a selection of distinctive points for which the characteristic scale is known. These points are invariant to scale, rotation and translation as well as robust to illumination changes and limited changes of viewpoint.

For indexing, the image is characterized by a set of scale invariant points; the scale associated with each point allows the computation of a scale invariant descriptor. Our descriptors are, in addition, invariant to image rotation, to affine illumination changes and robust to small perspective deformations. Experimental results for indexing show an excellent performance up to a scale factor of 4 for a database with more than 5000 images.

1 Introduction

The difficulty in object indexing is to determine the identity of an object under arbitrary viewing conditions in the presence of cluttered real-world scenes or occlusions. Local characterization has shown to be well adapted to this problem. The small size of the characteristic regions makes them robust against occlusion and background changes. To obtain robustness to changes of viewing conditions they should also be invariant to image transformations. Recent methods for indexing differ in the type of invariants used. Rotation invariants have been presented by [10], rotation and scale invariants by [8] and affine invariants by [13].

Schmid and Mohr [10] extract a set of interest points and characterize each of the points by rotationally invariant descriptors which are combinations of Gaussian derivatives. Robustness to scale changes is obtained by computing Gaussian derivatives at several scales. Lowe [8] extends these ideas to scale invariance by maximizing the output of difference-of-Gaussian filters in scale-space. Tuytelaars et al. [13] have developed affine invariant descriptors by searching for affine invariant regions and describing them by color invariants. To find these regions they simultaneously use interest points and contours. Instead of using an initial set of features, Chomat et al. [2] select the appropriate scale for every point in the image and compute descriptors at these scales. An object is represented by the set of these descriptors. All of the above methods are limited to a scale factor of 2.

Similar approaches exist for wide-baseline matching [1, 3, 5, 9, 12]. The problem is however more restricted. Additional constraints can be imposed and the search complexity is less prohibitive. For example, Prichett and Zisserman [9] first match regions bound by four line segments. They then use corresponding regions to compute the homography and grow the regions. Such an approach is clearly difficult to extend to the problem of indexing. Two of the papers on wide-baseline matching have specifically addressed the problem of scale. Hansen et al. [5] present a method that uses correlation of scale traces through multi-resolution images to find correspondence between images. A scale trace is a set of values for a pixel at different scales of computation. Dufournaud et al. [3] use a robust multi-scale framework to match images. Interest points and descriptors are computed at different scale levels. A robust homography based matching algorithm allows to select the correct scale. These two approaches are not usable in the context of indexing, as image to image comparison is necessary. In the context of indexing we need discriminant features which can be accessed directly. Storage of several levels of scale is prohibitive, as it gives rise to additional mismatches and increases the necessary storage space.

In this papers we propose an approach which allows indexing in the presence of scale changes up to a factor 4.

*This work was supported by the French project RNRT AGIR.

The success of this method is based on a repeatable and discriminant point detector. The detector is based on two results on scale space: 1) Interest points can be adapted to scale and give repeatable results [3]. 2) Local extrema over scale of normalized derivatives indicate the presence of characteristic local structures [7]. The first step of our approach is to compute interest points at several scale levels. We then select points at which a local measure (the Laplacian) is maximal over scales. This allows to select a subset of the points computed in scale space. For these points we know their scale of computation, that is their characteristic scale. Moreover, it allows to select the most distinctive points. Points are invariant to scale, rotation and translation as well as robust to illumination changes and limited changes of viewpoint. This detector is the main contribution of this paper. We show that its repeatability is better than the one of other approaches proposed in the literature and therefore allows to obtain better indexing results. The second contribution is the quality of our indexing and matching results.

Overview. This paper is organized as follows. In section 2 we introduce scale selection. In section 3 our scale invariant interest point detector is described and section 4 presents algorithms for matching and indexing. Experimental results are given in section 5.

2. Scale selection

In the following we briefly introduce the concept of scale-space and show how to select the characteristic scale. We then present experimental results for scale selection.

Scale-space. The scale-space representation is a set of images represented at different levels of resolutions [14]. Different levels of resolution are in general created by convolution with the Gaussian kernel: $L(\mathbf{x}, s) = G(s) * I(\mathbf{x})$ with I the image and $\mathbf{x} = (x, y)$. We can represent a feature (i.e. edges, corners) at different resolutions by applying the appropriate function (combinations of derivatives) at different scales. The amplitude of spatial derivatives, in general, decreases with scale. In the case of scale invariant forms, like step-edge, the derivatives should be constant over scales. In order to maintain the property of *scale invariance* the derivative function must be normalized with respect to the scale of observation. The scale normalized derivative D of order m is defined by:

$$D_{i_1 \ldots i_m}(\mathbf{x}, s) = s^m L_{i_1 \ldots i_m}(\mathbf{x}, s) = s^m G_{i_1 \ldots i_m}(s) * I(\mathbf{x})$$

Normalized derivatives behave nicely under scaling of the intensity pattern. Consider two images I and I' imaged at different scales. The relation between the two images is then defined by: $I(\mathbf{x}) = I'(\mathbf{x}')$, *where* $\mathbf{x}' = t\mathbf{x}$. Image derivatives are then related by:

$$s^m G_{i_1 \ldots i_m}(s) * I(\mathbf{x}) = t^m s^m G_{i_1 \ldots i_m}(ts) * I(\mathbf{x}')$$

Thus, for normalized derivatives we obtain:

$$D_{i_1 \ldots i_m}(\mathbf{x}, s) = D'_{i_1 \ldots i_m}(\mathbf{x}, ts)$$

We can see that the same values are obtained at corresponding relative scales.

To maintain uniform information change between successive levels of resolution the scale factor must be distributed exponentially. Let F be a function used to build the scale-space and normalized with respect to scale. The set of responses for a point $\mathbf{x}$ is then $F(\mathbf{x}, s_n)$ with $s_n = k^n s_0$. s_0 is the initial scale factor at the finest level of resolution and s_n denotes successive levels of the scale-space representation with k the factor of scale change between successive levels.

Characteristic scale. The properties of local characteristic scales were extensively studied in [7]. The idea is to select a characteristic scale by searching for a local extremum over scales. Given a point in an image we compute the function responses for several scale factors s_n, see Figure 1. The characteristic scale is the local maximum of the function. Note that there might be several maxima, therefore several characteristic scales. The characteristic scale is relatively independent of the image scale. The ratio of the scales, at which the extrema were found for corresponding points in two rescaled images, is equal to the scale factor between the images. Instead of detecting extrema we can also look for other *easy recognizable* signal shapes such as zero-crossings of the second derivative.

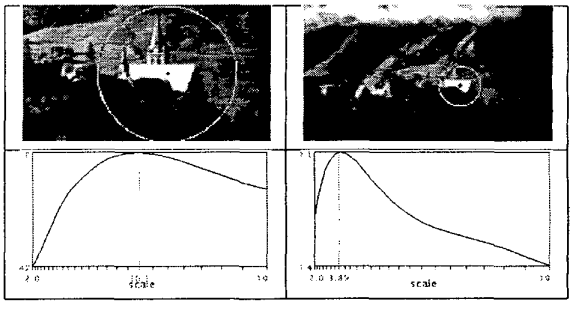


Figure 1: The top row shows two images taken with different focal lengths. The bottom row shows the response $F(\mathbf{x}, s_n)$ over scales where F is the normalized Laplacian (cf. eq.2). The characteristic scales are at 10.1 and 3.89 for the left and right image, respectively. The ratio corresponds to the scale factor (2.5) between the two images.

Several derivative based functions F can be used to compute a scale representation of an image. These functions should be rotation invariant. Illumination invariance is less critical because we are looking for extrema. In the following we present the differential expressions used for our experiments. Note that all expressions are scale normalized.

Square gradient $s^2(L_x^2(\mathbf{x}, s) + L_y^2(\mathbf{x}, s))$ (1)

Laplacian $|s^2(L_{xx}(\mathbf{x},s)+L_{yy}(\mathbf{x},s))|$ (2)

Difference-of-Gaussian $|I(\mathbf{x})*G(s_{n-1})-I(\mathbf{x})*G(s_n)|$ (3)

Harris function $\det(\mathbf{C})-\alpha\,\mathrm{trace}^2(\mathbf{C})$ (4)

with $\mathbf{C}(\mathbf{x},s,\tilde{s}) =$

$$s^2 G(\mathbf{x},\tilde{s}) \star \begin{bmatrix} L_x^2(\mathbf{x},s) & L_xL_y(\mathbf{x},s) \\ L_xL_y(\mathbf{x},s) & L_y^2(\mathbf{x},s) \end{bmatrix}$$

Experimental results. The scale selection technique based on local maxima has been evaluated for functions (1),(2),(3) and (4). The evaluation was conducted on several sequences with scale changes. The characteristic scale was selected for every point in the image. Figure 2 displays image points for which scale selection is possible (white and grey). Black points are points for which the function (Laplacian) has no maximum. Note that these points lie in homogeneous regions and have no maximum in the range of considered scales.

The selected scale for a point is correct if the ratio between characteristic scales in corresponding points is equal to the scale factor between the images. Corresponding points are determined by projection with the estimated transformation matrix. In the case of multiple scale maxima, the point is considered correct, if one of the maxima corresponds to the correct ratio. Points with correctly selected scales are displayed in white (cf. Figure 2).

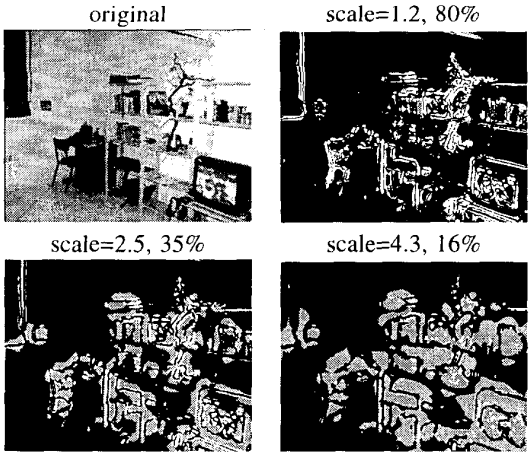


Figure 2: Characteristic scale of points. Black–no characteristic scale is detected. Gray–a characteristic scale is detected. White–a characteristic scale is detected and is correct. The scale of the images is given above the images and corresponds to $scale = \frac{original}{scaled}$. The scaled images were enlarged to increase the visibility.

We can observe that only a small percentage of selected scales are correct for large scale factors. In table 1 we have compared results for different functions F in the presence of a scale factor of 4.3. Results are averaged over several sequences. The first row shows the function used. The second shows the percentage of points for which a characteristic scale is detected. We can observe that most points are detected by the Laplacian. The percentage of correct points with respect to detected points is given in row three. The Laplacian and the DOG obtain the highest percentage. The last row shows the overall percentage of correct detection. Most correct points are detected by the Laplacian. The percentage is twice as high as for the gradient, and four times higher than for the Harris function. Results are similar to those of the DOG which is not surprising as this function is very similar to the Laplacian.

	Laplacian	DOG	gradient	Harris
detected	46%	38%	30%	16%
correct/ detected	29%	28%	22%	23%
correct	13.3%	10.6%	6.6%	3.4%

Table 1: Row 2: percentage of points for which a characteristic scale is detected. Row 3: percentage of points for which a correct scale is detected with respect to detected points. Row 4: percentage of correct / total.

We have observed that the performance degrades in the presence of large scale changes. This can be explained by the fixed search range of scales, which must be the same for all images if we have no a priori knowledge about the scale factor between the images. If the characteristic scale found in a coarse resolution image is near the upper limit of the scale range, the corresponding point at a finer scale is likely to be too far from significant signal changes to be detected in our scale limits. Our experiments shows, that characteristic scale found by searching for extrema only in the scale direction, are sensitive to this fact. Furthermore, we cannot apply too large a range of scales as we lose the local character, and the effect of image borders becomes too important.

3. Scale invariant interest points

The previous section shows that using all points gives unstable results. Feature points permit stabilizing the results.

Existing methods search for maxima in the 3D representation of an image (x, y and $scale$). A feature point represents a local maximum in the surrounding 3D cube and its value has to be higher than a certain threshold. In Figure 3 the point m is a feature point, if $\forall\bullet\ F(m,s_n) > F(\bullet,s_l)$ with $l \in \{n-1,n,n+1\}$ and $F(m,s_n) > t$.
Lindeberg [7] searches for 3D maxima of the Laplacian, as well as the magnitude of the gradient and Lowe [8] uses the difference-of-Gaussian.

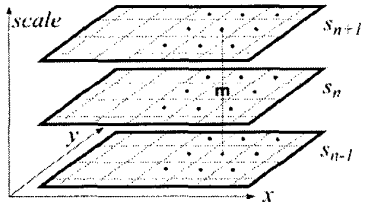


Figure 3: Searching for maxima in scale-space.

Our approach does not use a single function to search in 3D, but uses the Harris function (cf. eq. 4) to localize points in 2D and then selects points for which the Laplacian attains a maximum over scales. In the following, it is referred to as the Harris-Laplacian.

The Harris detector is used for 2D localization as it has shown to be most reliable in the presence of image rotation, illumination transformations and perspective deformations as shown in a comparative evaluation [11]. However, the repeatability of this detector fails when the resolution of images changes significantly. In order to deal with such changes, the Harris detector has to be adapted to the scale factor [3]. Repeatability results for such an adapted version are excellent. The remaining problem is scale selection. During our experiments we noticed that the adapted Harris function rarely attains maxima in 3D space. If too few points are detected, the image representation is not robust. Therefore, we propose to use a different function, the Laplacian, for scale maxima detection. We have seen in the previous section that this function allows to find the highest percentage of correct maxima.

Our detection algorithm works as follows. We first build a scale-space representation for the Harris function. At each level of the scale-space we detect interest points by detecting the local maxima in the image plane:

$$F(\mathbf{x}, s_n) > F(\mathbf{x_w}, s_n) \;\; \forall \mathbf{x_w} \in W$$
$$F(\mathbf{x}, s_n) > t_h$$

where W denotes the 8-neighbourhood of the point $\mathbf{x}$.

In order to obtain a more compact representation, we verify for each of the candidate points found on different levels if it forms a maximum in the scale direction. The Laplacian is used for selection.

$$F(\mathbf{x}, s_n) > F(\mathbf{x}, s_{n-1}) \wedge F(\mathbf{x}, s_n) > F(\mathbf{x}, s_{n+1})$$
$$F(\mathbf{x}, s_n) > t_l$$

Figure 5 shows the scale-space representation for two real images with points detected by the Harris-Laplacian method. For these two images of the same object imaged at different scales we present for each scale level the selected points. There are many point-to-point correspondences between the levels for which the scale ratio corresponds to the real scale change between the images (indicated by pointers). Additionally, very few points are detected in the same location but on different levels. Our points are therefore characteristic to the image plane and the scale dimension.

A comparative evaluation of different scale invariant interest point detectors is presented in the following. We compare the approaches of Lindeberg (Laplacian and gradient), Lowe as well as our Harris-Laplacian detector. To show the gain compared to the non-scale invariant method, we also present the results of the standard Harris detector. The stability of detectors is evaluated using the repeatability criteria introduced in [11]. The repeatability score is computed as a ratio between the number of point-to-point correspondences that can be established for detected points and the mean number of points detected in two images: $r_{1,2} = \frac{C(I_1,I_2)}{mean(m_1,m_2)}$ where $C(I_1, I_2)$ denotes the number of corresponding couples and m_1, m_2 the numbers of detected points in the images. Two points correspond if the error in relative location does not exceed 1.5 pixel in the coarse resolution image and the ratio of detected scales for these points does not differ from the real scale ratio by more than 20%. Figure 4 presents the repeatability score for the compared methods. The experiments were done on 10 sequences of real images. Each sequence consists of scaled and rotated images for which the scale factor varies from 1.2 up to 4.5. Best results are obtained for the Harris-Laplacian method. The results are 10% better than those of the second best detector, the Laplacian.

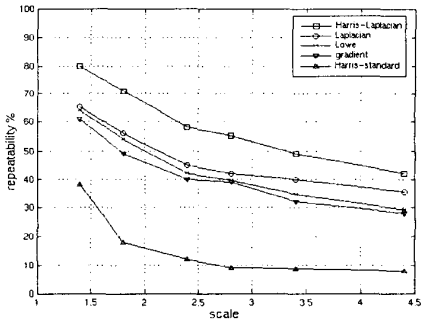


Figure 4: Repeatability of interest point detectors with respect to scale changes.

4. Robust matching and indexing

In the following we briefly describe our robust matching and indexing algorithms. The two algorithms are based on the same initial steps:

1. Extraction of Harris-Laplacian interest points (cf. section 3).
2. Computation of a descriptor for each point at its characteristic scale. Descriptors are invariant to image rotation and affine illumination changes. They are robust to small perspective deformations.

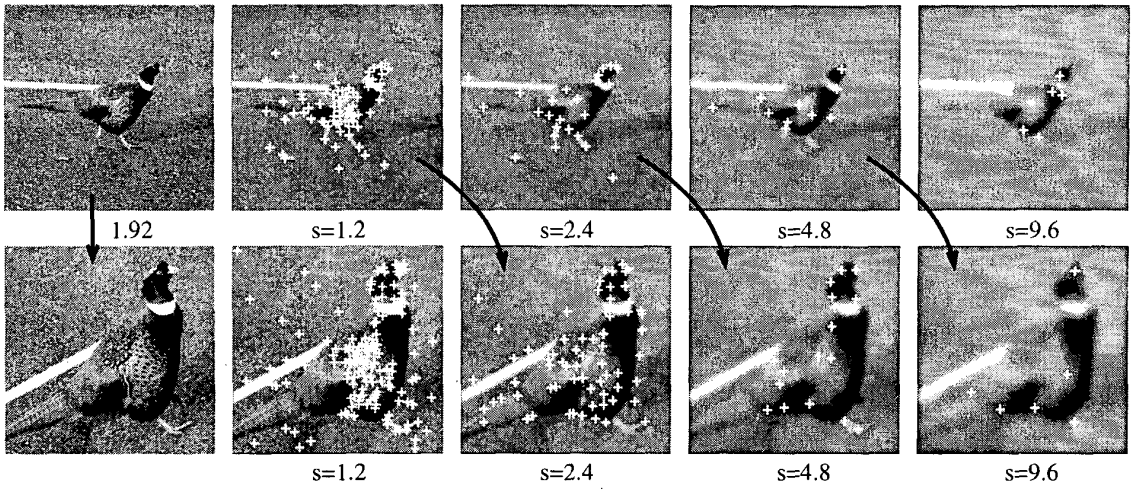


Figure 5: Points detected on different resolution levels with the Harris-Laplacian method.

3. Comparison of descriptors based on the Mahalanobis distance.

Interest points. To extract interest points we have used a scale representation with 17 resolution levels. The initial scale s_0 is 1.5 and the factor k between two levels of resolution is 1.2. The parameter α is set to 0.06 and the thresholds t_h and t_l are set to 1500 and 10, respectively.

Descriptors. Our descriptors are Gaussian derivatives which are computed at the characteristic scale. Invariance to rotation is obtained by "steering" the derivatives in the direction of the gradient [4]. To obtain a stable estimation of the gradient direction, we use the peak in a histogram of local gradient orientations. Invariance to the affine intensity changes is obtained by dividing the derivatives by the steered first derivative. Using up to 4th order derivatives, we obtain descriptors of dimension 12.

Comparison of descriptors. The similarity of descriptors is measured by the Mahalanobis distance. This distance requires the estimation of the covariance matrix Λ which encapsulates signal noise, variations in photometry, inaccuracy of interest point location, and so forth. Λ is estimated statistically over a large set of image samples.

Robust matching. To robustly match two images, we first determine point-to-point correspondences. We select for each descriptor in the first image the most similar descriptor in the second image based on the Mahalanobis distance. If the distance is below a threshold the match is kept. This allows us to obtain a set of initial matches. A robust estimation of the transformation between the two images based on RANdom SAmple Consensus (RANSAC) allows to reject inconsistent matches. For our experimental results the transformation is either a homography or a fundamental matrix. A model selection algorithm [6] can of course be used to automatically decide what transformation is the most appropriate one.

Indexing. A voting algorithm is used to select the most similar images in the database. This makes retrieval robust to mismatches as well as outliers. For each point of a query image, its descriptor is compared to the descriptors in the database. If the distance is less than a fixed threshold, a vote is added to the corresponding database image. Note that a point cannot vote several times for the same database image. The database image with the highest number of votes is the most similar one.

5. Experimental results

In the following, we validate our detection algorithm by matching and indexing results. Figure 6 illustrates the different steps of our matching algorithm. In this example the two images are taken from the same viewpoint, but with a change in focal length and image orientation. The top row shows the detected interest points. There are 190 and 213 points detected in the left and right images, respectively. The number of detected points is about equivalent to results obtained by a standard interest point detector. This clearly shows the selectivity of our point detection method. If no scale peak selection had been used, more than 2000 points would be detected. The middle row shows the 58 matches obtained during the initial matching phase. The bottom row displays the 32 inliers to the estimated homography, all of which are correct. The estimated scale factor between the two images is 4.9 and the estimated rotation angle is 19 degrees.

Figure 7 shows an example for a 3D scene where the fundamental matrix is used for verification. There are 180

and 176 detected points detected in the left and right images. The number of initial matches is 23 and there are 14 inliers to the robustly estimated fundamental matrix, all of them correct. Note that the images are taken from different viewpoints, the transformation includes a scale change, an image rotation as well as a change in the viewing angle. The building in the middle is almost half occluded.

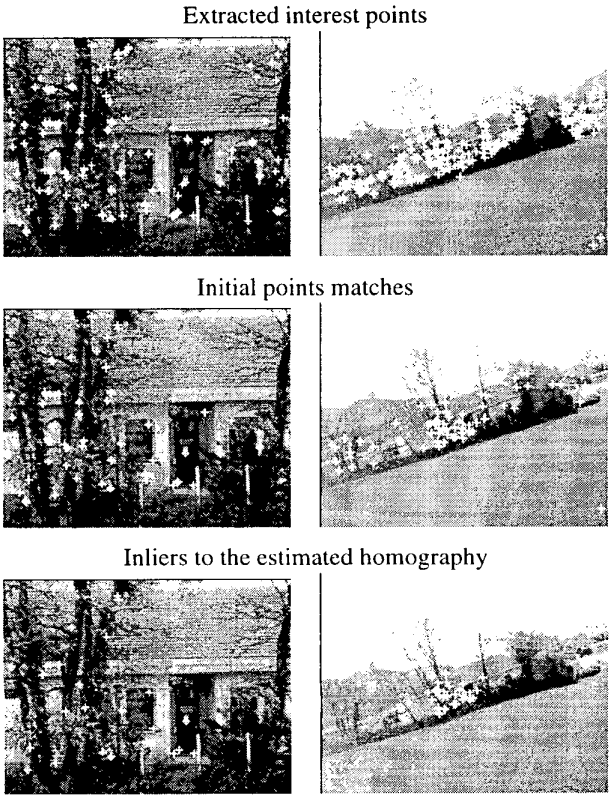


Figure 6: Robust matching: there are 190 and 213 points detected in the left and right images, respectively (top). 58 points are initially matched (middle). There are 32 inliers to the estimated homography (bottom), all of which are correct. The estimated scale factor is 4.9 and the estimated rotation angle is 19 degrees.

In the following we show the results for retrieval from a database with more than 5000 images. The images in the database are extracted from 16 hours of video sequences which include movies, sport events and news reports. Similar images are excluded by taking one image per 300 frames. Furthermore, the database contains one image from each of our 10 test sequences. The total number of descriptors in our database is 2539342.

The second row of figure 8 shows five images of the test sequences which are contained in the database. The top row displays images for which the corresponding image in the database (second row) was correctly retrieved, that is it was the most similar one. The approximate scale factor is given in row three. The changes between the image pairs (first and second row) include important changes in the focal length, for example 5.8 for the image pair (a). They also include important changes in viewpoint, for example for pair (b). Furthermore, they include important illumination changes (image pair (e)).

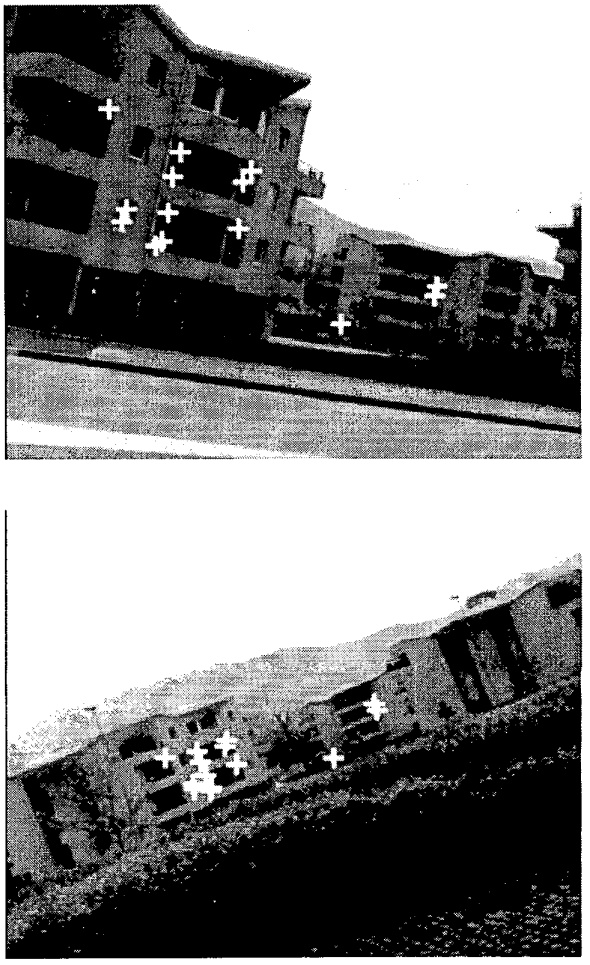
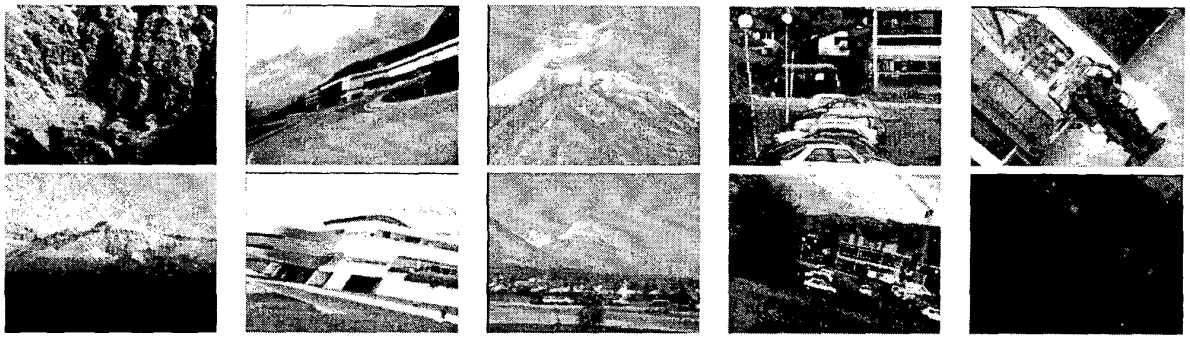


Figure 7: Example of images taken from different view points. There are 14 inliers to a robustly estimated fundamental matrix, all of them are correct. The estimated scale factor is 2.7.

The test sequences where used to systematically evaluate the performance of retrieval. Results are shown in table 2. For each of the 10 test sequences, we have evaluated the performance at different scale factors (1.4 to 4.4). For each scale factor, we have evaluated the percentage that the corresponding image is the most similar one or among the five or ten most similar images. We can see that up to a



(a) 1/5.8 (b) 3.7 (c) 1/4.4 (d) 1/4.1 (e) 5.7

Figure 8: The first row shows some of the query images. The second row shows the most similar images in the database, all of them are correct. The approximative scale factor between query image and database image is given in row three.

scale factor of 4.4, the performance is very good. At the scale of 4.4, 30% of the images are correctly retrieved, 50% are among the 5 best matches and 70% are among the 10 best matches. These results were obtained with 12 dimensional descriptors. If we use derivatives up to order 3, that is 7 dimensional descriptors, the results degrade significantly. This justifies using the fourth order derivatives.

# retrieved	scale factor					
	1.4	1.8	2.4	2.8	3.4	4.4
1	60	60	60	50	30	30
5	100	90	60	80	50	50
10	100	100	90	90	80	70

Table 2: Indexing results for our test sequences at different scale factors. The first row of the table gives the percentage of correct retrieval, that is the corresponding image is retrieved as the most similar one. The second/third row give percentages that the corresponding image is among the 5/10 most similar images.

6. Conclusions and perspectives

We have presented an algorithm for interest point detection that is invariant to important scale changes. A comparison with existing detectors shows that our interest point detector gives better results. Experimental validation for matching and indexing was carried out on a significant amount of data. Matching and indexing results are very good up to a scale factor of 4. To our knowledge none of the existing approach allows to deal with such scale factors in the context of indexing. Furthermore, our approach is invariant to image rotation and translation as well as robust to illumination changes and limited changes in viewpoint. Performance could be further improved by using more robust point descriptors. In our future research, we intend to focus on the problem of affine invariance of point descriptors.

References

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

[2] O. Chomat, V. C. de Verdière, D. Hall, and J. Crowley. Local scale selection for Gaussian based description techniques. In *ECCV*, pages 117–133, 2000.

[3] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *CVPR*, pages 612–618, 2000.

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

[5] B. B. Hansen and B. S. Morse. Multiscale image registration using scale trace correlation. In *CVPR*, pages 202–208, 1999.

[6] K. Kanatani. Geometric information criterion for model selection. *IJCV*, 26(3):171–189, 1998.

[7] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

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

[9] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760, 1998.

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

[11] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *ICCV*, pages 230–235, 1998.

[12] D. Tell and S. Carlsson. Wide baseline point matching using affine invariants computed from intensity profiles. In *ECCV*, pages 814–828, 2000.

[13] T. Tuytelaars and L. V. Gool. Content-based image retrieval based on local affinely invariant regions. In *Visual99*, pages 493–500, 1999.

[14] A. Witkin. Scale-space filtering. In *IJCAI*, pages 1019–1023, 1983.