International Journal of Computer Vision 60(1), 63–86, 2004

Scale & Affine Invariant Interest Point Detectors

KRYSTIAN MIKOLAJCZYK AND CORDELIA SCHMID
INRIA Rhne-Alpes GRAVIR-CNRS, 655 av. de l'Europe, 38330 Montbonnot, France
Krystian.Mikolajczyk@inrialpes.fr
Cordelia.Schmid@inrialpes.fr

Received January 3, 2003; Revised September 24, 2003; Accepted January 22, 2004

Abstract. In this paper we propose a novel approach for detecting interest points invariant to scale and affine transformations. Our scale and affine invariant detectors are based on the following recent results : (1) Interest points extracted with the Harris detector can be adapted to affine transformations and give repeatable results (geometrically stable). (2) The characteristic scale of a local structure is indicated by a local extremum over scale of normalized derivatives (the Laplacian). (3) The affine shape of a point neighborhood is estimated based on the second moment matrix.

Our scale invariant detector computes a multi-scale representation for the Harris interest point detector and then selects points at which a local measure (the Laplacian) is maximal over scales. This provides a set of distinctive points which are invariant to scale, rotation and translation as well as robust to illumination changes and limited changes of viewpoint. The characteristic scale determines a scale invariant region for each point. We extend the scale invariant detector to affine invariance by estimating the affine shape of a point neighborhood. An iterative algorithm modifies location, scale and neighborhood of each point and converges to affine invariant points. This method can deal with significant affine transformations including large scale changes. The characteristic scale and the affine shape of neighborhood determine an affine invariant region for each point.

We present a comparative evaluation of different detectors and show that our approach provides better results than existing methods. The performance of our detector is also confirmed by excellent matching results; the image is described by a set of scale/affine invariant descriptors computed on the regions associated with our points.

Keywords: interest points, local features, scale invariance, affine invariance, matching, recognition

1. Introduction

Local features have been shown to be well suited to matching and recognition as well as to many other applications as they are robust to occlusion, background clutter and other content changes. The difficulty is to obtain invariance to viewing conditions. Different solutions to this problem have been developed over the past few years and are reviewed in Section 1.1. These approaches first detect features and then compute a set of descriptors for these features. In the case of significant transformations, feature detection has to be adapted to the transformation, as at least a subset of the features must be present in both images in order to allow for correspondences. Features which have proved to be particularly appropriate are interest points. However, the Harris interest point detector is not invariant to scale and affine transformations (Schmid et al., 2000). In this paper we give a detailed description of a scale and an affine invariant interest point detector introduced in Mikolajczyk and Schmid (2001, 2002). Our approach combines the Harris detector with the Laplacian-based scale selection. The Harris-Laplace detector is then extended to deal with significant affine transformations. Previous detectors partially handle the problem of affine invariance since they

assume that the localization and scale are not affected by an affine transformation of the local image structures. The proposed improvements result in better repeatability and accuracy of interest points. Moreover, the scale invariant Harris-Laplace approach detects different regions than the DoG detector (Lowe, 1999). The latter one detects mainly blobs, whereas the Harris detector responds to corners and highly textured points, hence these detectors extract complementary features in images.

If the scale change between images is known, we can adapt the Harris detector to the scale change (Dufournaud et al., 2000) and we then obtain points, for which the localization and scale perfectly reflect the real scale change between two images. If the scale change between images is unknown, a simple way to deal with scale changes is to extract points at several scales and to use all these points to represent an image. The problem with a multi-scale approach is that in general a local image structure is present in a certain range of scales. The points are then detected at each scale within this range. As a consequence, there are many points, which represent the same structure, but the location and the scale of the points is slightly different. The unnecessarily high number of points increases the probability of mismatches and the complexity of the matching algorithms. In this case, efficient methods for rejecting the false matches and for verifying the results are necessary.

Our scale invariant approach solves this problem by selecting the points in the multi-scale representation which are present at *characteristic* scales. Local extrema over scale of normalized derivatives indicate the presence of characteristic local structures (Lindeberg, 1998). Here we use the Laplacian-of-Gaussian to select points localized at maxima in scale-space. This detector can deal with significant scale changes, as presented in Section 2. To obtain affine invariant points, we adapt the shape of the point neighborhood. The affine shape is determined by the second moment matrix (Lindeberg and Garding, 1997). We then obtain a truly affine invariant image description which gives stable/repeatable results in the presence of arbitrary viewpoint changes. Note that a perspective transformation of a smooth surface can be locally approximated by an affine transformation. Although smooth surfaces are almost never planar in the large, they are *always* planar in the small that is, sufficiently small surface patches can always be thought of as being comprised of coplanar points. Of course this does not hold if the point is localized on a depth boundary. However, such points are rejected during the subsequent steps, for example during matching. An additional post-processing method can be used to separate the foreground from the background (Borenstein and Ullman, 2002; Mikolajczyk and Schmid, 2003b). The affine invariant detector is presented in Section 3. To measure the accuracy of our detectors we introduce a repeatability criterion which we use to evaluate and compare our detectors to existing approaches. Section 4 presents the evaluation criteria and the results of the comparison, which shows that our detector performs better then existing ones. Finally, in Section 5 we present experimental results for matching.

1.1. Related Work

Many approaches have been proposed for extracting scale and affine invariant features. These are reviewed in the following.

Scale Invariant Detectors. There are a few approaches which are truly invariant to significant scale changes. Typically, such techniques assume that the scale change is the same in every direction, although they exhibit some robustness to weak affine deformations. Existing methods search for local extrema in the 3D scale-space representation of an image (x, y and *scale*). This idea was introduced in the early eighties by Crowley (1981) and Crowley and Parker (1984). In this approach the pyramid representation is computed using difference-of-Gaussian filters. A feature point is detected if a local 3D extremum is present and if its absolute value is higher than a threshold. The existing approaches differ mainly in the differential expression used to build the scale-space representation.

Lindeberg (1998) searches for 3D maxima of scale normalized differential operators. He proposes to use the Laplacian-of-Gaussian (LoG) and several other derivative based operators. The scale-space representation is built by successive smoothing of the high resolution image with Gaussian based kernels of different size. The LoG operator is circularly symmetric and it detects blob-like structures. The scale invariance of interest point detectors with automatic scale selection has also been explored by Bretzner and Lindeberg (1998) in the context of tracking.

Lowe (1999) proposed an efficient algorithm for object recognition based on local 3D extrema in

the scale-space pyramid built with difference-of-Gaussian (DoG) filters. The input image is successively smoothed with a Gaussian kernel and sampled. The difference-of-Gaussian representation is obtained by subtracting two successive smoothed images. Thus, all the DoG levels are constructed by combined smoothing and sub-sampling. The local 3D extrema in the pyramid representation determine the localization and the scale of the interest points. The DoG operator is a close approximation of the LoG function but the DoG can significantly accelerate the computation process (Lowe, 1999). A few images per second can be processed with this algorithm.

The common drawback of the DoG and the LoG representation is that local maxima can also be detected in the neighborhood of contours or straight edges, where the signal change is only in one direction. These maxima are less stable because their localization is more sensitive to noise or small changes in neighboring texture. A more sophisticated approach, solving this problem, is to select the scale for which the trace and the determinant of the Hessian matrix ($\mathcal{H}$) simultaneously assume a local extremum (Mikolajczyk, 2002). The trace of the $\mathcal{H}$ matrix is equal to the LoG but detecting simultaneously the maxima of the determinant penalizes points for which the second derivatives detect signal changes in only one direction. A similar idea is explored in the Harris detector, although it uses the first derivatives. The second derivative gives a small response exactly in the point where the signal change is most significant. Therefore the maxima are not localized exactly at the largest signal variation, but in its neighborhood.

A different approach for the scale selection was proposed by Kadir and Brady (2001). They explore the idea of using local complexity as a measure of saliency. The salient scale is selected at the entropy extremum of the local descriptors. The selected scale is therefore descriptor dependent. The method searches for scale localized features with high entropy, with the constraint that the scale is isotropic.

Affine Invariant Detectors. An affine invariant detector can be seen as a generalization of the scale invariant detector. In the case of an affine transformation the scaling can be different in each direction. The non-uniform scaling has an influence on the localization, the scale and the shape of a local structure. Therefore, the scale invariant detectors fail in the case of significant affine transformations.

An affine invariant algorithm for corner detection was proposed by Alvarez and Morales (1997). They apply affine morphological multi-scale analysis to extract corners. For each extracted point they build a chain of points detected at different scales, but associated with the same local image structure. The final location and orientation of the corner is computed using the bisector line given by the chain of points. A similar idea was previously explored by Deriche and Giraudon (1993). The main drawback of these approaches is that an interest point in images of natural scenes cannot be approximated by a model of a perfect corner, as it can take any form of a bi-directional signal change. The real points detected at different scales do not move along a straight bisector line as the texture around the points significantly influences the location of the local maxima. This approach cannot be a general solution to the problem of affine invariance but gives good results for images where the corners and multi-junctions are formed by straight or nearly straight step-edges. Our approach makes no assumption on the form of a local structure. It only requires a bi-directional signal change.

Recently, Tuytelaars and Van Gool (1999, 2000) proposed two approaches for detecting image features in an affine invariant way. The first one starts from Harris points and uses the nearby edges. Two nearby edges, which are required for each point, limit the number of potential features in an image. A parallelogram region is bounded by these two edges and the initial Harris point. Several intensity based functions are used to determine the parallelogram. In this approach, a reliable algorithm for extracting the edges is necessary. The second method is purely intensity-based and starts with extraction of local intensity extrema. Next, the algorithm investigates the intensity profiles along rays going out of the local extremum. An ellipse is fitted to the region determined by significant changes in the intensity profiles. A similar approach based on local intensity extrema was introduced by Matas et al. (2002). They use the water-shed algorithm to find intensity regions and fit an ellipse to the estimated boundaries.

Lindeberg and Garding (1997) developed a method for finding blob-like affine features with an iterative procedure in the context of shape from texture. The affine invariance of shape adapted fixed points was also used for estimating surface orientation from binocular data (shape from disparity gradients). This work provided the theory for the affine invariant detector presented in this paper. It explores the properties of the

second moment matrix and iteratively estimates the affine transformation of local patterns. The authors propose to extract the points using the maxima of a uniform scale-space representation and to iteratively modify the scale and the shape of points. However, the location of points is detected only at the initial step of the algorithm, by the circularly symmetric, not affine invariant Laplacian measure. Therefore, the spatial location of the maximum can be slightly different if the pattern undergoes a significant affine deformation. This method was also applied to detect elliptical blobs in the context of hand tracking (Laptev and Lindeberg, 2001). The affine shape estimation was used for matching and recognition by Baumberg (2000). He extracts interest points at several scales using the Harris detector and then adapts the shape of the point neighborhood to the local image structure using the iterative procedure proposed by Lindeberg. The affine shape is estimated for a fixed scale and fixed location, that is the scale and the location of the points are not extracted in an affine invariant way. The points as well as the associated regions are therefore not invariant in the case of significant affine transformations (see Section 4.1 for a quantitative comparison). Furthermore, there are many points repeated at the neighboring scale levels (Fig. 2), which increases the probability of false matches and the complexity. Recently, Schaffalitzky and Zisserman (2002) extended the Harris-Laplace detector (Mikolajczyk and Schmid, 2001) by affine normalization proposed by Baumberg (2000). However, the location and scale of points are provided by the scale invariant Harris-Laplace detector (Mikolajczyk and Schmid, 2001), which is not invariant to significant affine transformations.

2. Scale Invariant Interest Point Detector

The evaluation of interest point detectors presented in Schmid et al. (2000) demonstrate an excellent performance of the Harris detector compared to other existing approaches (Cottier, 1994; Forstner, 1994; Heitger et al., 1992; Horaud et al., 1990). However this detector is not invariant to scale changes. In this section we propose a new interest point detector that combines the reliable Harris detector (Harris and Stephens, 1988) with automatic scale selection (Lindeberg, 1998) to obtain a scale invariant detector. In Section 2.1 we introduce the methods on which we base the approach. In Section 2.2 we discuss in detail the scale invariant detector and present an example of extracted points.

2.1. Feature Detection in Scale-Space

Scale Adapted Harris Detector. The Harris detector is based on the second moment matrix. The second moment matrix, also called the auto-correlation matrix, is often used for feature detection or for describing local image structures. This matrix must be adapted to scale changes to make it independent of the image resolution. The scale-adapted second moment matrix is defined by:

$$\mu(\mathbf{x}, \sigma_I, \sigma_D) = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} = \sigma_D^2 \, g(\sigma_I) * \begin{bmatrix} L_x^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \quad (1)$$

where σ_I is the integration scale, σ_D is the differentiation scale and L_a is the derivative computed in the a direction. The matrix describes the gradient distribution in a local neighborhood of a point. The local derivatives are computed with Gaussian kernels of the size determined by the local scale σ_D (differentiation scale). The derivatives are then averaged in the neighborhood of the point by smoothing with a Gaussian window of size σ_I (integration scale). The eigenvalues of this matrix represent two principal signal changes in the neighborhood of a point. This property enables the extraction of points, for which both curvatures are significant, that is the signal change is significant in the orthogonal directions i.e. corners, junctions etc. Such points are stable in arbitrary lighting conditions and are representative of an image. One of the most reliable interest point detectors, the Harris detector (Harris and Stephens, 1988), is based on this principle. The Harris measure combines the trace and the determinant of the second moment matrix:

$$cornerness = \det(\mu(\mathbf{x}, \sigma_\mathbf{I}, \sigma_\mathbf{D})) - \alpha \mathrm{trace}^2(\mu(\mathbf{x}, \sigma_\mathbf{I}, \sigma_\mathbf{D})) \quad (2)$$

Local maxima of *cornerness* determine the location of interest points.

Automatic Scale Selection. Automatic scale selection and the properties of the selected scales have been extensively studied by Lindeberg (1998). The idea is to select the *characteristic* scale of a local structure, for which a given function attains an extremum over scales. In relation to automatic scale selection, the term *characteristic* originally referred to the fact that the selected

scale estimates the *characteristic length* of the corresponding image structures, in a similar manner as the notion of *characteristic length* is used in physics. The selected scale is characteristic in the quantitative sense, since it measures the scale at which there is maximum similarity between the feature detection operator and the local image structures. This scale estimate will (for a given image operator) obey perfect scale invariance under rescaling of the image pattern.

Given a point in an image and a scale selection operator we compute the operator responses for a set of scales σ_n (Fig. 1). The characteristic scale corresponds to the local extremum of the responses. Note that there might be several maxima or minima, that is several characteristic scales corresponding to different local structures centered on this point. The characteristic scale is relatively independent of the image resolution. It is related to the structure and not to the resolution at which the structure is represented. The ratio of the scales at which the extrema are found for corresponding points is the actual scale factor between the point neighborhoods. In Mikolajczyk and Schmid (2001) we compared several differential operators and we noticed that the scale-adapted Harris measure rarely attains maxima over scales in a scale-space representation. If too few interest points are detected, the image content is not reliably represented. Furthermore, the experiments showed that Laplacian-of-Gaussians finds the highest percentage of correct characteristic scales to be found.

$$|\mathrm{LoG}(\mathbf{x}, \sigma_n)| = \sigma_n^2 |L_{xx}(\mathbf{x}, \sigma_n) + L_{yy}(\mathbf{x}, \sigma_n)| \quad (3)$$

When the size of the LoG kernel matches with the size of a blob-like structure the response attains an extremum. The LoG kernel can therefore be interpreted as a matching filter (Duda and Hart, 1973). The LoG is well adapted to blob detection due to its circular symmetry, but it also provides a good estimation of the characteristic scale for other local structures such as corners, edges, ridges and multi-junctions. Many previous results confirm the usefulness of the Laplacian function for scale selection (Chomat et al., 2000; Lindeberg, 1993, 1998; Lowe, 1999).

2.2. *Harris-Laplace Detector*

In the following we explain in detail our scale invariant feature detection algorithm. The Harris-Laplace detector uses the scale-adapted Harris function (Eq. (2)) to localize points in scale-space. It then selects the points for which the Laplacian-of-Gaussian, Eq. (3), attains a maximum over scale. We propose two algorithms. The first one is an iterative algorithm which detects simultaneously the location and the scale of characteristic regions. The second one is a simplified algorithm, which is less accurate but more efficient.

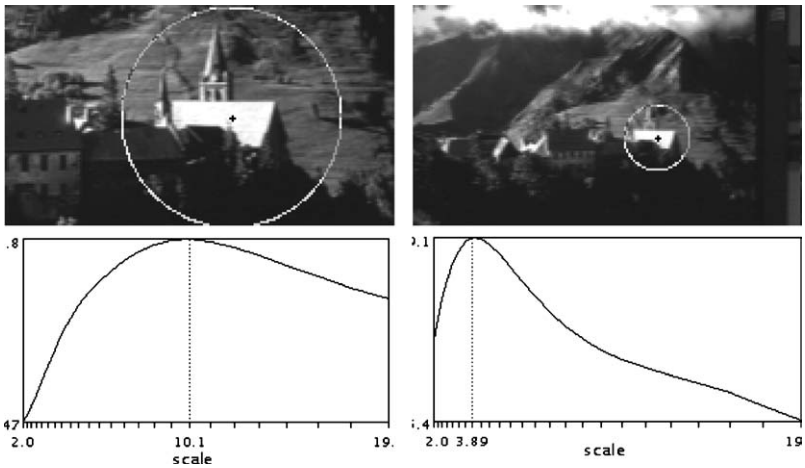


Figure 1. Example of characteristic scales. The top row shows two images taken with different focal lengths. The bottom row shows the response $F_{\mathrm{norm}}(\mathbf{x}, \sigma_n)$ over scales where F_{norm} is the normalized LoG (cf. Eq. (3)). The characteristic scales are 10.1 and 3.89 for the left and right image, respectively. The ratio of scales corresponds to the scale factor (2.5) between the two images. The radius of displayed regions in the top row is equal to 3 times the characteristic scale.

Harris-Laplace Detector. The algorithm consists of two steps: a multi-scale point detection and an iterative selection of the scale and the location. We first build a scale-space representation with the Harris function for pre-selected scales $\sigma_n = \xi^n \sigma_0$, where ξ is the scale factor between successive levels (set to 1.4 (Lindeberg, 1998; Lowe, 1999)). At each level of the representation we extract the interest points by detecting the local maxima in the 8-neighborhood of a point $\mathbf{x}$. A threshold is used to reject the maxima of small cornerness, as they are less stable under variations in imaging conditions. The matrix $\mu(\mathbf{x}, \sigma_n)$ is computed with the integration scale $\sigma_I = \sigma_n$ and the local scale $\sigma_D = s\sigma_n$, where s is a constant factor (set to 0.7 in our experiments). For each point we then apply an iterative algorithm that simultaneously detects the location and the scale of interest points. The extrema over scale of the LoG are used to select the scale of interest points. We reject the points for which the LoG response attains no extremum and for which the response is below a threshold. Given an initial point $\mathbf{x}$ with scale σ_I, the iteration steps are:

1. Find the local extremum over scale of the LoG for the point $\mathbf{x}^{(k)}$, otherwise reject the point. The investigated range of scales is limited to $\sigma_I^{(k+1)} = t\sigma_I^{(k)}$ with $t \in [0.7, \ldots, 1.4]$.
2. Detect the spatial location $\mathbf{x}^{(k+1)}$ of a maximum of the Harris measure nearest to $\mathbf{x}^{(k)}$ for the selected σ_I^{k+1}.
3. Go to Step 1 if $\sigma_I^{(k+1)} \neq \sigma_I^{(k)}$ or $\mathbf{x}^{(k+1)} \neq \mathbf{x}^{(k)}$.

The initial points are detected with the multi-scale Harris detector with a large change between two successive detection scales, i.e. 1.4. A small scale change (1.1) is used in the iterative algorithm and provides better accuracy for the location $\mathbf{x}$ and scale σ_I. Given the initial points detected with the scale interval $\xi = 1.4$, the iterative loop scans the range of scales $t\sigma_I$ with $t \in [0.7, \ldots, 1.4]$, which corresponds to the gap between two scale-space levels neighboring the initial point scale σ_I. Note that the initial points detected on the same local structure but at different scales converge to the same location and the same scale (see Fig. 6). It is straightforward to identify these points based on the coordinates and scales. To represent the structure it is sufficient to keep only one of them.

Simplified Harris-Laplace. The Harris-Laplace algorithm can be simplified in order to accelerate the detection of interest points (Mikolajczyk and Schmid, 2001). As before the initial points are detected with the multi-scale Harris detector; we build the scale-space representation with the Harris function and detect local maxima at each scale level. We then verify for each of the initial points whether the LoG attains a maximum at the scale of the point, that is the LoG response is lower for the finer and the coarser scale. We reject the points for which the Laplacian attains no extremum or the response is below a threshold. In this way we obtain a set of characteristic points with associated scales. For some points the scale peak might not correspond to the selected detection scales of an image. These points are either rejected, due to the lack of a maximum, or the location and the scale are not very accurate. Thus the scale interval between two successive levels should be small (i.e. 1.2) to find the location and scale of an interest point with high accuracy.

The Harris-Laplace approach provides a compact and representative set of points which are characteristic in the image and in the scale dimension. The first approach provides higher accuracy in the location and the scale of the interest points. The second approach is a trade-off between accuracy and computational complexity.

Example of Scale Invariant Points. In Fig. 2 we present two examples of points detected with the simplified Harris-Laplace method. The top row shows points detected with the multi-scale Harris detector used for initialization. Here, we manually selected the points corresponding to the same local structure. The detection scale is represented by a circle around the point with radius $3\sigma_I$. Note how the interest point, which is detected for the same image structure, changes its location relative to the detection scale in the gradient direction. One could determine the chain of points and select only one of them to represent the local structure (Alvarez and Morales, 1997; Deriche and Giraudon, 1993). Similar points are located in a small neighborhood and can be determined by comparing their descriptors. However, for local structures existing over a wide range of scales the information content can change (Kadir and Brady, 2001). In our approach the LoG measure is used to select the representative points for such structures. Moreover, the LoG enables the corresponding characteristic points to be selected (bottom row) even if the transformation between images is significant. Sometimes, two or more points are selected from the multi-scale set, but given no prior knowledge about the scale change between images we have to keep all the selected points. As we can see, the

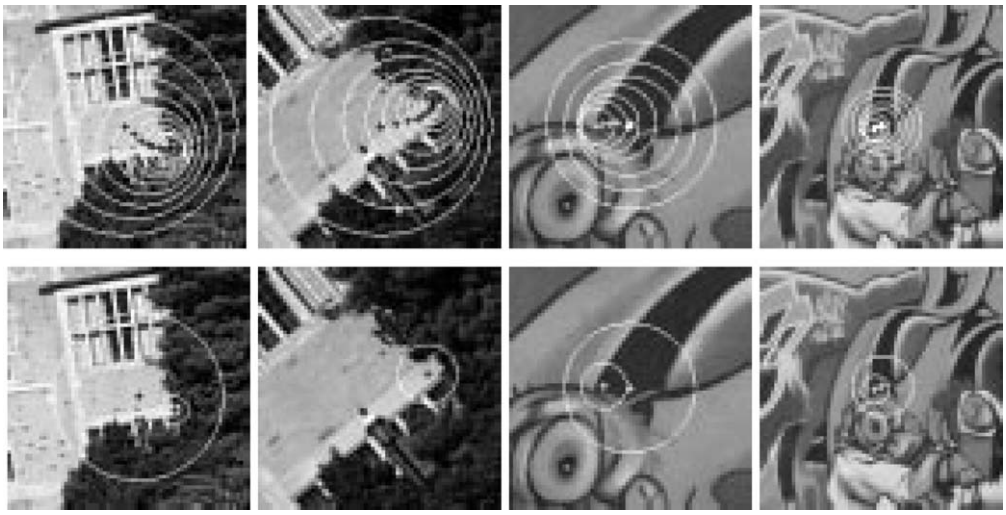


Figure 2. Scale invariant interest point detection: (Top) Initial multi-scale Harris points (selected manually) corresponding to one local structure. (Bottom) Interest points selected with the simplified Harris-Laplace approach.

location and the scale of points is correct with respect to the transformation between the images.

3. Affine Invariant Interest Point Detector

The scale invariant approach can be extended to make it affine invariant. In the following we show how the Harris-Laplace detector behaves in the case of affine transformations of the image. We then introduce the theory which provides a method for estimating the affine shape of a local structure. Each step of the detection algorithm is then discussed in detail and an outline of the iterative procedure is presented. An example of affine invariant points detected with this method is presented.

3.1. Motivation

In the case of affine transformations the scale change is, in general, different in each direction. The Harris-Laplace detector is designed to deal with uniform scale changes and it will therefore fail in the case of significant affine transformations. Figure 3 presents a pair of points detected in images between which there is an affine transformation. The top row shows points detected with the multi-scale Harris detector. The scale, selected with the LoG, is displayed in black. In the bottom row, the Harris-Laplace regions are displayed in black and the superposed white ellipses are the corresponding regions projected from the other image with the affine transformation. We can see that the regions detected with the Harris-Laplace approach do not cover the same part of the affine deformed image.

In the case of an affine transformation, when the scale change is not necessarily the same in every direction, automatically selected scales do not reflect the real transformation of a point. It is well known that the spatial locations of Harris maxima change relatively to the detection scale (Figs. 2 and 3). If the detection scales do not correspond to the real scale factor between the images a shift error is introduced between corresponding points and the associated regions do not correspond. The detection scales have to vary independently in orthogonal directions in order to deal with any affine scaling. Hence, we face the problem of computing the second moment matrix in affine Gaussian scale-space where a circular point neighborhood is replaced by an ellipse. In the next section we show how to deal with this problem.

3.2. Affine Second Moment Matrix

The second moment matrix can be used for estimating the anisotropic shape of a local image structure. This property was explored by Lindeberg (1998) and later by Baumberg (2000) to find the affine deformation of an isotropic structure. In the following we show how

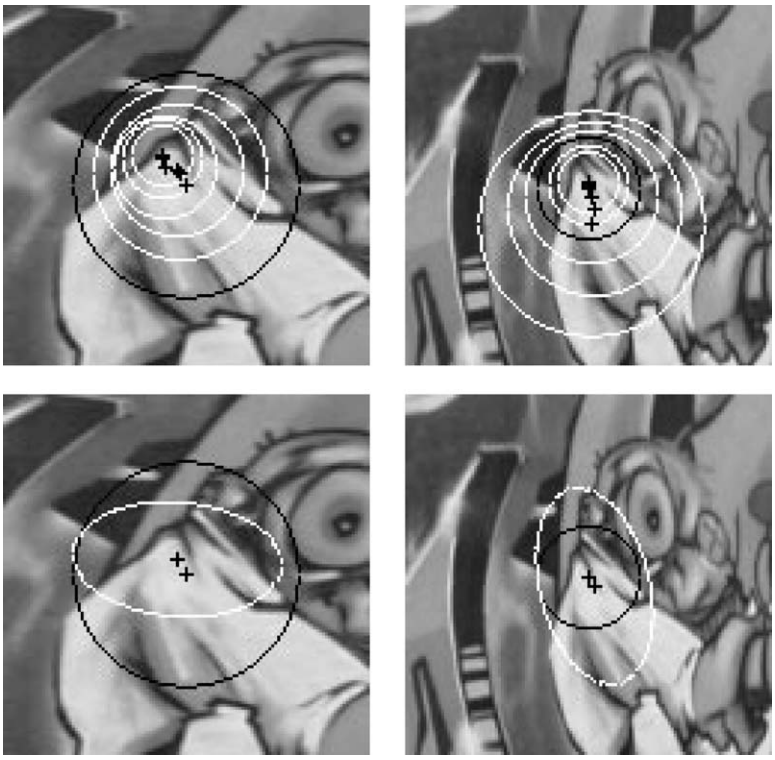


Figure 3. Scale invariant interest point detection in affine transformed images: (Top) Initial interest points detected with the multi-scale Harris detector and their characteristic scales selected by Laplacian scale peak (in black—Harris-Laplace). (Bottom) Characteristic point detected with Harris-Laplace (in black) and the corresponding point from the other image projected with the affine transformation (in white).

to determine the anisotropic shape of a point neighborhood.

In affine scale-space the second moment matrix μ, at a given point $\mathbf{x}$ is defined by:

$$\mu(\mathbf{x}, \Sigma_I, \Sigma_D) = \det(\Sigma_D)\, g(\Sigma_I) * ((\nabla L)(\mathbf{x}, \Sigma_D)(\nabla L)(\mathbf{x}, \Sigma_D)^T) \quad (4)$$

where Σ_I and Σ_D are the covariance matrices which determine the integration and differentiation Gaussian kernels. Clearly, it is unpractical to compute the matrix for all possible combinations of kernel parameters. With little loss of generality we can limit the number of degrees of freedom by setting $\Sigma_I = s\Sigma_D$, where s is a scalar. Hence, the differentiation and the integration kernels will differ only in size and not in shape.

Affine Transformation of Second Moment Matrix. Consider a point $\mathbf{x}_L$ transformed by a linear transformation $\mathbf{x}_R = A\mathbf{x}_L$. The matrix μ_L computed in the point $\mathbf{x}_L$ is then transformed in the following way:

$$\begin{aligned}\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) &= A^T \mu(\mathbf{x}_R, \Sigma_{I,R}, \Sigma_{D,R}) A \\ &= A^T \mu(A\mathbf{x}_L, A\Sigma_{I,L}A^T, A\Sigma_{D,L}A^T) A \end{aligned} \quad (5)$$

If we denote the corresponding matrices by:

$$\mu(\mathbf{x}_L, \Sigma_{I,L}, \Sigma_{D,L}) = M_L \quad \mu(\mathbf{x}_R, \Sigma_{I,R}, \Sigma_{D,R}) = M_R$$

these matrices are then related by:

$$M_L = A^T M_R A \quad M_R = A^{-T} M_L A^{-1} \quad (6)$$

In this case the differentiation and integration kernels are transformed by:

$$\Sigma_R = A\Sigma_L A^T$$

Let us suppose that the matrix M_L is computed in such a way that:

$$\Sigma_{I,L} = \sigma_I M_L^{-1} \quad \Sigma_{D,L} = \sigma_D M_L^{-1} \quad (7)$$

where the scalars σ_I and σ_D are the integration and differentiation scales respectively. We can then derive the following relation:

$$\begin{aligned}\Sigma_{I,R} &= A\Sigma_{I,L}A^T = \sigma_I\left(AM_L^{-1}A^T\right)\\ &= \sigma_I(A^{-T}M_LA^{-1})^{-1} = \sigma_I M_R^{-1}\\ \Sigma_{D,R} &= A\Sigma_{D,L}A^T = \sigma_D\left(AM_L^{-1}A^T\right)\\ &= \sigma_D(A^{-T}M_LA^{-1})^{-1} = \sigma_D M_R^{-1}\end{aligned} \tag{8}$$

This shows that imposing the conditions, defined in Eq. (7) leads to the relations 8, under the assumption that the points are related by an affine transformation and the matrices are computed for corresponding scales σ_I and σ_D. We can now invert the problem and suppose that we have two points related by an unknown affine transformation. If we estimate the matrices Σ_R and Σ_L such that the matrices verify conditions 7 and 8, then relation 6 will be true. This property enables the transformation parameters to be expressed directly by the matrix components. The affine transformation can then be defined by:

$$A = M_R^{-1/2}RM_L^{1/2}$$

where R is an orthogonal matrix which represents an arbitrary rotation or mirror transformation. In the next section we present an iterative algorithm for estimating the matrices Σ_R and Σ_L. The affine transformation can be estimated up to a rotation between two corresponding points without any prior knowledge about this transformation. Furthermore, the matrices M_L and M_R, computed under conditions 7 and 8, determine corresponding regions defined by $\mathbf{x}^TM\mathbf{x} = 1$. If the neighborhood of points $\mathbf{x}_R$ and $\mathbf{x}_L$ are normalized by transformations $\mathbf{x}'_R = M_R^{1/2}\mathbf{x}_R$ and $\mathbf{x}'_L = M_L^{1/2}\mathbf{x}_L$, respectively, the normalized regions are related by a simple rotation $\mathbf{x}'_L = R\mathbf{x}'_R$ (Baumberg, 2000; Garding and Lindeberg, 1994).

$$\begin{aligned}\mathbf{x}_R = A\mathbf{x}_L = M_R^{-1/2}RM_L^{1/2}\mathbf{x}_L,\\ M_R^{1/2}\mathbf{x}_R = RM_L^{1/2}\mathbf{x}_L\end{aligned} \tag{9}$$

The matrices M'_L and M'_R in the normalized frames are equal to a pure rotation matrix (see Fig. 4). In other words, the intensity patterns in the normalized frames are isotropic in terms of the second moment matrix.

Isotropy Measure. The second moment matrix can also be interpreted as an isotropy measure. Without loss of generality we suppose that a local anisotropic structure is an affine transformed isotropic structure. To compensate for the affine deformation, we have to find the transformation that projects the anisotropic pattern to the isotropic one. Note that rotation preserves the isotropy of an image patch, therefore, the affine deformation of an isotropic structure can be determined up to a rotation factor. This rotation can be recovered by methods based on the gradient orientation (Lowe, 1999; Mikolajczyk, 2002). The local isotropy can be measured by the eigenvalues of the second moment matrix $\mu(\mathbf{x}, \sigma_I, \sigma_D)$. If the eigenvalues are equal we consider the point isotropic. To obtain a normalized measure we use the eigenvalue ratio:

$$\mathcal{Q} = \frac{\lambda_{\min}(\mu)}{\lambda_{\max}(\mu)} \tag{10}$$

The value of $\mathcal{Q}$ varies in the range $[0 \ldots 1]$ with 1 for a perfect isotropic structure. This measure can give a slightly different response for different scales as the matrix μ is computed for a given integration and differentiation scale. These scales should be selected independently of the image resolution. The scale selection technique (see Section 2.1) gives the possibility to determine the integration scale related to the local image structure. The differentiation and integration scales can be related by a constant factor s, $\sigma_D = s\sigma_I$. For obvious reasons the differentiation scale should always be smaller than the integration scale. The factor

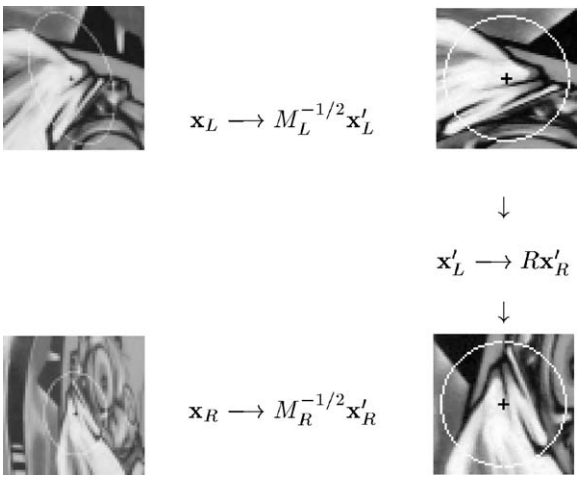


Figure 4. Diagram illustrating the affine normalization based on the second moment matrices. Image coordinates are transformed with matrices $M_L^{-1/2}$ and $M_R^{-1/2}$. The transformed images are related by an orthogonal transformation.

s should not be too small, otherwise the smoothing is too significant with respect to the differentiation. On the other hand s should be small enough, that a Gaussian window of size σ_I can average the covariance matrix $\mu(\mathbf{x}, \sigma_D, \sigma_I)$ in the point neighborhood. The idea is to suppress the noise without suppressing the anisotropic shape of the observed image structures. The solution is to select the differentiation scale σ_D independently of the scale σ_I, that is to vary factor s for example in the range $[0.5, \ldots, 0.75]$. These values are close to those chosen experimentally in the context of the Harris detector (Harris and Stephens, 1988; Schmid and Mohr, 1997). Given the integration scale we search for the scale σ_D for which the response of the isotropy measure attains a local maximum. Thus, the shape selected for the observed structure is closer to an isotropic one. A similar approach for selecting local scale was proposed by Almansa and Lindeberg (2000) and Lindeberg and Garding (1997).

3.3. Harris-Affine Interest Point Detector

In the following we describe our affine invariant approach. We initialize the affine detector with interest points extracted by the multi-scale Harris detector. To determine the *spatial localization* of the interest points we use the Harris detector, which is also based on the second moment matrix, thus it naturally fits in this framework. To obtain the *shape matrix* for each interest point we compute the second moment descriptor with automatically selected *integration* and *differentiation* scales. In our approach the integration and differentiation matrices are related by a scalar $\Sigma_D = s\Sigma_I$ to limit the search space. The outline of our detection method is presented in the following:

- the *spatial localization* of an interest point at a given scale and shape is determined by the local maximum of the Harris function,
- the *integration scale* is selected at the extremum over scale of the normalized Laplacian,
- the *differentiation scale* is selected at the maximum of normalized isotropy,
- the *shape adaptation matrix* is estimated with the second moment matrix and is used to normalize the point neighborhood.

In the following we discuss in detail each step of the algorithm.

Shape Adaptation Matrix. Our iterative shape adaptation method works in the transformed image domain. We transform the image and apply a circular kernel instead of applying the affine Gaussian kernel. This enables the use of a recursive implementation of the Gaussian filters for computing L_x and L_y. The second moment matrix is computed according to Eq. (1). A local window W is centered at interest point $\mathbf{x}$ and transformed by the matrix:

$$U = \prod_k \left(\mu^{-\frac{1}{2}}\right)^{(k)} U^{(0)} \tag{11}$$

in step (k) of the iterative algorithm. In the following we refer to this operation as U-transformation. Note, that a new μ matrix is computed at each iteration and the U matrix is the concatenation of square roots of the second moment matrices. We ensure that the original image is correctly sampled by setting the larger eigenvalue $\lambda_{\max}(U) = 1$, which implies that the image patch is enlarged in the direction of $\lambda_{\min}(U)$. For any given point, the integration and the differentiation scale determine the second moment matrix μ. These scale parameters are automatically selected in each iteration. Thus, the resulting μ matrix is independent of the initial scale and the resolution of the image.

Integration Scale. For any given spatial point we automatically select its characteristic scale. In order to preserve invariance to size changes we select the integration scale σ_I at which the normalized Laplacian (Eq. (3)) attains a local maximum over scale. In the presence of large affine deformations the scale change is very different in each direction. Thus, the characteristic scale detected in the original image and in its U-transformed version can be significantly different. Therefore, it is essential to select the integration scale in each iteration after applying the U transformation. We use a procedure similar to the one in the Harris-Laplace detector. The initial points converge toward a point where the scale and the second moment matrix do not change any more.

Differentiation Scale. We select the local differentiation scale using the integration scale and the isotropy measure $\mathcal{Q}$ (Section 3.2). This solution is motivated by the fact that the local scale has an important influence on the convergence of the second moment matrix. The iterative procedure converges toward a matrix with equal eigenvalues. The smaller the difference between

the eigenvalues ($\lambda_{\max}(\mu)$, $\lambda_{\min}(\mu)$) of the initial matrix, the closer the final solution and the faster the convergence. Note that the Harris measure (Eq. (2)) already selects the points with two large eigenvalues. A large difference between the eigenvalues leads to a large scaling in one direction by the U-transformation. In this case the point does not converge to a stable solution due to noise. The selection of the local scale enables a reasonable eigenvalue ratio to be obtained and the points to converge.

Note that the local differentiation scale can be set proportional to the integration scale $\sigma_D = s\sigma_I$, where s is a constant factor. This significantly accelerates the iterations but some points do not converge due to a large difference between the eigenvalues.

Spatial Localization. We have already shown how the local maxima of the Harris measure change their location if the detection scale changes (Fig. 2). We can also observe this effect when the scale change is different in each direction. In our approach the detection with different scales in x and y directions is replaced by applying the same scale in both directions on the transformed image. Consequently, we re-detect the maximum in the affine normalized window W. Thus, we obtain a vector of displacement to the nearest maximum in the U-normalized window W. The location of the initial point is corrected with the displacement vector back-transformed to the original image domain:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + U^{(k-1)} \cdot \left(\mathbf{x}_w^{(k)} - \mathbf{x}_w^{(k-1)}\right)$$

where $\mathbf{x}_w$ is the point in the coordinates of the U-transformed image.

Convergence Criterion. The important part of the iterative procedure is the stopping criterion. The convergence measure can be based on either the U or the μ matrix. If the criterion is based on μ computed in each iteration, we stop iterating when the matrix is sufficiently close to a pure rotation. This implies that $\lambda_{\max}(\mu)$ and $\lambda_{\min}(\mu)$ are equal. In practice we allow for a small error $\epsilon_C = 0.05$.

$$1 - \frac{\lambda_{\min}(\mu)}{\lambda_{\max}(\mu)} < \epsilon_C \tag{12}$$

Another possibility is to decompose the matrix $U = R^T \cdot D \cdot R$ into rotation R and scaling D and compare the consecutive U-transformations. We stop the iteration if the consecutive R and D transformations are sufficiently similar. Both termination criteria give the same final results. Another important point is to stop in the case of divergence. In theory there is a singular case when the eigenvalue ratio tends to infinity i.e. on a step-edge. Therefore, the point should be rejected if the ratio is too large (i.e. $\epsilon_l = 6$), otherwise it leads to unstable elongated structures.

$$\frac{\lambda_{\max}(D)}{\lambda_{\min}(D)} > \epsilon_l \tag{13}$$

The convergence properties of the shape adaptation algorithm has been extensively studied by Lindeberg and Garding (1997), who showed that except for the singular case the point of convergence is always unique. In general, the procedure converges provided that the initial estimate of the affine deformation is sufficiently close to the true deformation, and the integration scale is correctly selected with respect to the size of the local image structure.

Detection Algorithm. We propose an iterative procedure that allows the initial points to converge to affine invariant points and regions. To initialize our algorithm we use points extracted by the multi-scale Harris detector. These points are not detected in an affine invariant way due to a non-adapted Gaussian kernel, but provide an approximate location and scale for further search. For a given initial interest point $\mathbf{x}^{(0)}$ we apply the following procedure:

1. initialize $U^{(0)}$ to the identity matrix
2. normalize window $W(\mathbf{x}_w) = I(\mathbf{x})$ centered on $U^{(k-1)}\mathbf{x}_w^{(k-1)} = \mathbf{x}^{(k-1)}$
3. select *integration scale* σ_I at point $\mathbf{x}_w^{(k-1)}$
4. select *differentiation scale* $\sigma_D = s\sigma_I$, which maximizes $\frac{\lambda_{\min}(\mu)}{\lambda_{\max}(\mu)}$, with $s \in [0.5, \ldots, 0.75]$ and $\mu = \mu(\mathbf{x}_w{}^{(k-1)}, \sigma_I, \sigma_D)$
5. detect *spatial localization* $\mathbf{x}_w^{(k)}$ of a maximum of the Harris measure (Eq. (2)) nearest to $\mathbf{x}_w^{(k-1)}$ and compute the location of the interest point $\mathbf{x}^{(k)}$
6. compute $\mu_i^{(k)} = \mu^{-\frac{1}{2}}(\mathbf{x}_w^{(k)}, \sigma_I, \sigma_D)$
7. concatenate transformation $U^{(k)} = \mu_i^{(k)} \cdot U^{(k-1)}$ and normalize $U^{(k)}$ to $\lambda_{\max}(U^{(k)}) = 1$
8. go to Step 2 if $1 - \lambda_{\min}(\mu_i^{(k)})/\lambda_{\max}(\mu_i^{(k)}) \geq \epsilon_C$

Although the computation may seem to be very time consuming, note that most time is spent on computing L_x and L_y, which is done only once in each step if the relation between the integration and local scales is constant. The iteration loop begins with selecting the

Figure 5. Iterative detection of an affine invariant interest point in the presence of an affine transformation (top and bottom rows). The first column shows the points used for initialization. The consecutive columns shows the points and regions after iterations 1, 2, 3 and 4. Note that the points converge after 4 iterations and that the ellipses converge to corresponding image regions.

integration scale because we have noticed that this part of the algorithm is most robust to small localization errors of the interest point. However, scale σ_I changes if the shape of the patch is transformed. Given an initial approximate solution, the presented algorithm iteratively modifies the shape, the scale and the spatial location of a point and converges to a local structure. Figure 5 shows affine points detected in consecutive steps of the iterative procedure. After the fourth iteration the location, scale and shape of the point do not change any more. We can notice that the final ellipses cover the same image region despite strong affine deformation.

Selection of Similar Affine Points. We can suppose that the features are stable if they are present at a wide range of scales. These features are identified by several points which converge to the same structure. Provided that the normalized region is isotropic, there is one spatial maximum of the Harris measure and one characteristic scale for the considered local structure. Therefore, several initial points corresponding to the same feature but detected at different scale levels converge toward one point location and scale. It is straightforward to identify these points by comparing their location (x, y), scale σ_I, stretch $\lambda_{\min}(U)/\lambda_{\max}(U)$ and skew. The skew is recovered from the rotation matrix R, where $U = R^T \cdot D \cdot R$. We define a point as similar if each of these parameters is within a threshold to the parameters of the reference point. Finally, we compute the average parameters and select the most similar point from the identified set of points. As a result, for a given image we obtain a set of points where each one represents a different image location and structure.

Example of Affine Invariant Points. Figure 6 illustrates the detection of affine invariant points. Column (a) displays the points used for initialization, which are detected by the multi-scale Harris detector. The circles show the detection scales, where the radius of the circle is $3\sigma_I$. The circles in black show the points selected by the Harris-Laplace detector. Note that there is a significant displacement between points detected at different scales and the circles in corresponding images (top and bottom row) do not cover the same part of the image. In column (b) we show the Harris-Laplace points with estimated affine regions (in black) (Schaffalitzky and Zisserman, 2002). The scale and the location of points is constant during iterations. The projected corresponding regions are displayed in white and clearly show the difference in location and region shape. The initial scale is not correctly detected due to the use of a circular (not affine adapted) Laplacian operator. Similarly, the point locations differ by 3–4 pixels. The points in column (a), which correspond to the same physical structure, but are detected at different locations due to scale, converge to the same point location and region and are displayed in column (c). We can see that the method converges correctly even if the location and the scale of the initial point is relatively far from the point of convergence. Convergence is in general obtained in less than 10 it-

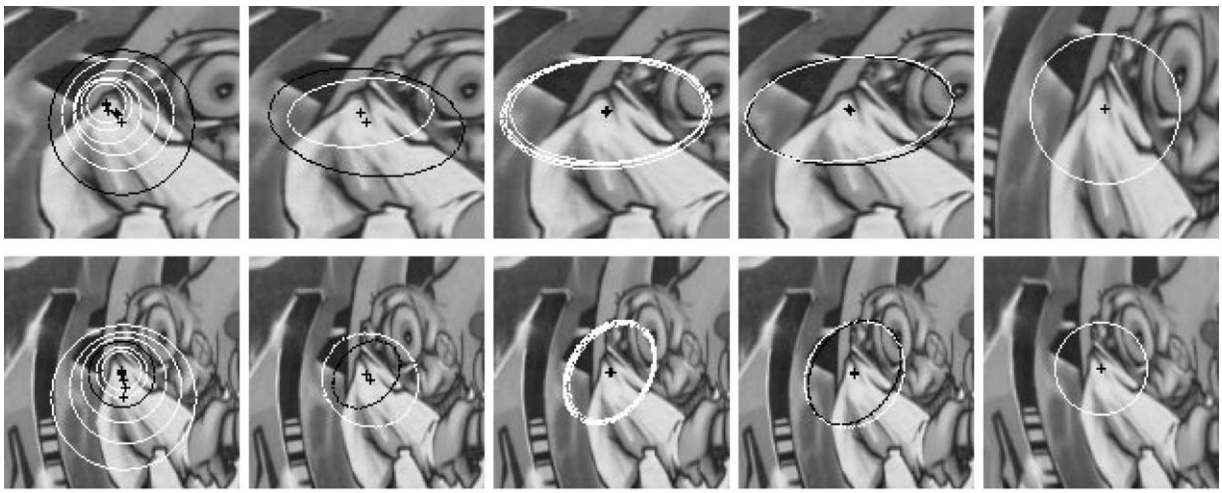


Figure 6. Affine invariant interest point detection: (a) Initial interest points detected with the multi-scale Harris detector and their characteristic scale selected by the Laplacian scale peak (in black—Harris-Laplace). (b) Affine regions detected for the Harris-Laplace points (in black) and the regions projected from the corresponding image (in white). (c) Points and corresponding affine regions obtained with the iterative algorithm applied to the initial multi-scale Harris points. Note that points corresponding to the same structure converge to the same solution. (d) Selected *average* affine points (in black) and its corresponding projected points (in white). (e) Point neighborhoods normalized with the estimated matrices to remove stretch and skew.

erations. Typically, about 40% of the initial points do not converge due to the lack of characteristic scales or to the large difference between the eigenvalues of the matrix U ($\lambda_{\max}(U)/\lambda_{\min}(U) > 6$). About 30% of the remaining points are selected by the similarity measure. About 20–30% of the initial multi-scale Harris points are then used to represent an image. Column (d) displays the selected points (in black) and projected points from the corresponding image (in white). The minor differences between the regions in column (d) are caused by the imprecision of the scale estimation and the error ϵ_C. Column (e) shows the selected points normalized with the estimated matrices to remove the stretch and the skew. We can clearly see that the regions correspond between the two images (top and bottom row).

4. Comparative Evaluation of Interest Points

In this section we compare our scale and affine invariant detectors to other existing approaches presented in Section 1.1. The stability and accuracy of detectors is evaluated using the repeatability criterion introduced in Schmid et al. (2000). We also discuss the performance of different detectors. The important parameters characterizing a feature detector are:

1. The average number of corresponding points detected in images under different geometric and photometric transformations.
2. The accuracy of localization and region estimation.

We present quantitative measures in Section 4.1.

Another important parameter is the distinctiveness of the feature, however this is also a function of the descriptor used. The reader is referred to Mikolajczyk and Schmid (2003a), for a detailed evaluation of different descriptors computed on scale and affine invariant regions.

4.1. Repeatability

Repeatability Criterion. The repeatability score for a given pair of images is computed as the ratio between the number of point-to-point correspondences and the minimum number of points detected in the images. We take into account only the points located in the part of the scene present in both images. We use test images with homographies to find the corresponding regions. We consider that two points $\mathbf{x}_a$ and $\mathbf{x}_b$ correspond if:

1. The error in relative point location is less than 1.5 pixel: $\|\mathbf{x}_a - H \cdot \mathbf{x}_b\| < 1.5$, where H is the homography between the images.

2. The error in the image surface covered by point neighborhoods is $\epsilon_S < 0.4$. In the case of scale invariant points the surface error is:

$$\epsilon_S = \left| 1 - s^2 \frac{\min(\sigma_a^2, \sigma_b^2)}{\max(\sigma_a^2, \sigma_b^2)} \right|$$

where σ_a and σ_b are the selected point scales and s is the actual scale factor recovered from the homography between the images ($s > 1$).

The surface error for affine regions is:

$$\epsilon_S = 1 - \frac{\mu_a \cap (A^T \mu_b A)}{(\mu_a \cup A^T \mu_b A)}$$

where μ_a and μ_b are the elliptic regions defined by $x^T \mu x = 1$. The union of the regions is $(\mu_a \cup (A^T \mu_b A))$ and $(\mu_a \cap (A^T \mu_b A))$ is their intersection. A is the locally linearized homography H in point $\mathbf{x}_b$. The location error of 1.5 pixel is tolerated by descriptors and can be neglected because it introduces a relatively small error between corresponding regions compared to the error introduced by the inaccuracy of the shape estimation. Given the scale interval 1.4 between two successive scale-space levels the maximum scale estimation inaccuracy is $\sqrt{1.4}$. We allow for a slightly larger error 1.3, that is $\epsilon_s < |1 - 1/1.3^2|$, which corresponds to $\epsilon_s < 0.4$.

Data Set. The evaluation is done on real images[1] taken by a digital camera. A significant amount of noise is added during the acquisition process (zoom, viewpoint, light changes, Jpeg compression). The zoom changes involve a change in pixel intensity as automatic camera settings are used. Jpeg compression additionally introduces artifacts. Some of the image pairs are displayed in Section 5.2. In order to use a homography for verification we used planar scenes or 3D scenes with a fixed camera position. The homography between images was estimated using manually selected corresponding points. Each scale change sequence consists of scaled and rotated images, for which the scale factor varies from 1.4 to 4.5. For the viewpoint change sequences the viewpoint varies in the horizontal direction between 0 and 70 degrees. There are 10 images in each sequence representing different scenes. The experiments were carried out using 10 scale change sequences and 6 viewpoint change sequences of real images, one of the sequences is displayed in Fig. 9. There are 160 images in total and approximately 100 000 interest points are detected in these images and used to evaluate the detectors.

Scale Invariant Detectors. In the following we compute the repeatability score for different scale invariant detectors. We compare the detection methods proposed by Lindeberg and Garding (1997) (Laplacian, Hessian and gradient), Lowe (1999) (DoG) as well as our Harris-Laplace and Harris-Affine detector. To show the gain obtained by scale invariance, we also present the results for the standard Harris detector (not adapted to scale changes). Figure 7 shows the repeatability score for the compared methods. The best results are obtained for the Harris-Laplace method. Its repeatability score is 68% for a scale factor of 1.4. The repeatability is not 100% because some points cannot be detected in the corresponding image due to the fixed range of detection scales, which is the same for each image. The points which are extracted at finer scales in the high resolution image and at coarser scales in the coarse resolution image do not have corresponding points. The repeatability score is also influenced by rotation and illumination changes as well as the camera noise. The repeatability of the non-adapted Harris detector is acceptable only for scale changes up to a factor of 1.4. As we might expect LoG and DoG give similar results. The slightly better results for the LoG are due to the artifacts and inaccuracy introduced by sampling of pyramid levels in the DoG approach (Lowe, 1999). The scale invariant detectors perform better than the

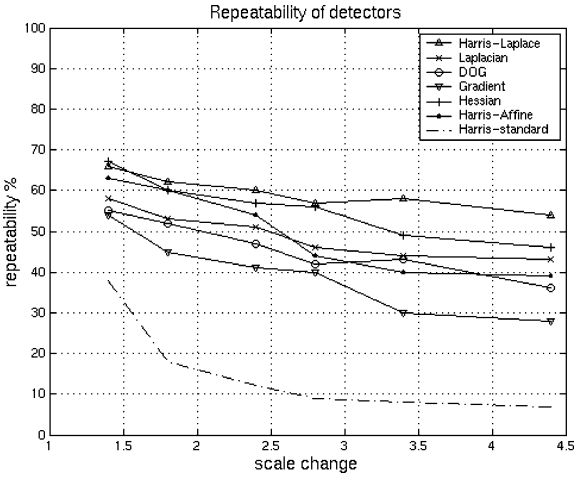


Figure 7. Repeatability of interest point detectors with respect to scale changes. The regions extracted by the detectors are different, therefore the detectors are complementary.

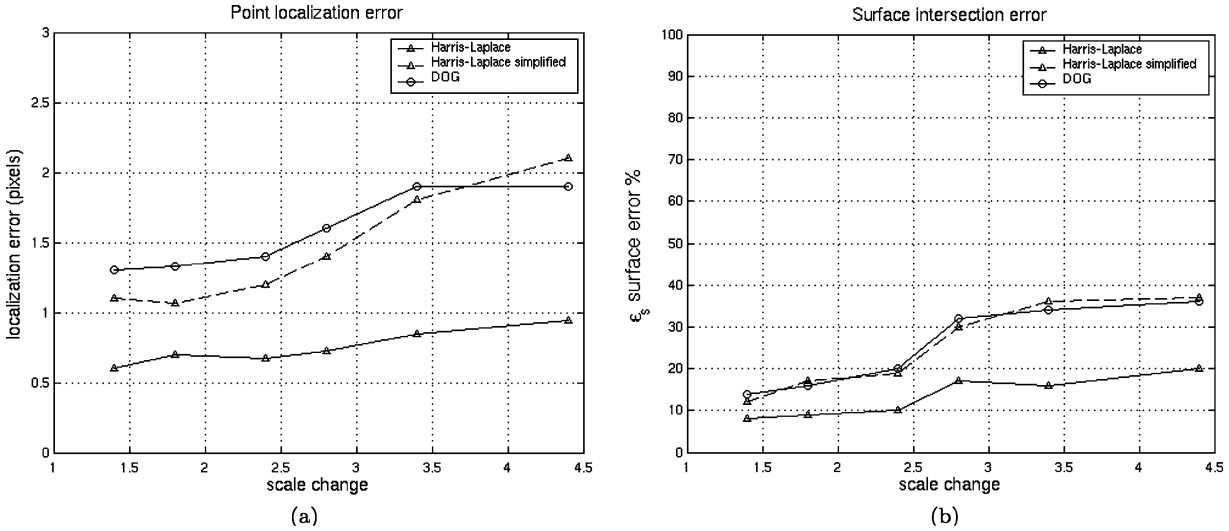


Figure 8. Detection error of corresponding points extracted with scale invariant detectors: (a) relative location and (b) surface intersection ϵ_S.

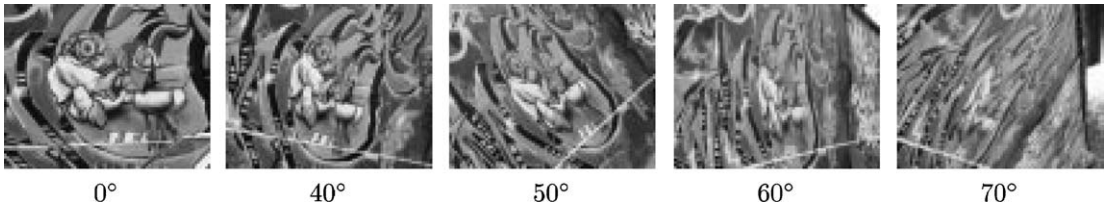


Figure 9. Images of one test sequence with perspective deformations. The corresponding viewpoint angles are indicated below the images.

Harris-Affine approach, but these detectors are appropriate for the uniform scale changes, whereas the affine detector can handle more complex image transformations. Figure 8 shows the accuracy of point locations and scale estimation for Harris-Laplace and the simplified Harris-Laplace. The accuracy is limited by the scale interval which is 1.1 for Harris-Laplace and 1.4 for the simplified Harris-Laplace. In order to measure the accuracy of the localization (Fig. 8(a)) we accept points with localization errors up to 3 pixels. Similarly, for the error of region intersection (Fig. 8(b)), we accept points with the surface error up to 60% and then compute the average error value. We can notice the gain in scale accuracy obtained with iterative Harris-Laplace. The errors are systematically smaller than for the simplified Harris-Laplace.

Affine Invariant Detectors. We have done a similar comparison for Harris-Affine, Harris-Laplace and the approach proposed by Schaffalitzky and Zisserman (2002) referred to as Harris-AffineRegions. Harris-AffineRegions applies the iterative estimation of the affine point neighborhood to Harris-Laplace points. The location and scale of a point remain fixed during iterations.

Figure 10 displays the repeatability rate and Fig. 11 shows the localization and the intersection error for corresponding points. Corresponding points used for computing these errors are determined by the homography. We used the same criteria to compute the localization and intersection error as for the scale invariant detectors. The affine transformation for the error estimation is computed with a local approximation of the homography.

We notice in Fig. 10 that our affine detector significantly improves the results in the case of strong affine deformations. We can notice the breakdown point of the Harris-Laplace detector at a viewpoint change of 40 degrees. The performance of Harris-Laplace continue to decrease, whereas Harris-Affine still provides

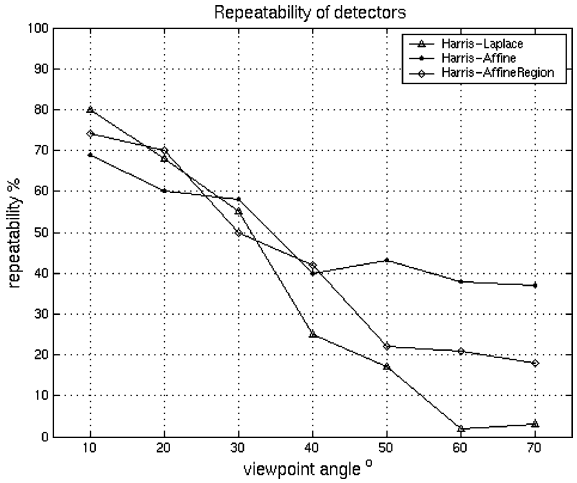


Figure 10. Repeatability of detectors: Harris-Affine—approach proposed in this paper, Harris-AffineRegions—Harris-Laplace detector with affine normalization of the point neighborhood, Harris-Laplace—multi-scale Harris detector with characteristic scale selection.

sufficient corresponding features. The accuracy of the feature localization and shape is critical for local descriptors, for example, differential descriptors fail if this error is significant (Mikolajczyk and Schmid, 2003a). The improvement is with respect to localization as well as region intersection (Fig. 11). These results clearly show that the location of the maximum of the Harris measure and the extremum over scale are significantly influenced by affine transformations. In the presence of weak affine distortions the Harris-Laplace and the Harris-AffineRegions detectors achieve the best results. The localization error is the same for these two detectors. The difference in the surface error is insignificant for small viewpoint changes. The affine adaptation does not improve the location, the scale, and the region shape because the scaling is almost the same in every direction. The circular Gaussian kernel is well suited for this case. The other scale invariant detectors give worse results than those of Harris-Laplace, if applied on images with affine transformations. Note, that the relative rank of detectors does not change compared to Fig. 7. For clarity we show the results only for the Harris-Laplace.

4.2. *Computational Complexity*

The complexity and efficiency of a feature detector is an important issue in particular when applying the detectors to image sequences or large image databases. Table 1 shows a comparison of the computation time required by the detectors. Here, each detector is applied to an image of size 800 × 640 (displayed in Fig. 12). Detection is done on a Pentium II 500 MHz. The first column lists the detectors and the second column shows the main operations required for detecting the initial points. The points are detected at 12 scale levels. Note that to obtain the Hessian or the second

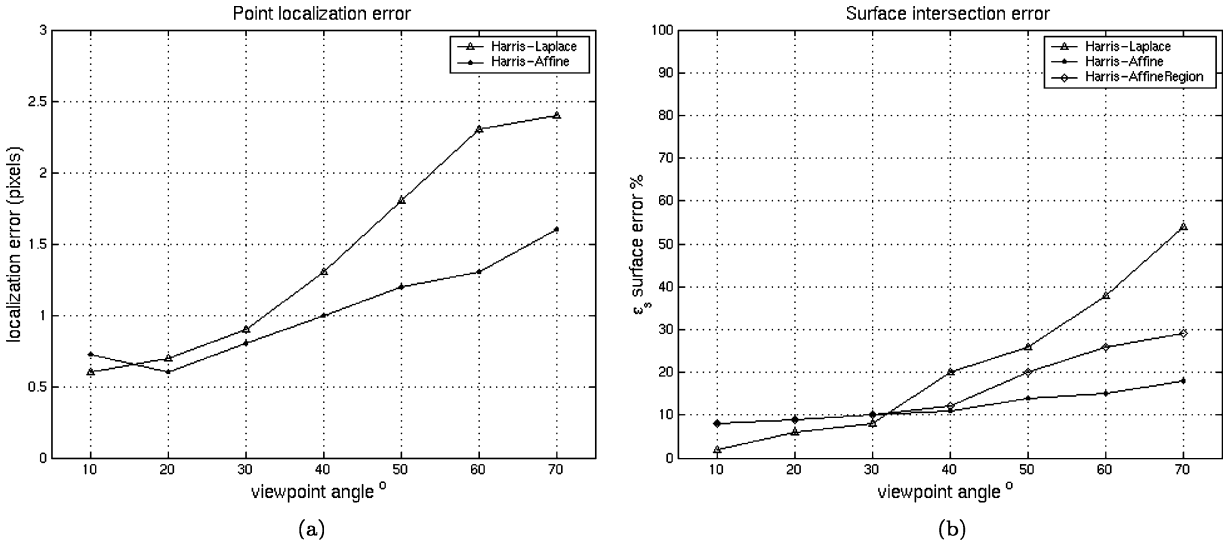


Figure 11. Detection error of corresponding points extracted with affine invariant detectors: (a) relative location (the same for Harris-Laplace and Harris-AffineRegions) and (b) surface intersection ϵ_S.

Table 1. Complexity of the detectors. $g(I)$ denotes Gaussian smoothing. $H(I)$ denotes the Hessian matrix and $\mu(I)$ the second moment matrix computed for every image point. $(d_{xx} + d_{yy})$ is a convolution of a point neighborhood with a 2D Laplacian kernel. #n denotes the number of iterations per point patch, and can vary for different initial points.

Detector	Operation on image (initial points)	Operation on patch (scale)	Operation on patch (shape)	Run time seconds	Number of points
DoG	#12 $g(I)$			0.7	1527
Hessian	#12 $H(I)$			0.9	1832
H-L simplified	#12 $\mu(I)$	#3 $(d_{xx} + d_{yy})$		1.4	1625
H-L	#12 $\mu(I)$	#n $(d_{xx} + d_{yy})$		7	1438
H-AR	#12 $\mu(I)$	#3 $(d_{xx} + d_{yy})$	#n $\mu(\mathbf{x})$	12	1463
H-A	#12 $\mu(I)$	#7n $(d_{xx} + d_{yy})$	#5n $\mu(\mathbf{x})$	36	1123

moment matrix we compute and smooth the derivatives for each image point. In this implementation we use recursive filters to accelerate the Gaussian filtering. We have compared this recursive implementation with non-optimized Gaussian filtering. The number of detected points differ by 0.5% due to slightly different responses of regular Gaussian filters. The shape of the second moment matrices remains the same. Every initial point is processed independently. The simplified Harris-Laplace approach requires 3 convolutions $(\sigma_{n-1}, \sigma_n, \sigma_{n+1})$ of a point neighborhood with a 2D Laplacian kernel to select the scale (third column). The number of convolutions is larger for the iterative Harris-Laplace method and varies for each initial point. Typically, #n is less than 5, and the maximum number of iterations is limited to 10. The Harris-AffineRegion method selects the scale and then iterates on local shape, therefore it computes the second moment matrix at each iteration step. Typically, #n is less than 10, and the maximum number of iterations is limited to 15. The Harris-Affine approach probes 7 integration scales (third column) and 5 differentiation scales (fourth column) at each iteration to find local extrema. The number of iterations is similar to the Harris-AffineRegion method. The fifth column shows the run time in seconds and the sixth the number of points provided by the detectors. The run time is the computational time required by a Pentium II 500 MHz to detect features in a 800×640 image. This time can slightly vary depending on the number of features in the image.

The fastest detector is DoG since it only smooths, subtracts and samples the image. The Harris-Affine (H-A) detector is the one with the highest complexity. It can be significantly accelerated by fixing the ratio between the differentiation and integration scales. This will reduce the number of iterations on $\mu(\mathbf{x})$ from #5n to #n times, where 5 is the number of probed differentiation scales. The scale selection and the point localization can be done at the first iteration only, in a similar manner to the Harris-AffineRegion method. All these simplifications can significantly reduce the detection time but at the cost of accuracy.

5. Applications

In this section we present an example application for our interest point detectors and show how they can be used to match image pairs with significant scale or viewpoint changes. For examples of other applications the reader is referred to Lazebnik et al. (2003), Rothganger et al. (2003), and Schaffalitzky and Zisserman (2002). In Section 5.1 we describe our matching approach. Section 5.2 shows the results for scale and affine invariant features.

5.1. Matching Algorithm

Given an image we detect a set of interest points and compute the point descriptors. The descriptors are then compared with a similarity measure. The resulting similarity is used for finding the corresponding points.

Descriptors and Similarity Measure. Our descriptors are Gaussian derivatives computed in the local neighborhood of interest points. Derivatives are computed on image patches normalized with the matrix U (Eq. (11)), which is estimated independently for each point. Invariance to rotation is obtained by

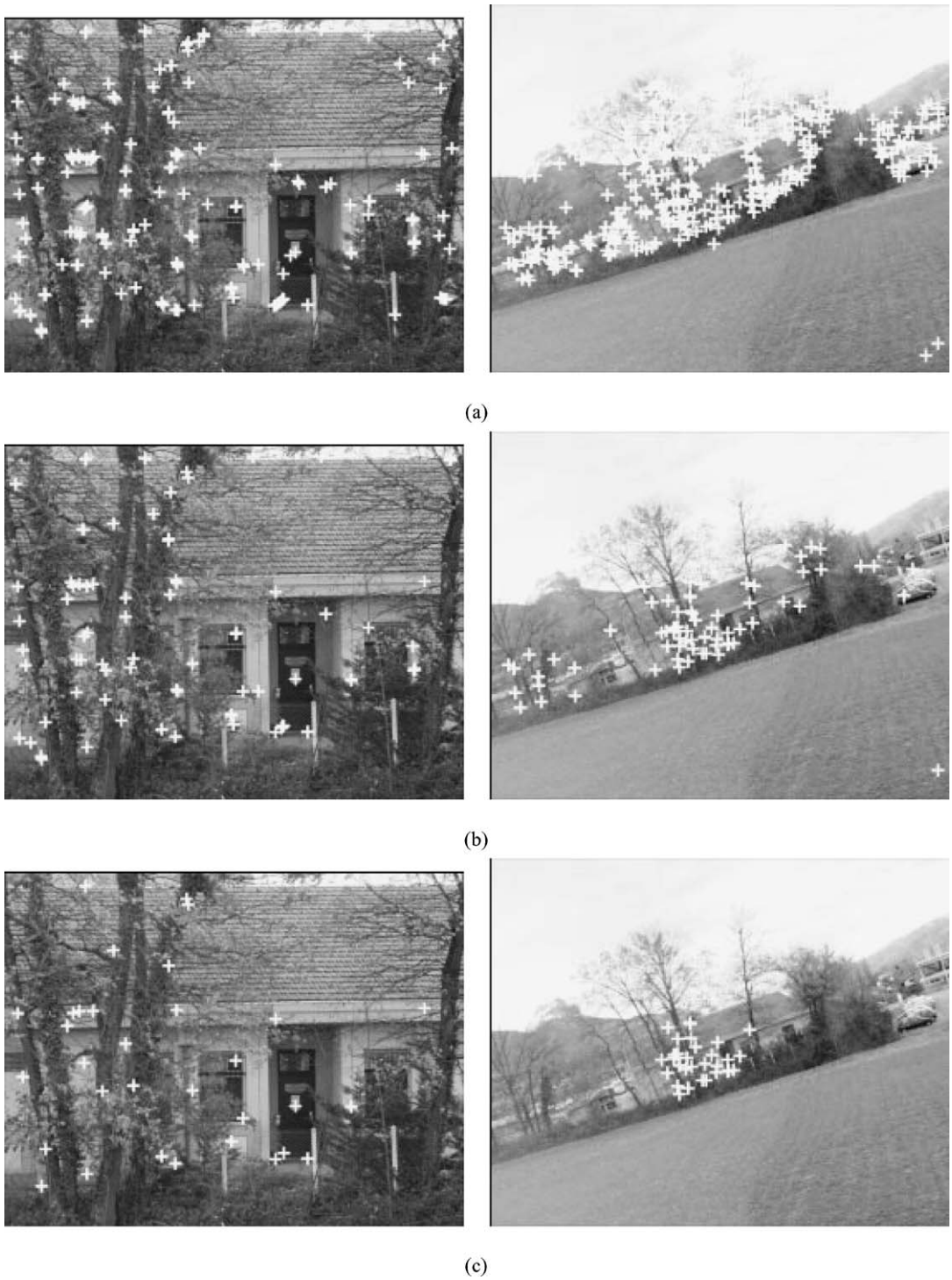


Figure 12. Robust matching: Harris-Laplace detects 190 and 213 points in the left and right images, respectively (a). 58 points are initially matched (b). There are 32 inliers to the estimated homography (c), all of which are correct. The estimated scale factor is 4.9 and the estimated rotation angle is 19 degrees.

"steering" the derivatives in the direction of the gradient (Freeman and Adelson, 1991). To obtain a stable estimate of the gradient direction, we use the average gradient orientation in a point neighborhood (Mikolajczyk, 2002). Invariance to affine intensity changes is obtained by dividing the higher order derivatives by the first derivative. We obtain descriptors of dimension 12 by using derivatives up to 4th order.

To measure the similarity between the descriptors we use the Mahalanobis distance. The covariance matrix is estimated over a large set of images and incorporates signal noise, variations in photometry as well as inaccuracy of the interest point location.

Matching. To robustly match the images, we first determine point-to-point correspondences using the similarity measure. We select for each descriptor in the first image the most similar descriptor in the second image using the Mahalanobis distance. If the distance is below a threshold the match is potentially correct. A set of initial matches is obtained. In the second step of verification we apply cross-correlation, which rejects low-score matches. Finally, a robust estimation of the transformation between the two images based on RANdom SAmple Consensus (RANSAC) enables the selection of the inliers. In our experiments the transformation is either a homography or a fundamental matrix. A model selection algorithm (Kanatani, 1998; Triggs, 2001) can be used to automatically decide which transformation is the most appropriate.

5.2. Experimental Results for Matching

In this section, we present matching results in the presence of scale and viewpoint changes. The results are obtained with the Harris-Laplace and the Harris-Affine detector. We show the matched points which are inliers to the estimated transformations. The number of correctly matched descriptors is limited by the number of corresponding features provided by the detector and depends on the accuracy of the detectors. The matching approach is based on the distance measure between the descriptors and RANSAC. If the fraction of inliers among the initial matches is too small then RANSAC fails. Note that there are points which are correctly detected but are rejected by the distance measure. However, these points could be matched by using a more distinctive descriptor or by applying semi-local constraints.

Scale Change. Figure 12 illustrates the consecutive steps of the matching algorithm. In this example two images are taken from the same viewpoint, but with a zoom change and camera rotation. The multi-scale Harris detector provides 1382 and 926 points for the images, respectively. The best ratio inliers/initial matches obtained by varying the distance threshold was 41/220. The fraction of outliers is too significant and RANSAC fails. This ratio for Harris (not adapted to scale changes) is 4/140. Moreover, these 4 points are accidentally matched since the size of the point neighborhood used to compute the descriptors is the same for both images. This clearly shows that the multi-scale Harris detector needs a more efficient matching strategy and the non-adapted Harris detector cannot deal with scale changes. The ratio inliers/initial matches for Harris-Laplace is 32/58 with a distance threshold fixed for all image pairs. The top row shows the interest points detected with the Harris-Laplace detector. There are 190 and 213 points detected in the left and right images respectively. These numbers are about equivalent to the number of points which are usually detected with the standard Harris detector applied at the finest level of the scale-space representation. Note that there are about 10 times more points if the multi-scale Harris detector is used. This clearly shows the selectivity of our method. Row (b) shows the 58 matches obtained by the initial matching with the similarity measure. Row (c) 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 rotation angle is 19 degrees.

Another example is displayed in Fig. 14(a). There is a scale change of 3.9 and a rotation of 17° between the images. There are 118 correctly matched points. In the presence of uniform scale changes the Harris-Laplace detector performs better than the Harris-Affine detector. The Harris-Affine approach estimates the affine deformation of features, which rejects many points with correct scale and location but with highly anisotropic shape. The affine invariant points are also less distinctive.

Viewpoint Change. Figure 13 illustrates the matching results with features provided by Harris-Affine detector. In order to separate the detection and the matching results, we present in row (a) all the possible point-to-point correspondences established with the estimated homography. There are 78 corresponding pairs among the 287 and 325 points detected in the first and the second image, respectively. After matching with the

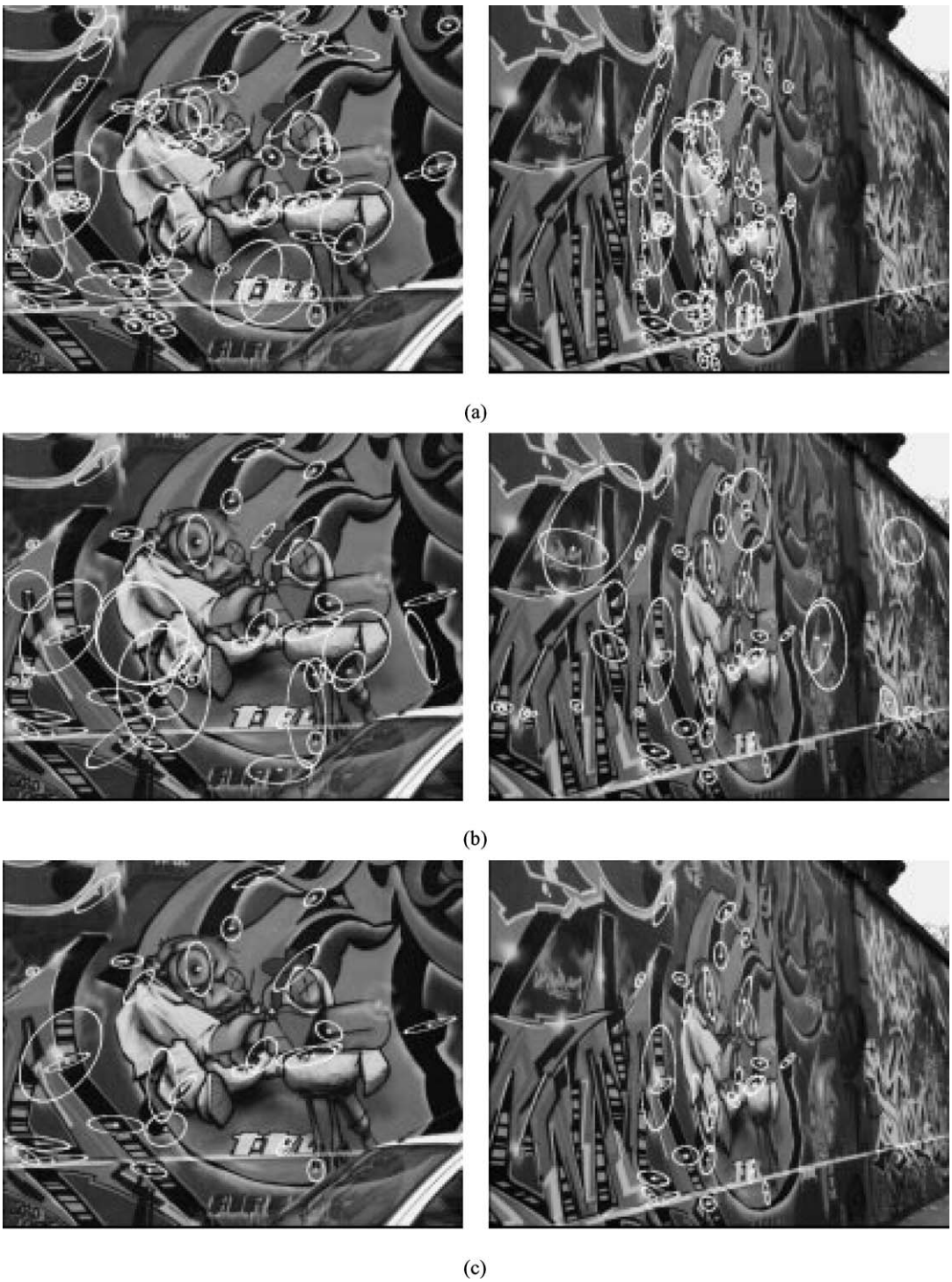
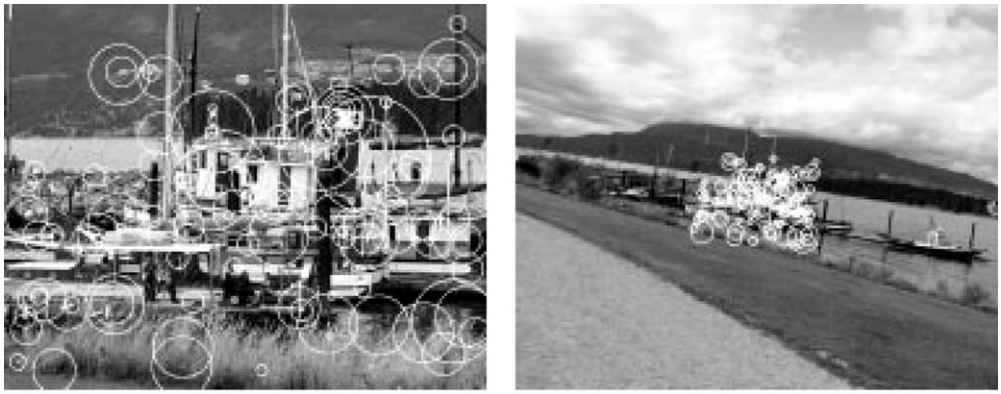


Figure 13. Robust matching: (a) 78 pairs of possible matches are found among the 287 and 325 points detected by Harris-Affine. (b) 43 points are matched based on the descriptors and the cross-correlation score. 27 of these matches are correct. (c) 27 are inliers to the estimated homography. All of them correct.



(a) Scale change of 3.9 and rotation of 17°.

(b) Scale change of 1.8 and viewpoint change of 30°

(c) Scale change of 1.7 and viewpoint change of 50°

Figure 14. Correctly matched images using scale and affine regions. The displayed matches are the inliers to a robustly estimated homography or fundamental matrix. There are (a) 118 matches (b) 34 matches and (c) 22 matches. All of them are correct.

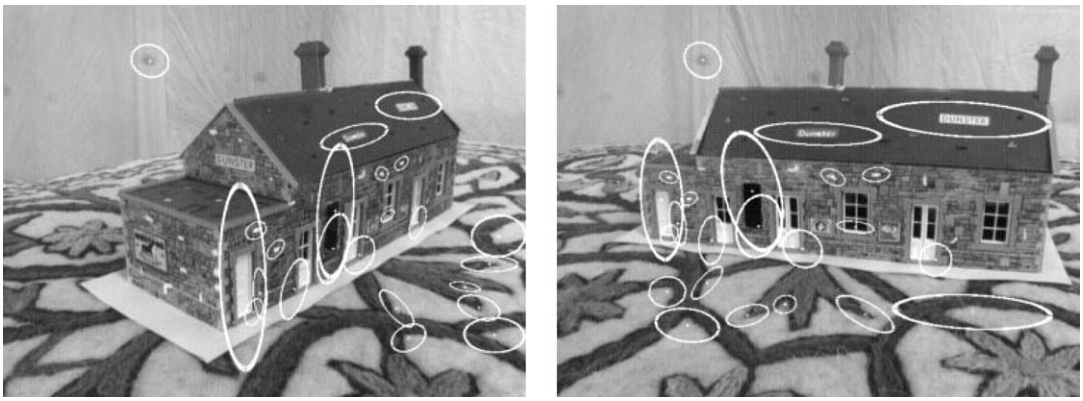


Figure 15. Example of an image pair, for which our matching approach fails. However, there are correctly detected corresponding points which we have manually selected. The failure is therefore due to descriptors.

similarity measure, we obtain 53 matches (29 correct and 24 incorrect). Next, we apply the additional verification based on the cross-correlation of affine normalized image patches. This verification rejects 10 matches (2 correct and 8 incorrect). The remaining 43 matches (27 correct and 16 incorrect) are displayed in row (b). Finally, there are 27 inliers to the robustly estimated homography, which are presented in row (c). Note, that there is a large perspective transformation between the images. The limited benefit of using cross-correlation can be explained by a high sensitivity of this method to different types of errors introduced by the feature detector such as inaccuracy in the feature localization, scale and affine normalization. Other examples are presented in Fig. 14(b) and (c). The images show a 3D scene and a planar scene taken from different viewpoints. Points are detected with Harris-Affine and there are 34 inliers to a robustly estimated fundamental matrix (Fig. 14(b)) and 22 inliers to a homography (Fig. 14(c)).

In Fig. 15, we show a pair of images for which our matching procedure fails. It shows that there are at least 23 similar regions that could be matched. The failure is therefore not due to the Harris-Affine detector, but to the matching procedure. It is true that affine-invariant descriptors are less distinctive. For example, corners of sharp or wide angles, of light or dark intensity are almost the same once normalized to be geometrically as well as photometrically invariant. Therefore, improving the matching is necessary to match these two images. This can be achieved by using (i) more distinctive descriptors (see Mikolajczyk and Schmid, 2003a for a performance evaluation of different descriptors computed for affine-invariant regions) or (ii) semi-local geometric consistency (Dufournaud et al., 2000; Pritchett and Zisserman, 1998; Tell and Carlsson, 2002).

6. Conclusions and Future Work

In this paper we have proposed two novel approaches for scale and affine invariant interest point detection. Our algorithm simultaneously adapts location, scale and shape of a point neighborhood to obtain affine invariant points. None of the previous methods simultaneously solves for all of these parameters in a feature extraction algorithm. The experimental results for wide baseline matching show the performance of our approach. The scale invariant detector can deal with larger scale changes than the affine invariant detector but it fails for images with large affine transformations. The affine invariant points provide for reliable matching even for images with significant perspective deformations. However, the stability and convergence of affine regions is the subject of further investigation as well as their robustness to occlusions.

The invariance to geometric and photometric affine transformations removes some of the information that the points convey, therefore the design of a more robust and distinctive descriptor is required. It might then be combined with semi-local constraints (Dufournaud et al., 2000; Pritchett and Zisserman, 1998; Schmid and Mohr, 1997; Tell and Carlsson, 2002) to improve the results. A future area of work will also be the use of the proposed approaches in different applications, as for example, shot matching in a video sequence, recognition of object classes and tracking.

Note

1. http://www.inrialpes.fr/lear/people/Mikolajczyk/Database

References

Almansa, A. and Lindeberg, T. 2000. Fingerprint enhancement by shape adaptation of scale-space operators with automatic scale selection. *IEEE Transactions on Image Processing*, 9(12):2027–2042.

Alvarez, L. and Morales, F. 1997. Affine morphological multiscale analysis of corners and multiple junctions. *International Journal of Computer Vision*, 2(25):95–107.

Baumberg, A. 2000. Reliable feature matching across widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, pp. 774–781.

Borenstein, E. and Ullman, S. 2002. Class-specific, top-down segmentation. In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, pp. 202–215.

Brand, P. and Mohr, R. 1994. Accuracy in image measure. In *Proceedings of the SPIE Conference on Videometrics III*, S.F. El-Hakim (Ed.). Boston, Massachusetts, USA, vol. 2350, pp. 218–228.

Bretzner, L. and Lindeberg, T. 1998. Feature tracking with automatic selection of spatial scales. *Computer Vision and Image Understanding*, 71(3):385–392.

Brown, M. and Lowe, D.G. 2002. Invariant features from interest point groups. In *The 13th British Machine Vision Conference*, Cardiff University, UK, pp. 253–262.

Chomat, O., de Verdière, V.C., Hall, D., and Crowley, J. 2000. Local scale selection for Gaussian based description techniques. In *Proceedings of the 6th European Conference on Computer Vision*, Dublin, Ireland, pp. 117–133.

Cottier, J. 1994. Extraction et appariements robustes des points d'intérêt de deux images non etalonnées.

Crowley, J. 1981. A representation for visual information. PhD thesis, Carnegie Mellon University.

Crowley, J. and Parker, A. 1984. A representation for shape based on peaks and ridges in the difference of low pass transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):156–170.

Deriche, R. and Giraudon, G. 1993. A computational approach for corner and vertex detection. *International Journal of Computer Vision*, 10(2):101–124.

Duda, R. and Hart, P. 1973. *Pattern Classification and Scene Analysis*. Wiley-Interscience.

Dufournaud, Y., Schmid, C., and Horaud, R. 2000. Matching images with different resolutions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina, USA, pp. 612–618.

Förstner, W. 1994. A framework for low level feature extraction. In *Proceedings of the 3rd European Conference on Computer Vision*, Stockholm, Sweden, pp. 383–394.

Förstner, W. and Gülchm, E. 1987. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Intercommission Conference on Fast Processing of Photogrammetric Data*, Interlaken, Switzerland, pp. 281–305.

Freeman, W. and Adelson, E. 1991. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906.

Garding, J. and Lindeberg, T. 1994. Direct estimation of local surface shape in a fixating binocular vision system. In *Proceedings of the 3rd European Conference on Computer Vision*, Stockholm, Sweden, pp. 365–376.

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

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., and Kuebler, O. 1992. Simulation of neural contour mechanism: From simple to end-stopped cells. *Vision Research*, 32(5):963–981.

Horaud, R., Skordas, T., and Veillon, F. 1990. Finding geometric and relational structures in an image. In *Proceedings of the 1st European Conference on Computer Vision*, Antibes, France, pp. 374–384.

Kadir, T. and Brady, M. 2001. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105,

Kanatani, K. 1998. Geometric information criterion for model selection. *International Journal of Computer Vision*, 26(3):171–189.

Laptev, I. and Lindeberg, T. 2001. Tracking of multi-state hand models using particle filtering and a hierarchy of multi-scale image features. In *Proceedings of Scale Space and Morphology Workshop*, Vancouver, Canada, vol. 2106. Lecture Notes in Computer Science, pp. 63–74.

Lazebnik, S., Schmid, C., and Ponce, J. 2003. Sparse texture representation using affine-invariant neighborhoods. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, pp. 319–324.

Lindeberg, T. 1993. Detecting salient blob-like image structures and their scales with a scale-space primal sketch—A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318.

Lindeberg, T. 1998. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116.

Lindeberg, T. and Garding, J. 1997. 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.

Lowe, D.G. 1999. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision*, Kerkyra, Greece, pp. 1150–1157.

Matas, J., Chum, O., Urban, M., and Pajdla, T. 2002. Robust wide baseline stereo from maximally stable extremal regions. In *The 13th British Machine Vision Conference*, Cardiff University, UK, pp. 384–393.

Mikolajczyk, K. 2002. Interest point detection invariant to affine transformations. PhD thesis, Institut National Polytechnique de Grenoble.

Mikolajczyk, K. and Schmid, C. 2001. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision*, Vancouver, Canada, pp. 525–531.

Mikolajczyk, K. and Schmid, C. 2002. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, vol. I, pp. 128–142.

Mikolajczyk, K. and Schmid, C. 2003. An performance evaluation of local descriptors. In *Proceedings of the Conference on*

Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, pp. 257–264.

Mikolajczyk, K. and Schmid, C. 2003. Shape recognition with edge based features. In *Proceedings of the 13th British Machine Vision Conference*, Norwich, UK, pp. 779–788.

Pritchett, P. and Zisserman, A. 1998. Wide baseline stereo matching. In *Proceedings of the 6th International Conference on Computer Vision*, Bombay, India. IEEE Computer Society Press, pp. 754–760.

Rothganger, F., Lazebnik, S., Schmid, C., and Ponce, J. 2003. 3D Object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, pp. 272–277.

Schaffalitzky, F. and Zisserman, A. 2001. Viewpoint invariant texture matching and wide baseline stereo. In *Proceedings of the 8th International Conference on Computer Vision*, Vancouver, Canada, pp. 636–643.

Schaffalitzky, F. and Zisserman, A. 2002. Multi-view matching for unordered image sets. In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, pp. 414–431.

Schmid, C. and Mohr, R. 1997. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534.

Schmid, C., Mohr, R., and Bauckhage, C. 2000. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172.

Tell, D. and Carlsson, S. 2002. Combining appearance and topology for wide baseline matching. In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, pp. 814–828.

Triggs, B. 2001. Joint feature distributions for image correspondence. In *Proceedings of the 8th International Conference on Computer Vision*, Vancouver, Canada, pp. 201–208.

Tuytelaars, T. and Gool, L.V. 1999. Content-based image retrieval based on local affinely invariant regions. In *Int. Conf. on Visual Information Systems*, pp. 493–500.

Tuytelaars, T. and Van Gool, L. 2000. Wide baseline stereo matching based on local, affinely invariant regions. In *The 11th British Machine Vision Conference*, University of Bristol, UK, pp. 412–425.