International Journal of Computer Vision 65(1/2), 43–72, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.
DOI: 10.1007/s11263-005-3848-x

A Comparison of Affine Region Detectors

K. MIKOLAJCZYK
University of Oxford, OX1 3PJ, Oxford, United Kingdom
km@robots.ox.ac.uk

T. TUYTELAARS
University of Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
tuytelaa@esat.kuleuven.be

C. SCHMID
INRIA, GRAVIR-CNRS, 655, av. de l'Europe, 38330, Montbonnot, France
schmid@inrialpes.fr

A. ZISSERMAN
University of Oxford, OX1 3PJ, Oxford, United Kingdom
az@robots.ox.ac.uk

J. MATAS
Czech Technical University, Karlovo Namesti 13, 121 35, Prague, Czech Republic
matas@cmp.felk.cvut.cz

F. SCHAFFALITZKY AND T. KADIR
University of Oxford, OX1 3PJ, Oxford, United Kingdom
fsm@robots.ox.ac.uk
tk@robots.ox.ac.uk

L. VAN GOOL
University of Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
vangool@esat.kuleuven.be

Received August 20, 2004; Revised May 3, 2005; Accepted May 11, 2005

First online version published in January, 2006

Abstract. The paper gives a snapshot of the state of the art in affine covariant region detectors, and compares their performance on a set of test images under varying imaging conditions. Six types of detectors are included: detectors based on affine normalization around Harris (Mikolajczyk and Schmid, 2002; Schaffalitzky and Zisserman, 2002) and Hessian points (Mikolajczyk and Schmid, 2002), a detector of 'maximally stable extremal regions', proposed by Matas et al. (2002); an edge-based region detector (Tuytelaars and Van Gool, 1999) and a detector based on intensity extrema (Tuytelaars and Van Gool, 2000), and a detector of 'salient regions', proposed by Kadir, Zisserman and Brady (2004). The performance is measured against changes in viewpoint, scale, illumination, defocus and image compression.

The objective of this paper is also to establish a reference test set of images and performance software, so that future detectors can be evaluated in the same framework.

Keywords: affine region detectors, invariant image description, local features, performance evaluation

1. Introduction

Detecting regions covariant with a class of transformations has now reached some maturity in the computer vision literature. These regions have been used in quite varied applications including: wide baseline matching for stereo pairs (Baumberg, 2000; Matas et al., 2002; Pritchett and Zisserman, 1998; Tuytelaars and Van Gool, 2000), reconstructing cameras for sets of disparate views (Schaffalitzky and Zisserman, 2002), image retrieval from large databases (Schmid and Mohr, 1997; Tuytelaars and Van Gool, 1999), model based recognition (Ferrari et al., 2004; Lowe, 1999; Obdržálek and Matas, 2002; Rothganger et al., 2003), object retrieval in video (Sivic and Zisserman, 2003; Sivic et al., 2004), visual data mining (Sivic and Zisserman, 2004), texture recognition (Lazebnik et al., 2003a,b), shot location (Schaffalitzky and Zisserman, 2003), robot localization (Se et al., 2002) and servoing (Tuytelaars et al., 1999), building panoramas (Brown and Lowe, 2003), symmetry detection (Turina et al., 2001), and object categorization (Csurka et al., 2004; Dorko and Schmid, 2003; Fergus et al., 2003; Opelt et al., 2004).

The requirement for these regions is that they should correspond to the same pre-image for different viewpoints, i.e., their shape is not fixed but automatically adapts, based on the underlying image intensities, so that they are the projection of the same 3D surface patch. In particular, consider images from two viewpoints and the geometric transformation between the images induced by the viewpoint change. Regions detected after the viewpoint change should be the same, modulo noise, as the transformed versions of the regions detected in the original image–image transformation and region detection commute. As such, even though they have often been called *invariant* regions in the literature (e.g., Dorko and Schmid, 2003; Lazebnik et al., 2003a; Sivic and Zisserman, 2004; Tuytelaars and Van Gool, 1999), in principle they should be termed *covariant* regions since they change covariantly with the transformation. The confusion probably arises from the fact that, even though the regions themselves are covariant, the normalized image pattern they cover and the feature descriptors derived from them are typically invariant.

Note, our use of the term ‘region’ simply refers to a set of pixels, i.e. any subset of the image. This differs from classical segmentation since the region boundaries do not have to correspond to changes in image appearance such as colour or texture. All the detectors presented here produce simply connected regions, but in general this need not be the case.

For viewpoint changes, the transformation of most interest is an *affinity*. This is illustrated in Fig. 1. Clearly, a region with fixed shape (a circular example is shown in Fig. 1(a) and (b)) cannot cope with the geometric deformations caused by the change in viewpoint. We can observe that the circle does not cover the same image content, i.e., the same physical surface. Instead, the shape of the region has to be adaptive, or covariant with respect to affinities (Fig. 1(c)–close-ups shown in Fig. 1(d)–(f)). Indeed, an affinity is sufficient to locally model image distortions arising from viewpoint changes, provided that (1) the scene surface can be locally approximated by a plane or in case of a rotating camera, and (2) perspective effects are ignored, which are typically small on a local scale anyway. Aside from the geometric deformations, also photometric deformations need to be taken into account. These can be modeled by a linear transformation of the intensities.

To further illustrate these issues, and how affine covariant regions can be exploited to cope with the geometric and photometric deformation between wide baseline images, consider the example shown in Fig. 2. Unlike the example of Fig. 1 (where a circular region was chosen for one viewpoint) the elliptical image regions here are detected *independently* in each viewpoint. As is evident, the pre-images of these affine

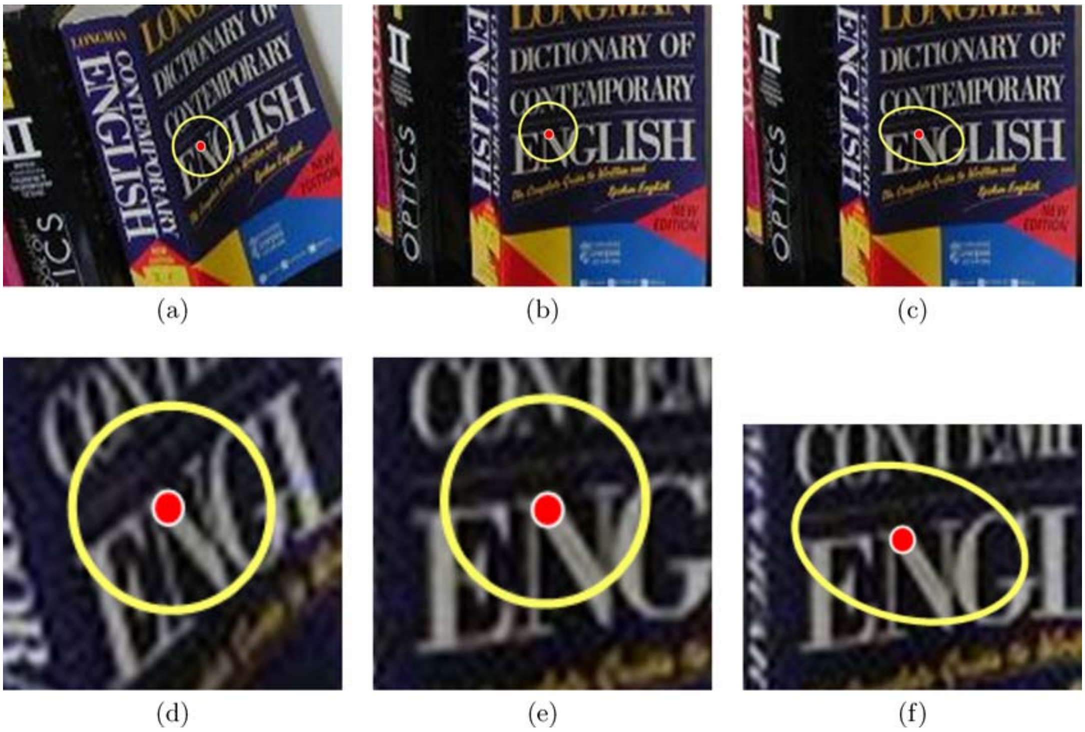


Figure 1. Class of transformations needed to cope with viewpoint changes. (a) First viewpoint; (b, c) second viewpoint. Fixed size circular patches (a, b) clearly do not suffice to deal with general viewpoint changes. What is needed is an anisotropic rescaling, i.e., an affinity (c). Bottom row shows close-up of the images of the top row.

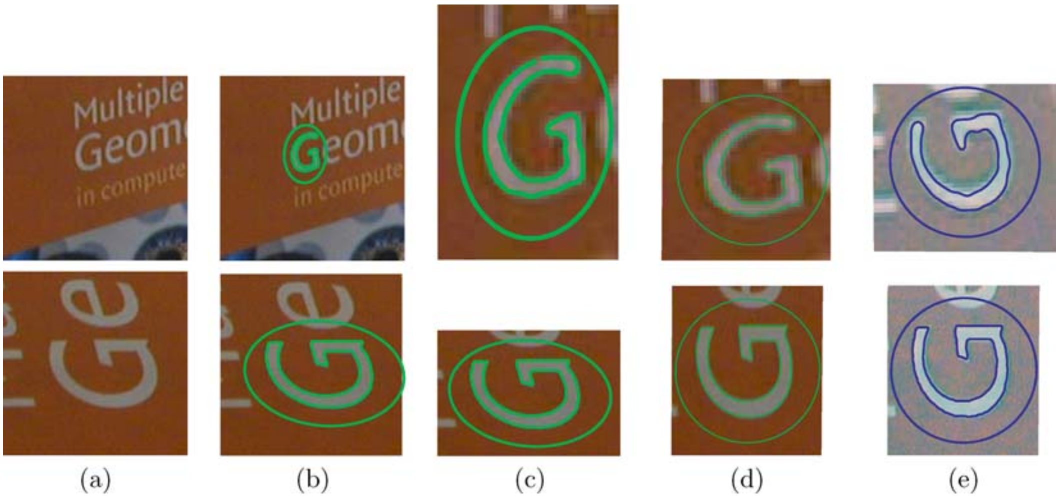


Figure 2. Affine covariant regions offer a solution to viewpoint and illumination changes. First row: one viewpoint; second row: other viewpoint. (a) Original images, (b) detected affine covariant regions, (c) close-up of the detected regions. (d) Geometric normalization to circles. The regions are the same up to rotation. (e) Photometric and geometric normalization. The slight residual difference in rotation is due to an estimation error.

covariant regions correspond to the same surface region. Given such an affine covariant region, it is then possible to normalize against the geometric and photometric deformations (shown in Fig. 2(d), (e)) and to obtain a viewpoint and illumination invariant description of the intensity pattern within the region.

In a typical matching application, the regions are used as follows. First, a set of covariant regions is detected in an image. Often a large number, perhaps hundreds or thousands, of possibly overlapping regions are obtained. A vector descriptor is then associated with each region, computed from the intensity pattern within the region. This descriptor is chosen to be invariant to viewpoint changes and, to some extent, illumination changes, and to discriminate between the regions. Correspondences may then be established with another image of the same scene, by first detecting and representing regions (independently) in the new image; and then matching the regions based on their descriptors. By design the regions commute with viewpoint change, so by design, corresponding regions in the two images will have similar (ideally identical) vector descriptors. The benefits are that correspondences can then be easily established and, since there are multiple regions, the method is robust to partial occlusions.

This paper gives a snapshot of the state of the art in affine covariant region detection. We will describe and compare six methods of detecting these regions on images. These detectors have been designed and implemented by a number of researchers and the comparison is carried out using binaries supplied by the authors. The detectors are: (i) the 'Harris-Affine' detector (Mikolajczyk and Schmid, 2002, 2004; Schaffalitzky and Zisserman, 2002); (ii) the 'Hessian-Affine' detector (Mikolajczyk and Schmid, 2002, 2004); (iii) the 'maximally stable extremal region' detector (or MSER, for short) (Matas et al., 2002, 2004); (iv) an edge-based region detector (Tuytelaars and Van Gool, 1999, 2004) (referred to as EBR); (v) an intensity extrema-based region detector (Tuytelaars and Van Gool, 2000, 2004) (referred to as IBR); and (vi) an entropy-based region detector (Kadir et al., 2004) (referred to as *salient regions*).

To limit the scope of the paper we have not included methods for detecting regions which are covariant only to similarity transformations (i.e., in particular scale), such as (Lowe, 1999, 2004; Mikolajczyk and Schmid, 2001; Mikolajczyk et al., 2003), or other methods of computing affine invariant descriptors, such as image lines connecting interest points (Matas et al., 2000; Tell and Carlson, 2000, 2002), or invariant vertical line segments (Goedeme et al., 2004). Also the detectors proposed by Lindeberg and Gårding (1997) and Baumberg (2000) have not been included, as they come very close to the Harris-Affine and Hessian-Affine detectors.

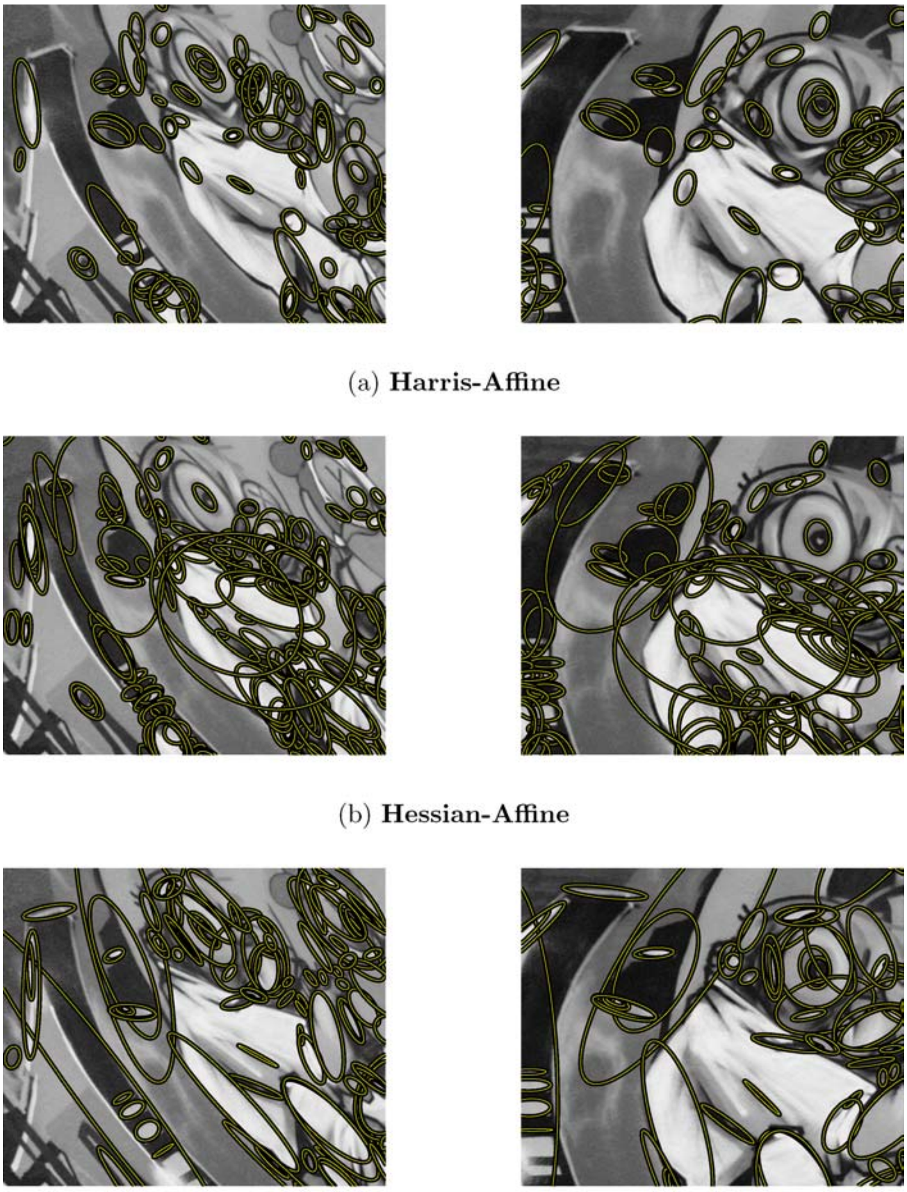
The six detectors are described in Section 2. They are compared on the data set shown in Fig. 9. This data set includes structured and textured scenes as well as different types of transformations: viewpoint changes, scale changes, illumination changes, blur and JPEG compression. It is described in more detail in Section 3. Two types of comparisons are carried out. First, in Section 10, the repeatability of the detector is measured: how well does the detector determine corresponding scene regions? This is measured by comparing the overlap between the ground truth and detected regions, in a manner similar to the evaluation test used in Mikolajczyk and Schmid (2002), but with special attention paid to the effect of the different scales (region sizes) of the various detectors' output. Here, we also measure the accuracy of the regions' shape, scale and localization. Second, the distinctiveness of the detected regions is assessed: how distinguishable are the regions detected? Following (Mikolajczyk and Schmid, 2003, 2005), we use the SIFT descriptor developed by Lowe (1999), which is an 128-dimensional vector, to describe the intensity pattern within the image regions. This descriptor has been demonstrated to be superior to others used in literature on a number of measures (Mikolajczyk and Schmid, 2003).

Our intention is that the images and tests described here will be a benchmark against which future affine covariant region detectors can be assessed. The images, Matlab code to carry out the performance tests, and binaries of the detectors are available from http://www.robots.ox.ac.uk/∼vgg/research/affine.

2. Affine Covariant Detectors

In this section we give a brief description of the six region detectors used in the comparison. Section 2.1 describes the related methods Harris-Affine and Hessian-Affine. Sections 2.2 and 2.3 describe methods for detecting edge-based regions and intensity extrema-based regions. Finally, Sections 2.4 and 2.5 describe MSER and salient regions.

For the purpose of the comparisons the output region of all detector types are represented by a common shape, which is an ellipse. Figures 3 and 4 show the ellipses for all detectors on one pair of images. In order not to overload the images, only some of the corresponding regions that were actually detected in both images have been shown. This selection is obtained by increasing the threshold.



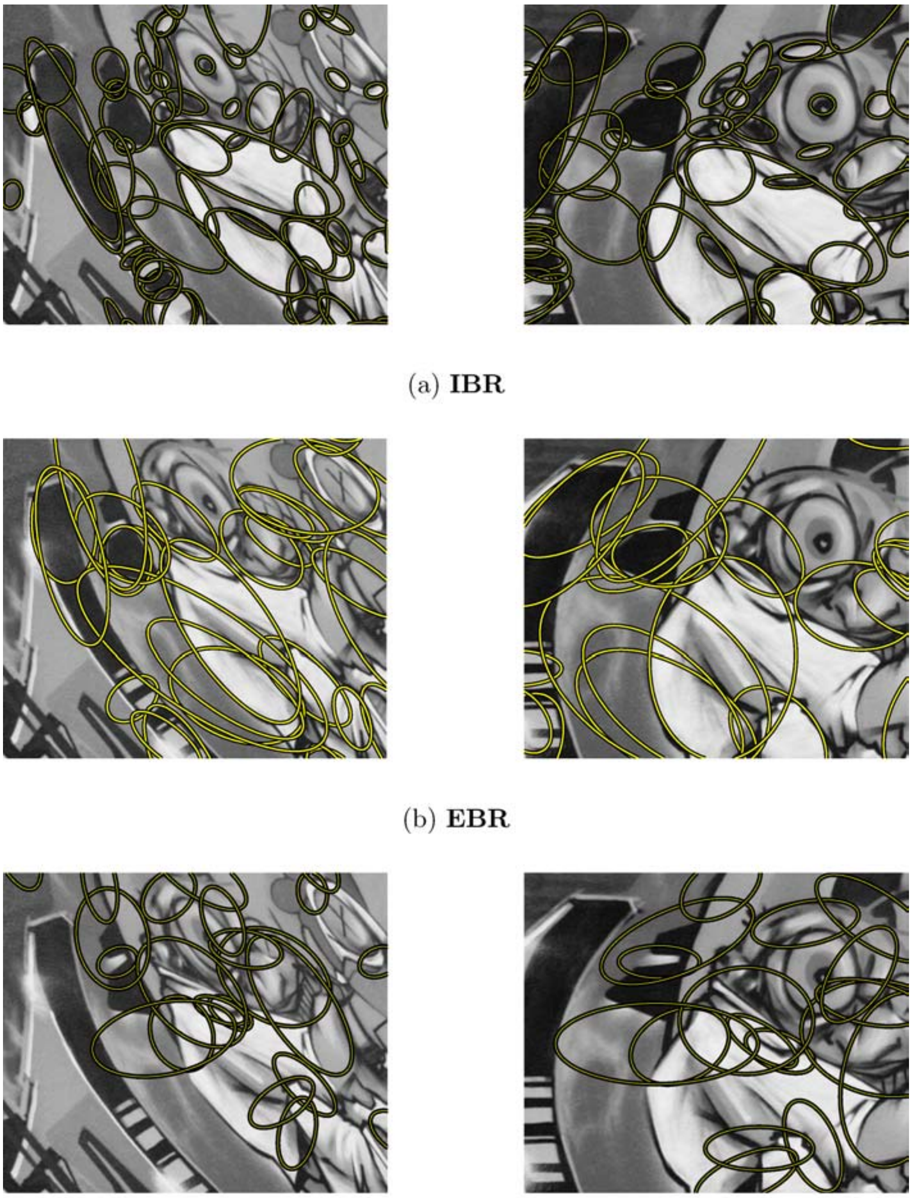
(a) **Harris-Affine**

(b) **Hessian-Affine**

(c) **MSER**

Figure 3. Regions generated by different detectors on corresponding sub-parts of the first and third graffiti images of Fig. 9(a). The ellipses show the original detection size.

In fact, for most of the detectors the output shape is an ellipse. However, for two of the detectors (edge-based regions and MSER) it is not, and information is lost by this representation, as ellipses can only be matched up to a rotational degree of freedom. Examples of the original regions detected by these two methods are given in Fig. 5. These are parallelogram-shaped regions for the edge-based region detector, and arbitrarily shaped regions for the MSER detector. In the following the representing ellipse is chosen to have the same first and second moments as the originally detected region, which is an affine covariant construction method.



(a) **IBR**

(b) **EBR**

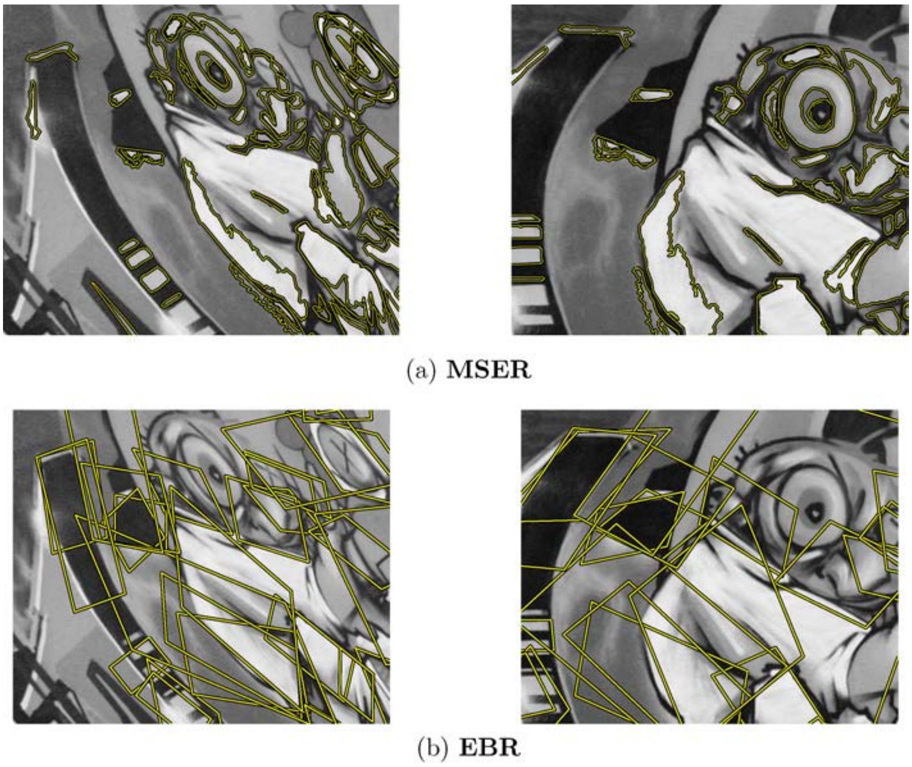
(c) **Salient regions**

Figure 4. Regions generated by different detectors continued.

2.1. Detectors Based on Affine Normalization— Harris-Affine & Hessian-Affine

We describe here two related methods which detect interest points in scale-space, and then determine an elliptical region for each point. Interest points are either detected with the Harris detector or with a detector based on the Hessian matrix. In both cases scale-selection is based on the Laplacian, and the shape of the elliptical region is determined with the second moment matrix of the intensity gradient (Baumberg, 2000; Lindeberg and Gårding, 1997).

The second moment matrix, also called the auto-correlation matrix, is often used for feature detection or for describing local image structures. Here it is used both in the Harris detector and the elliptical



(a) **MSER**

(b) **EBR**

Figure 5. Originally detected region shapes for the regions shown in Figs. 3(c) and 4(b).

shape estimation. This matrix describes the gradient distribution in a local neighbourhood of a point:

$$M = \mu(\mathbf{x}, \sigma_I, \sigma_D) = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix}$$
$$= \sigma_D^2\, g(\sigma_I) * \begin{bmatrix} I_x^2(\mathbf{x}, \sigma_D) & I_x I_y(\mathbf{x}, \sigma_D) \\ I_x I_y(\mathbf{x}, \sigma_D) & I_y^2(\mathbf{x}, \sigma_D) \end{bmatrix} \qquad (1)$$

The local image derivatives are computed with Gaussian kernels of scale σ_D (differentiation scale). The derivatives are then averaged in the neighbourhood of the point by smoothing with a Gaussian window of scale σ_I (integration scale). The eigenvalues of this matrix represent two principal signal changes in a neighbourhood of the point. This property enables the extraction of points, for which both curvatures are significant, that is the signal change is significant in orthogonal directions. 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.

A similar idea is explored in the detector based on the Hessian matrix:

$$H = H(\mathbf{x}, \sigma_D) = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$
$$= \begin{bmatrix} I_{xx}(\mathbf{x}, \sigma_D) & I_{xy}(\mathbf{x}, \sigma_D) \\ I_{xy}(\mathbf{x}, \sigma_D) & I_{yy}(\mathbf{x}, \sigma_D) \end{bmatrix} \qquad (2)$$

The second derivatives, which are used in this matrix give strong responses on blobs and ridges. The regions are similar to those detected by a Laplacian operator (trace) (Lindeberg, 1998; Lowe, 1999) but a function based on the determinant of the Hessian matrix penalizes very long structures for which the second derivative in one particular orientation is very small. A local maximum of the determinant indicates the presence of a blob structure.

To deal with scale changes a scale selection method (Lindeberg, 1998) is applied. The idea is to select the *characteristic* scale of a local structure, for which a given function attains an extremum over scales (see Fig. 6). The selected scale is characteristic in the quantitative sense, since it measures the scale

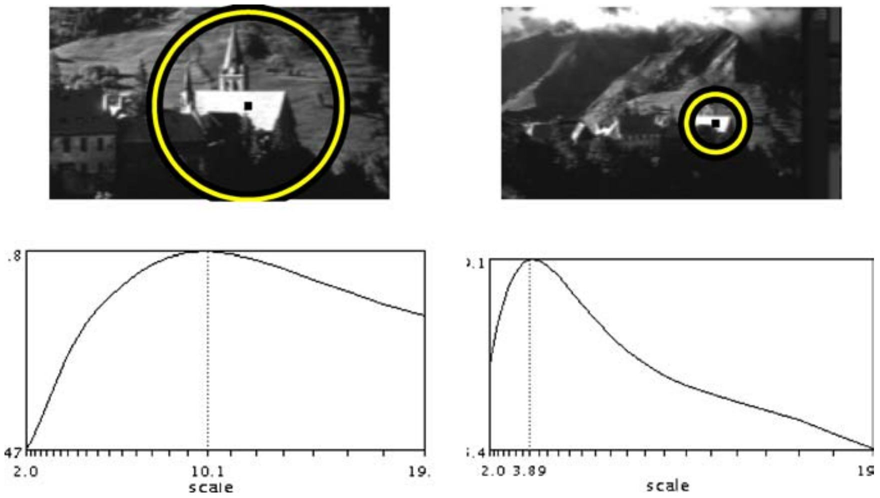


Figure 6. Example of characteristic scales. Top row shows images taken with different zoom. Bottom row shows the responses of the Laplacian over scales. The characteristic scales are 10.1 and 3.9 for the left and right image, respectively. The ratio of scales corresponds to the scale factor (2.5) between the two images. The radius of displayed regions in the top row is equal to 3 times the selected scales.

at which there is maximum similarity between the feature detection operator and the local image structures. The size of the region is therefore selected independently of image resolution for each point. The Laplacian operator is used for scale selection in both detectors since it gave the best results in the experimental comparison in Mikolajczyk and Schmid (2001).

Given the set of initial points extracted at their characteristic scales we can apply the iterative estimation of elliptical affine region (Lindeberg and Gårding, 1997). The eigenvalues of the second moment matrix are used to measure the affine shape of the point neighbourhood. To determine the affine shape, we find the transformation that projects the affine pattern to the one with equal eigenvalues. This transformation is given by the square root of the second moment matrix $M^{1/2}$. If the neighbourhood 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; Lindeberg and Gårding, 1997). The matrices M'_L and M'_R computed in the normalized frames are equal to a rotation matrix (see Fig. 7). Note that rotation preserves the eigenvalue ratio for an image patch, therefore, the affine deformation can be determined up to a rotation factor. This factor can be recovered by other methods, for example normalization based on the dominant gradient orientation (Lowe, 1999; Mikolajczyk and Schmid, 2002).

The estimation of affine shape can be applied to any initial point given that the determinant of the second moment matrix is larger than zero and the signal to noise ratio is insignificant for this point. We can therefore use this technique to estimate the shape of initial regions provided by the Harris and Hessian based detector.

The outline of the iterative region estimation:

1. Detect initial region with Harris or Hessian detector and select the scale.
2. Estimate the shape with the second moment matrix
3. Normalize the affine region to the circular one
4. Go to step 2 if the eigenvalues of the second moment matrix for new point are not equal.

Examples of Harris-Affine and Hessian-Affine regions are displayed on Fig. 3(a) and (b).

2.2. An Edge-Based Region Detector

We describe here a method to detect affine covariant regions in an image by exploiting the edges present in the image. The rationale behind this is that edges are typically rather stable features, that can be detected over a range of viewpoints, scales and/or illumination changes. Moreover, by exploiting the edge geometry, the dimensionality of the problem can be significantly reduced. Indeed, as will be shown next, the 6D search problem over all possible affinities (or 4D, once the

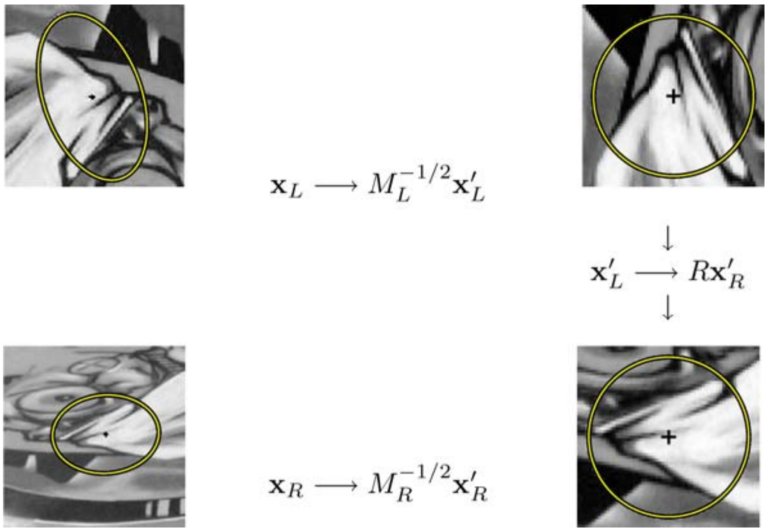

$\mathbf{x}_L \longrightarrow M_L^{-1/2}\mathbf{x}'_L$
$\downarrow$
$\mathbf{x}'_L \longrightarrow R\mathbf{x}'_R$
$\downarrow$
$\mathbf{x}_R \longrightarrow M_R^{-1/2}\mathbf{x}'_R$

Figure 7. Diagram illustrating the affine normalization using the second moment matrices. Image coordinates are transformed with matrices $M_L^{-1/2}$ and $M_R^{-1/2}$.

center point is fixed) can further be reduced to a one-dimensional problem by exploiting the nearby edges geometry. In practice, we start from a Harris corner point **p** (Harris and Stephens, 1988) and a nearby edge, extracted with the Canny edge detector (Canny, 1986). To increase the robustness to scale changes, these basic features are extracted at multiple scales. Two points $\mathbf{p_1}$ and $\mathbf{p_2}$ move away from the corner in both directions along the edge, as shown in Fig. 8(a). Their relative speed is coupled through the equality of relative affine invariant parameters l_1 and l_2:

$$l_i = \int \text{abs}\left(\left|\mathbf{p_i}^{(1)}(s_i) \quad \mathbf{p} - \mathbf{p_i}(s_i)\right|\right) ds_i \tag{3}$$

with s_i an arbitrary curve parameter (in both directions), $\mathbf{p_i}^{(1)}(s_i)$ the first derivative of $\mathbf{p_i}(s_i)$ with respect to s_i, abs() the absolute value and $|\ldots|$ the determinant. This condition prescribes that the areas between the joint $\langle \mathbf{p}, \mathbf{p_1} \rangle$ and the edge and between the joint $\langle \mathbf{p}, \mathbf{p_2} \rangle$ and the edge remain identical. This is an affine invariant criterion indeed. From now on, we simply use l when referring to $l_1 = l_2$.

For each value l, the two points $\mathbf{p_1}(l)$ and $\mathbf{p_2}(l)$ together with the corner **p** define a parallelogram $\Omega(l)$: the parallelogram spanned by the vectors $\mathbf{p_1}(l) - \mathbf{p}$ and $\mathbf{p_2}(l) - \mathbf{p}$. This yields a one dimensional family of parallelogram-shaped regions as a function of l. From this 1D family we select one (or a few) parallelogram for which the following photometric quantities of the texture go through an extremum.

$$\text{Inv}_1 = abs\left(\frac{|\mathbf{p_1}-\mathbf{p_g} \quad \mathbf{p_2}-\mathbf{p_g}|}{|\mathbf{p}-\mathbf{p_1} \quad \mathbf{p}-\mathbf{p_2}|}\right)\frac{M^1_{00}}{\sqrt{M^2_{00}M^0_{00} - (M^1_{00})2}}$$

$$\text{Inv}_2 = abs\left(\frac{|\mathbf{p}-\mathbf{p_g} \quad \mathbf{q}-\mathbf{p_g}|}{|\mathbf{p}-\mathbf{p_1} \quad \mathbf{p}-\mathbf{p_2}|}\right)\frac{M^1_{00}}{\sqrt{M^2_{00}M^0_{00} - (M^1_{00})2}}$$

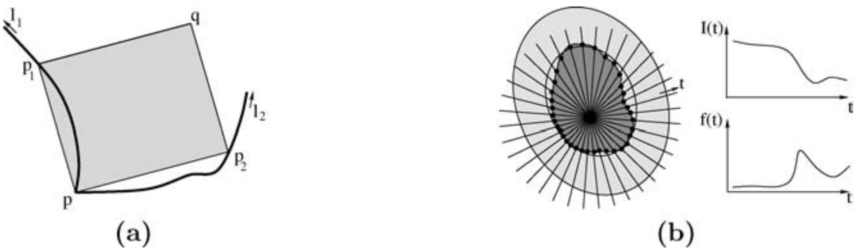


Figure 8. Construction methods for EBR and IBR. (a) The edge-based region detector starts from a corner point **p** and exploits nearby edge information; (b) The intensity extrema-based region detector starts from an intensity extremum and studies the intensity pattern along rays emanating from this point.

with

$$M^n_{pq} = \int_\Omega I^n(x, y) x^p y^q \, dx dy \qquad (4)$$

$$\mathbf{p_g} = \left(\frac{M^1_{10}}{M^1_{00}}, \frac{M^1_{01}}{M^1_{00}} \right)$$

with M^n_{pq} the nth *order*, $(p+q)$th *degree* moment computed over the region $\Omega(l)$, $\mathbf{p_g}$ the center of gravity of the region, weighted with intensity $I(x, y)$, and $\mathbf{q}$ the corner of the parallelogram opposite to the corner point $\mathbf{p}$ (see Fig. 8(a)). The second factor in these formula has been added to ensure invariance under an intensity offset.

In the case of straight edges, the method described above cannot be applied, since $l = 0$ along the entire edge. Since intersections of two straight edges occur quite often, we cannot simply neglect this case. To circumvent this problem, the two photometric quantities given in Eq. (4) are combined and locations where *both* functions reach a minimum value are taken to fix the parameters s_1 and s_2 along the straight edges. Moreover, instead of relying on the correct detection of the Harris corner point, we can simply use the straight lines intersection point instead. A more detailed explanation of this method can be found in Tuytelaars and Van Gool (1999, 2004). Examples of detected regions are displayed in Fig. 5(b).

For easy comparison in the context of this paper, the parallelograms representing the invariant regions are replaced by the enclosed ellipses, as shown in Fig. 4(b). However, in this way the orientation-information is lost, so it should be avoided in a practical application, as discussed in the beginning of Section 2.

2.3. Intensity Extrema-Based Region Detector

Here we describe a method to detect affine covariant regions that starts from intensity extrema (detected at multiple scales), and explores the image around them in a radial way, delineating regions of arbitrary shape, which are then replaced by ellipses.

More precisely, given a local extremum in intensity, the intensity function along rays emanating from the extremum is studied, as shown in Fig. 8(b). The following function is evaluated along each ray:

$$f_I(t) = \frac{abs(I(t) - I_0)}{\max\left(\frac{\int_0^t abs(I(t) - I_0) dt}{t}, d \right)}$$

with t an arbitrary parameter along the ray, $I(t)$ the intensity at position t, I_0 the intensity value at the extremum and d a small number which has been added to prevent a division by zero. The point for which this function reaches an extremum is invariant under affine geometric and linear photometric transformations (given the ray). Typically, a maximum is reached at positions where the intensity suddenly increases or decreases. The function $f_I(t)$ is in itself already invariant. Nevertheless, we select the points where this function reaches an extremum to make a robust selection. Next, all points corresponding to maxima of $f_I(t)$ along rays originating from the same local extremum are linked to enclose an affine covariant region (see Fig. 8(b)). This often irregularly-shaped region is replaced by an ellipse having the same shape moments up to the second order. This ellipse-fitting is again an affine covariant construction. Examples of detected regions are displayed in Fig. 4(a). More details about this method can be found in Tuytelaars and Van Gool (2000, 2004).

2.4. Maximally Stable Extremal Region Detector

A Maximally Stable Extremal Region (MSER) is a connected component of an appropriately thresholded image. The word ‘extremal’ refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The ‘maximally stable’ in MSER describes the property optimized in the threshold selection process.

The set of extremal regions $\mathcal{E}$, i.e., the set of all connected components obtained by thresholding, has a number of desirable properties. Firstly, a monotonic change of image intensities leaves $\mathcal{E}$ unchanged, since it depends only on the ordering of pixel intensities which is preserved under monotonic transformation. This ensures that common photometric changes modelled locally as linear or affine leave $\mathcal{E}$ unaffected, even if the camera is non-linear (gamma-corrected). Secondly, continuous geometric transformations preserve topology–pixels from a single connected component are transformed to a single connected component. Thus after a geometric change locally approximated by an affine transform, homography or even continuous non-linear warping, a matching extremal region will be in the transformed set $\mathcal{E}'$. Finally, there are no more extremal regions than there are pixels in the image. So a set of regions was defined that is preserved under a broad class of geometric and photometric changes and yet has the same cardinality as e.g. the set of fixed-sized square windows commonly used in narrow-baseline matching.

Implementation Details. The enumeration of the set of extremal regions $\mathcal{E}$ is very efficient, almost linear in the number of image pixels. The enumeration proceeds as follows. First, pixels are sorted by intensity. After sorting, pixels are marked in the image (either in decreasing or increasing order) and the list of growing and merging connected components and their areas is maintained using the union-find algorithm (Sedgewick, 1988). During the enumeration process, the area of each connected component as a function of intensity is stored. Among the extremal regions, the 'maximally stable' ones are those corresponding to thresholds were the relative area change as a function of relative change of threshold is at a local minimum. In other words, the MSER are the parts of the image where local binarization is stable over a large range of thresholds. The definition of MSER stability based on relative area change is only affine invariant (both photometrically and geometrically). Consequently, the process of MSER detection is affine covariant.

Detection of MSER is related to *thresholding*, since every extremal region is a connected component of a thresholded image. However, no global or 'optimal' threshold is sought, all thresholds are tested and the stability of the connected components evaluated. The output of the MSER detector is not a binarized image. For some parts of the image, multiple stable thresholds exist and a system of nested subsets is output in this case.

Finally we remark that the different sets of extremal regions can be defined just by changing the ordering function. The MSER described in this section and used in the experiments should be more precisely called intensity induced MSERs.

2.5. Salient Region Detector

This detector is based on the pdf of intensity values computed over an elliptical region. Detection proceeds in two steps: first, at each pixel the entropy of the pdf is evaluated over the three parameter family of ellipses centred on that pixel. The set of entropy extrema over scale and the corresponding ellipse parameters are recorded. These are candidate salient regions. Second, the candidate salient regions over the entire image are ranked using the magnitude of the derivative of the pdf with respect to scale. The top P ranked regions are retained.

In more detail, the elliptical region $\mathcal{E}$ centred on a pixel $\mathbf{x}$ is parameterized by its scale s (which specifies the major axis), its orientation θ (of the major axis), and the ratio of major to minor axes λ. The pdf of intensities $p(I)$ is computed over $\mathcal{E}$. The entropy $\mathcal{H}$ is then given by

$$\mathcal{H} = -\sum_I p(I) \log p(I)$$

The set of extrema over scale in $\mathcal{H}$ is computed for the parameters s, θ, λ for each pixel of the image. For each extrema the derivative of the pdf $p(I; s, \theta, \lambda)$ with s is computed as

$$\mathcal{W} = \frac{s^2}{2s-1} \sum_I \left| \frac{\partial p(I; s, \theta, \lambda)}{\partial s} \right|,$$

and the *saliency* $\mathcal{Y}$ of the elliptical region is computed as $\mathcal{Y} = \mathcal{HW}$. The regions are ranked by their saliency $\mathcal{Y}$. Examples of detected regions are displayed in Fig. 4(c). More details about this method can be found in Kadir et al. (2004).

3. The Image Data Set

Figure 9 shows examples from the image sets used to evaluate the detectors. Five different changes in imaging conditions are evaluated: viewpoint changes (a) & (b); scale changes (c) & (d); image blur (e) & (f); JPEG compression (g); and illumination (h). In the cases of viewpoint change, scale change and blur, the same change in imaging conditions is applied to two different scene types. This means that the effect of changing the image conditions can be separated from the effect of changing the scene type. One scene type contains homogeneous regions with distinctive edge boundaries (e.g. graffiti, buildings), and the other contains repeated textures of different forms. These will be referred to as *structured* versus *textured* scenes respectively.

In the viewpoint change test the camera varies from a fronto-parallel view to one with significant foreshortening at approximately 60 degrees to the camera. The scale change and blur sequences are acquired by varying the camera zoom and focus respectively.

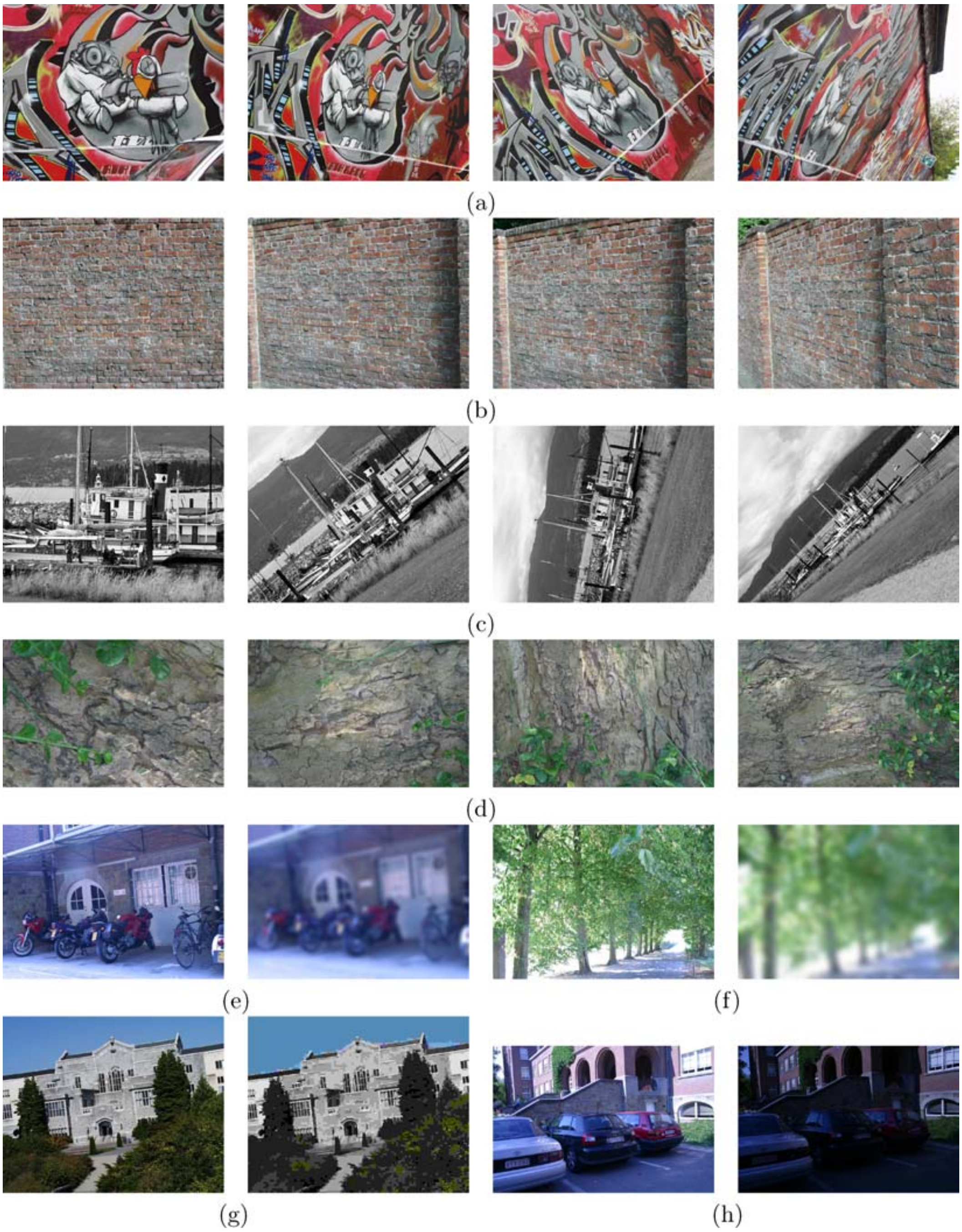


Figure 9. Data set. (a), (b) Viewpoint change, (c), (d) Zoom+rotation, (e), (f) Image blur, (g) JPEG compression, (h) Light change. In the case of viewpoint change, scale change and blur, the same change in imaging conditions is applied to two different scene types: structured and textured scenes. In the experimental comparisons, the left most image of each set is used as the reference image.

The scale changes by about a factor of four. The light changes are introduced by varying the camera aperture. The JPEG sequence is generated using a standard xv image browser with the image quality parameter varying from 40 to 2%. Each of the test sequences contains 6 images with a gradual geometric or photometric transformation. All images are of medium resolution (approximately 800 × 640 pixels).

The images are either of planar scenes or the camera position is fixed during acquisition, so that in all cases the images are related by homographies (plane projective transformations). This means that the mapping relating images is known (or can be computed), and this mapping is used to determine ground truth matches for the affine covariant detectors.

The homographies between the reference (left most) image and the other images in a particular dataset are computed in two steps. First, a small number of point correspondences are selected manually between the reference and other image. These correspondences are used to compute an approximate homography between the images, and the other image is warped by this homography so that it is roughly aligned with the reference image. Second, a standard small-baseline robust homography estimation algorithm is used to compute an accurate residual homography between the reference and warped image (using hundreds of automatically detected and matched interest points) (Hartley and Zisserman, 2004). The composition of these two homographies (approximate and residual) gives an accurate homography between the reference and other image. The root-mean-square error is less than 1 pixel for every image pair.

Of course, the homography could be computed directly and automatically using correspondences of the affine covariant regions detected by any of the methods of Section 2. The reason for adopting this two step approach is to have an estimation method independent of all the detectors that are being evaluated.

All the images as well as the computed homographies are available on the website.

3.1. Discussion

Before we compare the performance of the different detectors in more detail in the next section, a few more general observations can already be made, simply by examining the output of the different detectors for the images shown in Figs. 3 and 4. For all our experiments (unless explicitly mentioned), the same set of parameters are used for each detector. These parameters are the default parameters given by the authors.

First of all, note that the ellipses in the left and right images of Figs. 3 and 4 do indeed cover more or less the same scene regions. This is the key requirement for covariant operators, and seems to be fulfilled for at least a subset of the detected regions for all detectors. Some other key observations are summarized below.

Complexity and Required Computation Time. The computational complexity of the algorithm finding initial points in the *Harris-Affine* and *Hessian-Affine* detectors is $\mathcal{O}(n)$, where n is the number of pixels. The complexity of the automatic scale selection and shape adaptation algorithm is $\mathcal{O}((m + k)p)$, where p is the number of initial points, m is a number of investigated scales in the automatic scale selection and k is a number of iterations in the shape adaptation algorithm.

For the *intensity extrema-based region detector*, the algorithm finding intensity extrema is $\mathcal{O}(n)$, where n is again the number of pixels. The complexity of constructing the actual region around the intensity extrema is $\mathcal{O}(p)$, where p is the number of intensity extrema.

For the *edge-based region detector*, the algorithm finding initial corner points and the algorithm finding edges in the image are both $\mathcal{O}(n)$, where n is again the number of pixels. The complexity of constructing the actual region starting from the corners and edges is $\mathcal{O}(pd)$, where p is the number of corners and d is the average number of edges nearby a corner.

For the *salient region detector*, the complexity of the first step of the algorithm is $\mathcal{O}(nl)$, where l is the number of ellipses investigated at each pixel (the three discretized parameters of the ellipse shape). The complexity of the second step is $\mathcal{O}(e)$, where e is the number of extrema detected in the first step.

For the *MSER detector*, the computational complexity of the sorting step is $\mathcal{O}(n)$ if the range of image values is small, e.g. the typical $\{0, \ldots, 255\}$, since the sort can be implemented as BINSORT. The complexity of the union-find algorithm is $\mathcal{O}(n \log\log n)$, i.e., fast.

Computation times vary widely, as can be seen in Table 1. The computation times mentioned in this table have all been measured on a Pentium 4 2 GHz Linux PC, for the leftmost image shown in Fig. 9(a), which is 800 × 640 pixels. Even though the timings are for not heavily optimized code and may change depending on the implementation as well as on the image content, we believe the table gives a reasonable indication of typical computation times.

Region Density. The various detectors generate very different numbers of regions, c.f. Table 1. The number of regions also strongly depends on the scene type, e.g. for the MSER detector there are about 2600 regions for the textured blur scene (Fig. 9(f)) and only 230 for the light change scene (Fig. 9(h)). Similar behaviour can be observed for other detectors.

The variation in numbers between detector type is to be expected since the detectors respond to different

Table 1. Computation times for the different detectors for the leftmost image of Fig. 9(a) (size 800 × 640).

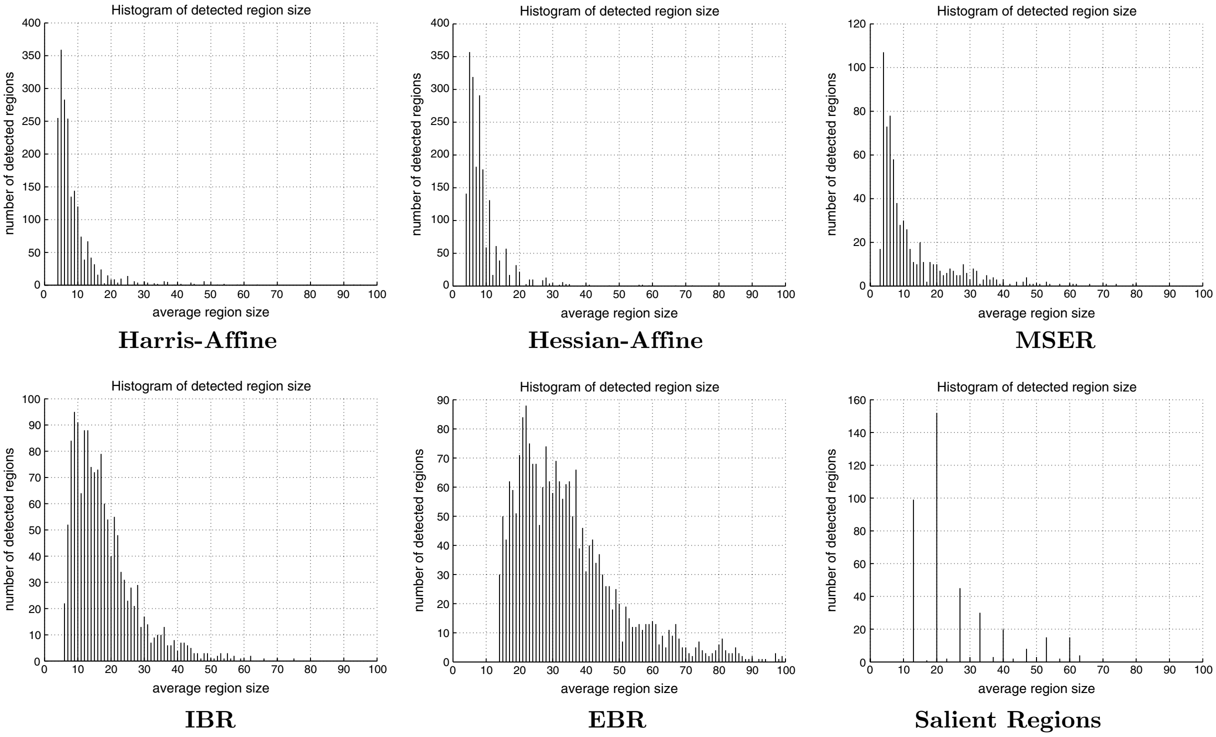
Detector	Run time (min:sec)	Number of regions
Harris-Affine	0:01.43	1791
Hessian-Affine	0:02.73	1649
MSER	0:00.66	533
IBR	0:10.82	679
EBR	2:44.59	1265
Salient Regions	33:33.89	513

features and the images contain different numbers for a given feature type. For example, the edge-based region detector requires curves, and if none of sufficient length occur in a particular image, then no regions of this type can be detected.

However, this variety is also a virtue: the different detectors are *complementary*. Some respond well to structured scenes (e.g. MSER and the edge-based regions), others to more textured scenes (e.g. Harris-Affine and Hessian-Affine). We will return to this point in Section 4.2.

Region Size. Also the size of the detected regions significantly varies depending on the detector. Typically, Harris-Affine, Hessian-Affine and MSER detect many very small regions, whereas the other detectors only yield larger ones. This can also be seen in the examples shown in Figs. 3 and 4. Figure 10 shows histograms of region size for the different region detectors. The size of the regions is measured as the geometric average of the half-length of both axes of the ellipses, which corresponds to the radius of a circular region with the same area. Larger regions typically have a higher discriminative power, as they contain more information, which makes them easier to match, at the cost of a higher risk of being occluded or not covering a planar part of the scene. Also, as will be shown in the next section (cf. Fig. 11), large regions automatically have better chances of overlapping other regions.

Distinguished Regions Versus Measurement Regions. As a final note, we would like to draw the attention of the reader to the fact that given a detected affine covariant region, it is possible to associate with it any number of new affine regions that are obtained



Harris-Affine Hessian-Affine MSER

IBR EBR Salient Regions

Figure 10. Histograms of region size for the different detectors for the reference image of Fig. 9(a). Note that the *y* axes do not have the same scalings in all cases.

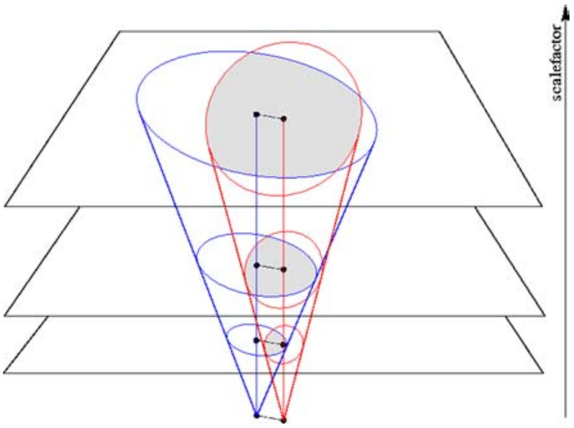


Figure 11. Rescaling regions has an effect on their overlap.

by affine covariant constructions, such as scaling, taking the convex hull or fitting an ellipse based on second order moments. In this respect, one should make a distinction between a *distinguished region* and a *measurement region*, as first pointed out in Matas et al. (2002), where the former refers to the set of pixels that have effectively contributed to the affine detector response while the latter can be any region obtained by an affine covariant construction. Here, we focus on the original distinguished regions (except for the ellipse fitting for edge-based and MSER regions, to obtain the same shape for all detectors), as they determine the intrinsic quality of a detector. In a practical matching setup however, it may be advantageous to use a different measurement region (see also Section 5 and the discussion on scale in next section).

4. Overlap Comparison Using Homographies

The objective of this experiment is to measure the repeatability and accuracy of the detectors: to what extent do the detected regions overlap exactly the same scene area (i.e., are the pre-images identical)? How often are regions detected in one image without the corresponding region being detected in another? Quantitative results are obtained by making these questions precise (see below). The ground truth in all cases is provided by mapping the regions detected on the images in a set to a *reference image* using homographies. The basic measure of accuracy and repeatability we use is the relative amount of overlap between the detected region in the reference image and the region detected in the other image, projected onto the reference image using the homography relating the images. This gives a good indication of the chance that the region can be matched correctly. In the tests the reference image is always the image of highest quality and is shown as the leftmost image of each set in Fig. 9.

Two important parameters characterize the performance of a region detector:

1. the *repeatability*, i.e., the average number of corresponding regions detected in images under different geometric and photometric transformations, both in absolute and relative terms (i.e., percentage-wise), and
2. the *accuracy* of localization and region estimation.

However, before describing the overlap test in more detail, it is necessary to discuss the effect of region size and region density, since these affect the outcome of the overlap comparison.

A Note on the Effect of Region Size. Larger regions automatically have a better chance of yielding good overlap scores. Simply rescaling the regions, i.e., using a different measurement region (e.g. doubling the size of all regions) suffices to boost the overlap performance of a region detector. This can be understood as follows. Suppose the distinguished region is an ellipse, and the measurement region is also an ellipse centered on the distinguished region but with an arbitrary scaling s. Then from a geometrical point of view, varying the scaling defines a cone out of the image plane (with elliptical cross-section), and with s a distance on the cone axis. In the reference image there are two such cones–one from the distinguished region in that image, and the other from the mapped distinguished region from the other image, as illustrated in Fig. 11. Clearly as the scaling goes to zero there is no intersection of the cones, and as the scaling goes to infinity the relative amount of overlap, defined as the ratio of the intersection to the union of the ellipses approaches unity.

To measure the intrinsic quality of a region detector, we need to define an overlap criterion that is insensitive to such rescaling. Focusing on the original distinguished regions would unproportionally favour detectors with large distinguished regions. Instead, the solution adopted here is to apply a scaling s that normalizes the reference region to a fixed region size prior to computing the overlap measure. It should be noted though that this is only for reason of comparison of different detectors. It may result in increased or decreased repeatability scores compared to what one might get in a typical matching experiment, where such normalization typically is not performed (and is not desirable either).

A Note on the Effect of Region Density. Also the region density, i.e., the number of detected regions per fixed amount of pixel area, may have an effect on the repeatability score of a detector. Indeed, if only a few regions are detected, the thresholds can be set very sharply, resulting in very stable regions, which typically perform better than average. At the other extreme, if the number of regions becomes really huge, the image might get so cluttered with regions that some of them may be matched by accident rather than by design. In the limit, one would get an (affine) scale space approach rather than an affine covariant region detector.

One way out would be to tune the parameters of the detectors such that they all output a similar number of regions. However, this is difficult to achieve in practice, since the number of detected regions also depends on the scene type. Moreover, it is not straightforward for all detectors to come up with a single parameter that can be varied to obtain the desired number of regions in a meaningful way, i.e., representing some kind of 'quality measure' for the regions. So we use the default parameters supplied by the authors. To give an idea of the number of regions, both absolute and relative repeatability scores are given. In addition, for several detectors, the repeatability is computed versus the number of detected regions, which is reported in Section 4.3.

4.1. Repeatability Measure

Two regions are deemed to correspond if the *overlap error*, defined as the error in the image area covered by the regions, is sufficiently small:

$$1 - \frac{R_{\mu_a} \cap R_{(H^T \mu_b H)}}{(R_{\mu_a} \cup R_{H^T \mu_b H})} < \epsilon_O$$

where R_μ represents the elliptic region defined by $x^T \mu x = 1$. H is the homography relating the two images. The union of the regions is $R_{\mu_a} \cup R_{(H^T \mu_b H)}$, and $R_{\mu_a} \cap R_{(H^T \mu_b H)}$ is their intersection. The area of the union and the intersection of the regions are computed numerically.

The *repeatability score* for a given pair of images is computed as the ratio between the number of region-to-region correspondences and the smaller of the number of regions in the pair of images. We take into account only the regions located in the part of the scene present in both images.

To compensate for the effect of regions of different sizes, as mentioned in the previous section, we first rescale the regions as follows. Based on the region detected in the reference image, we determine the scale factor that transforms it into a region of normalized size (corresponding to a radius 30, in our experiments). Then, we apply this scale factor to both the region in the reference image and the region detected in the other image which has been mapped onto the reference image, before computing the actual overlap error as described above. The precise procedure is given in the Matlab code on http://www.robots.ox.ac.uk/∼vgg/research/affine.

Examples of the overlap errors are displayed in Fig. 12. Note that an overlap error of 20% is very small as it corresponds to only 10% difference between the regions' radius. Regions with 50% overlap error can still be matched successfully with a robust descriptor.

4.2. Repeatability Under Various Transformations

In a first set of experiments, we fix the overlap error threshold to 40% and the normalized region size to a radius of 30 pixels, and check the repeatability of the different region detectors for gradually increasing transformations, according to the image sets shown in Fig. 9. In other words, we measure how the number of correspondences depends on the transformation between the reference and other images in the set. Both the relative and actual number of corresponding regions is recorded. In general we would like a detector to have a high repeatability score and a large number of correspondences. This test allows to measure the robustness of the detectors to changes in viewpoint, scale, illumination, etc.

The results of these tests are shown in Figs. 13–20 (a) and (b). Figures 13–20(c) and (d) show matching results, which are discussed in Section 4.4. A detailed discussion is given below, but we first make some general comments. The ideal plot for repeatability would be a horizontal line at 100%. As can be seen in all cases, neither a horizontal line nor 100% are achieved. Indeed the performance generally decreases with the severity of the transformation, and the best performance achieved is 95% for JPEG compression (Fig. 19). The reasons for this lack of 100% performance are

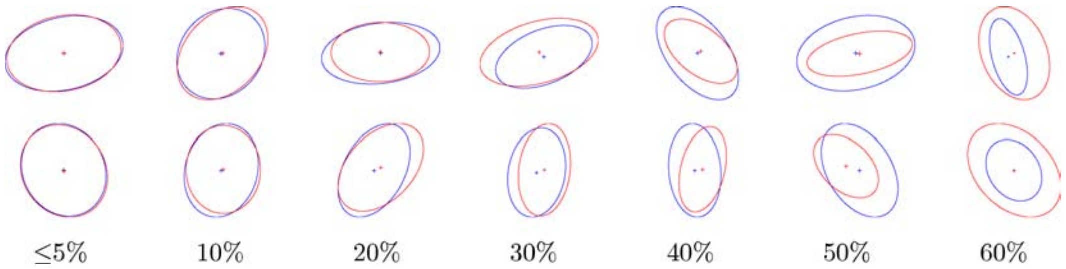


Figure 12. Overlap error ϵ_O. Examples of ellipses projected on the corresponding ellipse with the ground truth transformation. (bottom) Overlap error for above displayed ellipses. Note that the overlap error comes from different size, orientation and position of the ellipses.

sometimes specific to detectors and scene types (discussed below), and sometimes general—the transformation is outside the range for which the detector is designed, e.g. discretization errors, noise, non-linear illumination changes, projective deformations etc. Also the limited 'range' of the regions shape (size, skewness, ...) can partially explain this effect. For instance, in case of a zoomed out test image, only the large regions in the reference image will survive the transformation, as the small regions will have become too small for accurate detection. The same holds for other types of transformations: very elongated regions in the reference image may become undetectable if the inferred affine transformation stretches them even further but, on the other hand, allow for very large viewpoint changes if the inferred affine transformation makes them rounder.

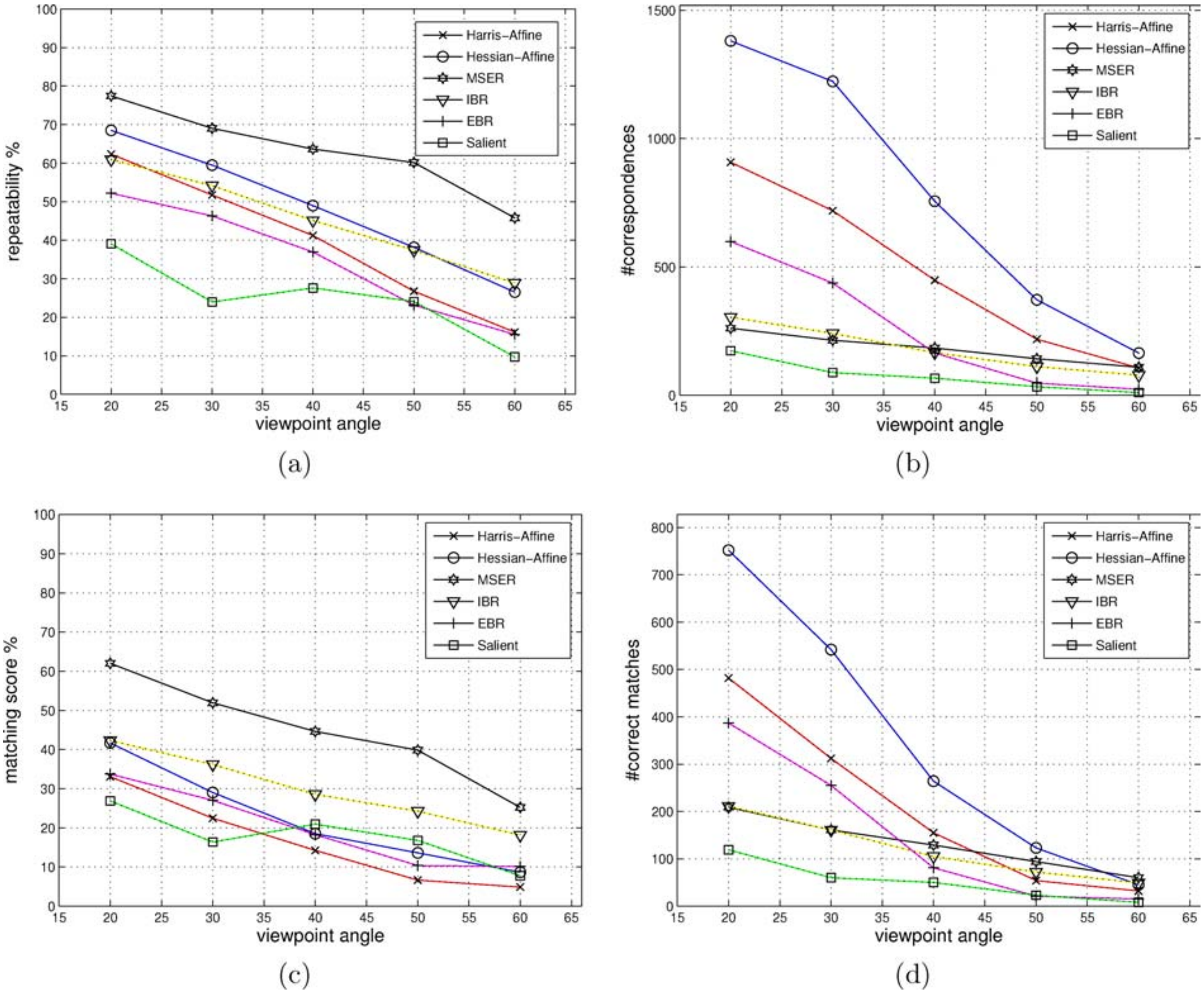
The left side of each figure typically represents small transformations. The repeatability score obtained in this range indicates how well a given detector performs for the given scene type and to what extent the detector is affected by a small transformation of this scene. The invariance of the detector under the studied transformation, on the other hand, is reflected in the slope of the curves, i.e., how much does a given curve degrade with increasing transformations.

The absolute number of correspondences typically drops faster than the relative number. This can be understood by the fact that in most cases larger transformations result in lower quality images and/or smaller commonly visible parts between the reference image and the other image, and hence a smaller number of regions are detected.

Viewpoint Change. The effect of changing viewpoint for the structured graffiti scene from Fig. 9(a) are displayed in Fig. 13. Figure 13(a) shows the repeatability score and Fig. 13(b) the absolute number of correspondences. The results for images containing repeated texture motifs (Fig. 9(b)) are displayed in Fig. 14. The best results are obtained with the MSER detector for both scene types. This is due to the high detection accuracy especially on the homogeneous regions with distinctive boundaries. The repeatability score for a viewpoint change of 20 degrees varies between 40% and 78% and decreases for large viewpoint angles to $10\% - 46\%$. The largest number of corresponding regions is given by Hessian-Affine (1300) detector followed by Harris-Affine (900) detector for the structured scene, and given by Harris-Affine (1200), MSER (1200) and EBR (1300) detectors for the textured scene. These numbers decrease to less than 200/400 for the structured/textured scene for large viewpoint angle.

Scale Change. Figure 15 shows the results for the structured scene from Fig. 9(c), while Fig. 16 shows the results for the textured scene from Fig. 9(d). The main image transformation is a scale change and in-plane rotation. The Hessian-Affine detector performs best, followed by MSER and Harris-Affine detectors. This confirms the high performance of the automatic scale selection applied in both Hessian-Affine and Harris-Affine detectors. These plots clearly show the sensitivity of the detectors to the scene type. For the textured scene, the edge-based region detector gives very low repeatability scores (below 20%), whereas for the structured scene, its results are similar to the other detectors, with score going from 60% down to 28%. The unstable repeatability score of the salient region detector for the textured scene is due to the small number of detected regions in this type of images.

Blur. Figures 17 and 18 show the results for the structured scene from Fig. 9(e) and the textured one from Fig. 9(f), both undergoing increasing amounts of



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

Figure 13. Viewpoint change for the structured scene (Graffiti sequence Fig. 9(a)). (a) Repeatability score for viewpoint change (default settings-overlap 40%, normalized size = 30 pixels). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

image blur. The results are better than for viewpoint and scale changes, especially for the structured scene. All detectors have nearly horizontal repeatability curves, showing a high level of invariance to image blur, except for the MSER detector, which is clearly more sensitive to this type of transformation. This is because the region boundaries become smooth, and the segmentation process is less accurate. The number of corresponding regions detected on structured scene is much lower than for the textured scene and it changes by a different factor for different detectors. This clearly shows that the detectors respond to different features. The repeatability for the EBR detector is very low for the textured scene. This can be explained by the lack of stable edges, on which the region extraction is based.

JPEG Artifacts. Figure 19 shows the score for the JPEG compression sequence from Fig. 9(g). For this type of structured scene (buildings), with large homogeneous areas and distinctive corners, Hessian-Affine and Harris-Affine are clearly best suited. The degradation under increasing compression artefacts is similar for all detectors.

Light Change. Figure 20 shows the results for light changes for the images on Fig. 9(h). All curves are nearly horizontal, showing good robustness to

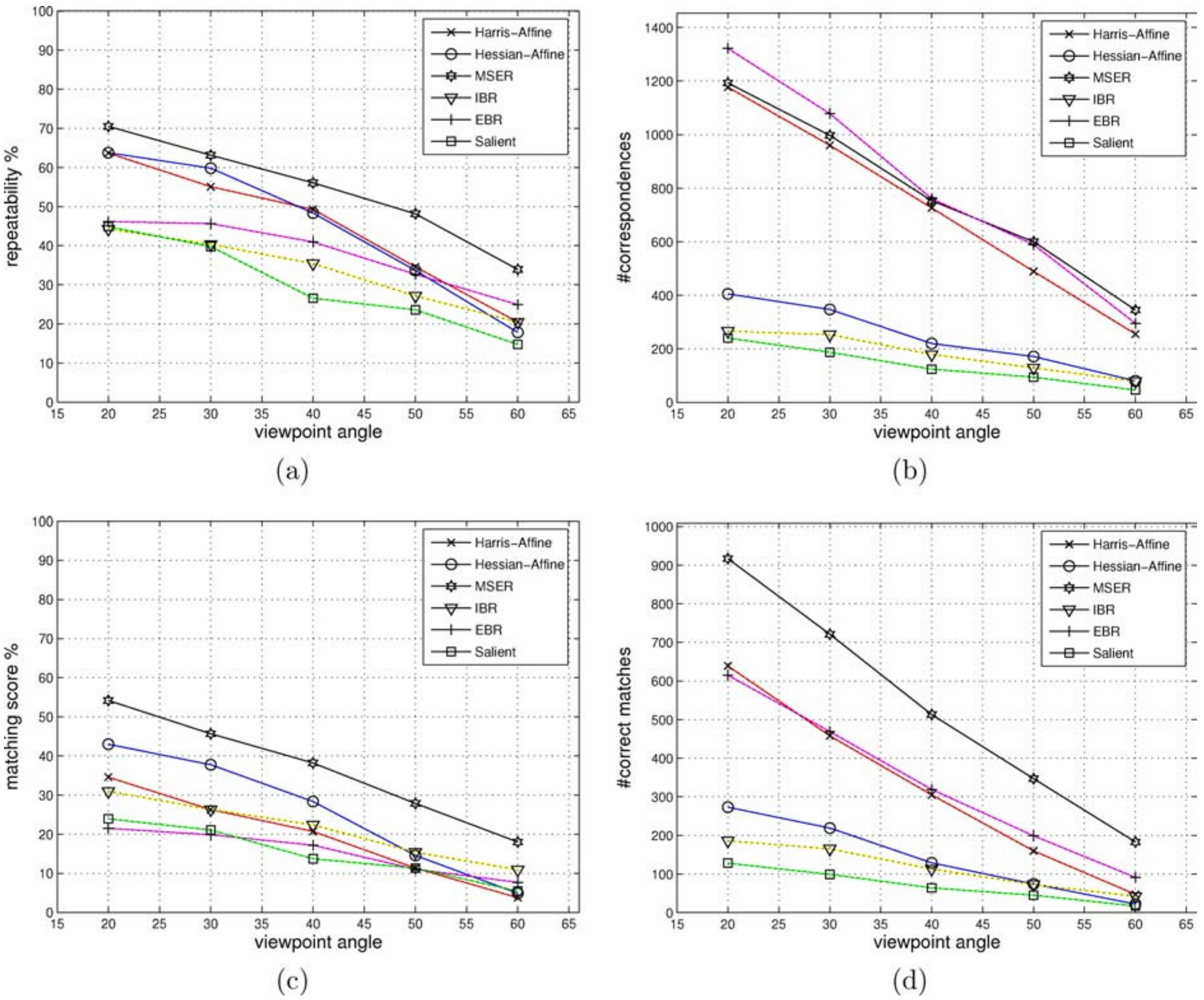


Figure 14. Viewpoint change for the textured scene (Wall sequence Fig. 9(b)). (a) Repeatability score for viewpoint change (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

illumination changes, although the MSER obtains the highest repeatability score for this type of scene. The absolute score shows how a small transformation of this type of a scene can affect the repeatability of different detectors.

General Conclusions. For most experiments the MSER regions or Hessian-Affine obtain the best repeatability score and are followed by Harris-Affine. Salient regions give relatively low repeatability. For the edge-based region detector, it largely depends on the scene content, i.e., whether the image contains stable curves or not. The intensity extrema-based region detector gives average scores. Results largely depend on the type of scene used for the experiments. Again, this illustrates the complementarity of the various detectors. Depending on the application, a combination of detectors is probably prudent.

Viewpoint changes are the most difficult type of transformation to cope with, followed by scale changes. All detectors behave similarly under the different types of transformations, except for the blur sequence of Fig. 17, where MSER performs significantly worse than the others.

In the majority of the examples Hessian-Affine and Harris-Affine detector provide several times more corresponding regions than the other detectors. The Hessian-Affine detector almost systematically outperforms the Harris-Affine detector, and the same holds for MSER with respect to IBR.

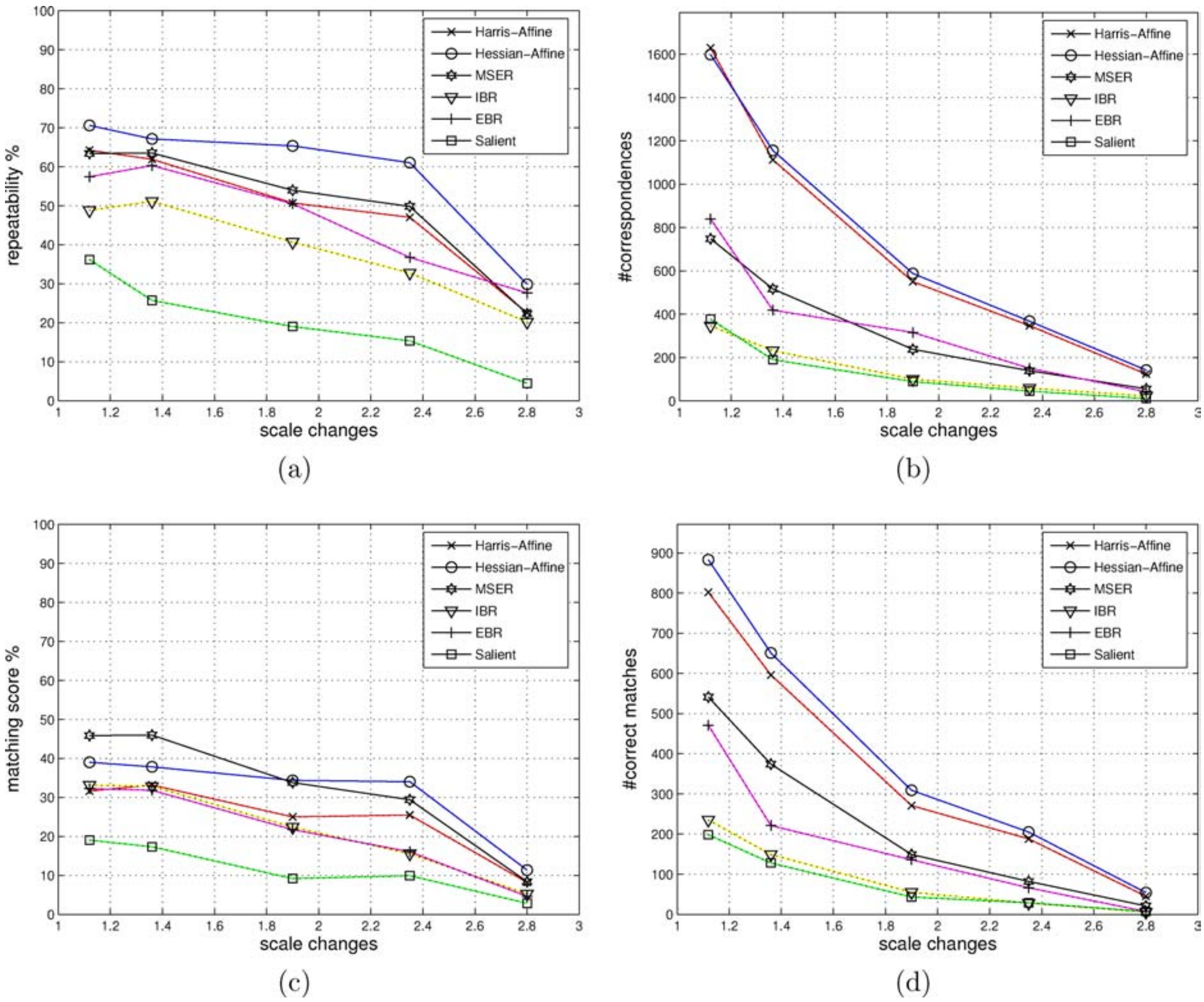


Figure 15. Scale change for the structured scene (Boat sequence Fig. 9(c)). (a) Repeatability score for scale change (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

4.3. More Detailed Tests

To further validate our experimental setup and to obtain a deeper insight in what is actually going on, a more detailed analysis is performed on one image pair with a viewpoint change of 40 degrees, namely the first and third column of the graffiti sequence shown in Fig. 9(a).

Accuracy of the Detectors. First, we test the effect of our choice for the overlap error threshold. This was fixed to 40% in all the previous experiments. Choosing a lower threshold results in more accurate regions, (see Fig. 12). Figure 21(a) shows the repeatability score as a function of the overlap error. Clearly, as the required overlap is relaxed, more regions are qualified as corresponding, and the repeatability scores go up. The relative ordering of the various detectors remains virtually the same, except for the Harris-Affine and Hessian region detectors. They improve their ranking with increasing overlap error, which means that these detectors are less accurate than the others–at least for this type of scene.

Choice of Normalized Region Size. Next, we test the effect of our choice of the normalized region size. This was fixed to a radius of 30 pixels in all the previous experiments. Figure 21(b) shows how the repeatability scores vary as a function of the normalized region size, with the overlap error threshold fixed to 40%. The relative ordering of the different detectors stays the

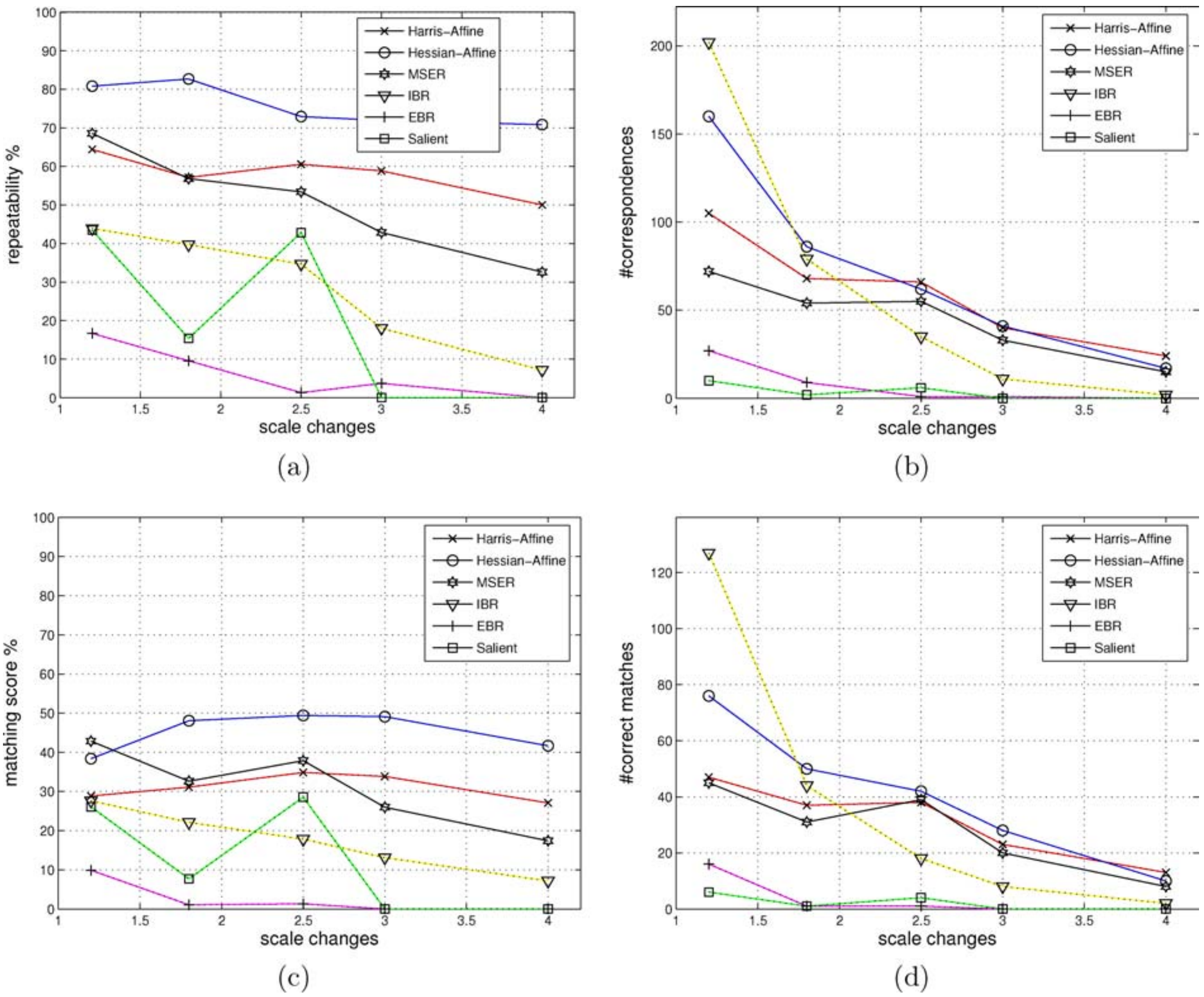


Figure 16. Scale change for the textured scene (Bark sequence Fig. 9(d)). (a) Repeatability score for scale change (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

same, which indicates that our experimental setup is not very sensitive to the choice of the normalized region size. With a larger normalized region size, we obtain lower overlap errors and the curves increase slightly (see also Fig. 11).

Varying the Region Density. For some detectors, it is possible to vary the number of detected regions, simply by changing the value of one significant parameter. This makes it possible to compensate for the effect that different region densities might have on the repeatability scores and compare different detectors when they output similar number of regions. Figure 21(c) shows that the repeatability of MSER (92%) and IBR (63%) is high for a small number of regions (70) and decreases to 68% and 50% respectively for 350 detected regions, unlike the repeatability for Hessian-Laplace, Harris-Laplace and salient regions which is low for a small number of regions and increases for more than 300 regions. However, the rank of the detectors remains the same in the range of available threshold settings, therefore the order of the detectors in the experiments in the previous section is not affected by the density of regions. Depending on the application and the required number of regions one can set the appropriate threshold to optimize the performance.

Repeatability Score as a Function of Region Size. Rather than normalizing all regions to a fixed region size prior to computing the overlap error, an alternative

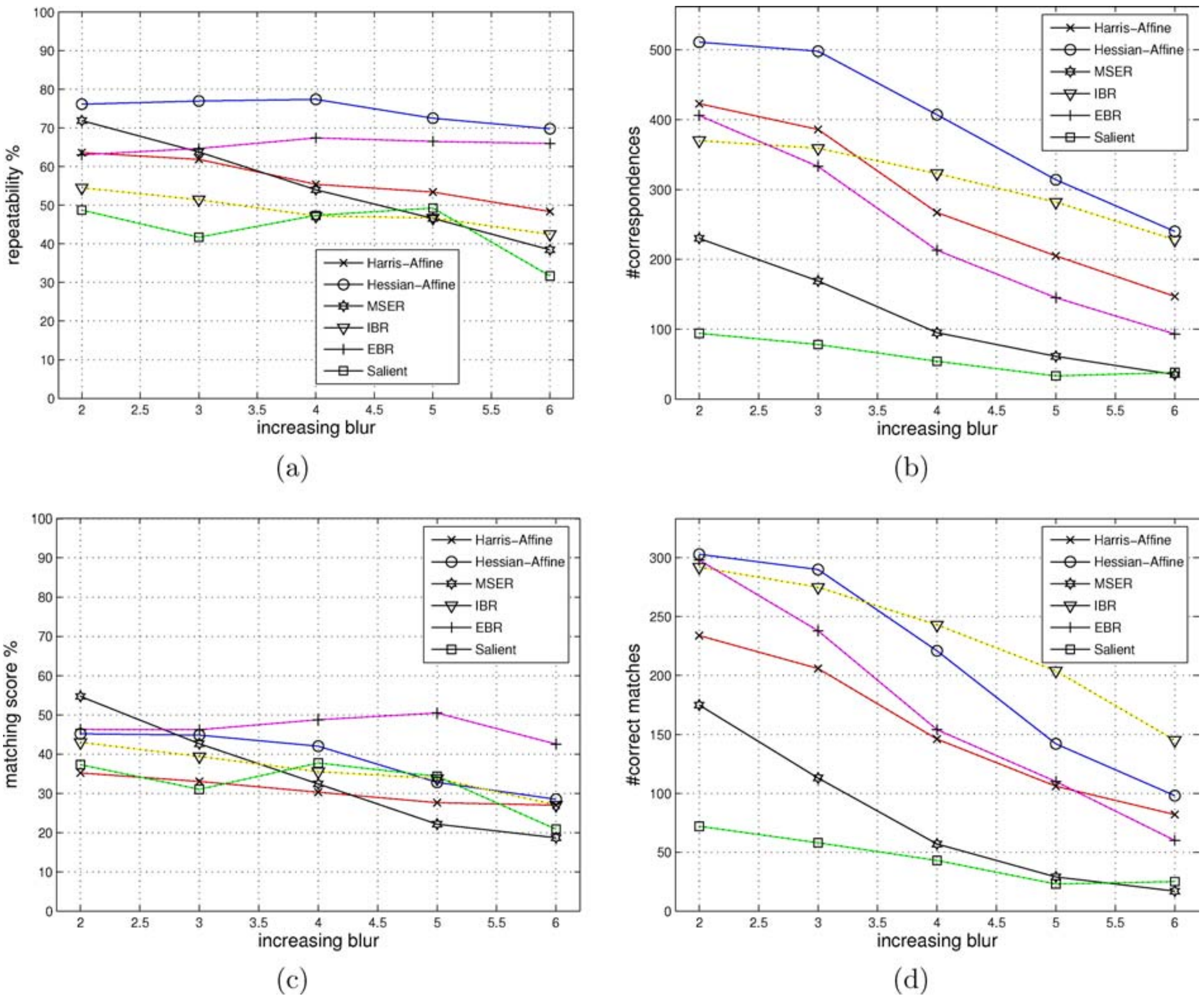


Figure 17. Blur for the structured scene (Bikes sequence Fig. 9(e)). (a) Repeatability score for blur change (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

approach would be to only compare regions of similar sizes. This results in a plot showing the repeatability scores for different detectors as a function of region size. Large regions typically yield higher repeatability scores, not only because of their intrinsic stability, but also because they automatically yield lower overlap errors. Figure 21(d) shows the repeatability with respect to detected region size. MSER detector has the highest repeatability score and it is nearly the same for different size of the detected regions. The results for Hessian-Affine, Harris-Affine and IBR are similar. The repeatability is low for small regions, then it increases for medium size regions and slightly decreases for larger regions except that the score for Harris-Affine decreases more rapidly. The repeatability for EBR and salient regions is small for small and medium size regions and increases for large regions. Note, that the repeatability for different region size depends also on the type of image transformation i.e., for large scale changes only the small regions from one image will match with the large regions from the other one.

5. Matching Experiments

In the previous section, the performance of the different region detectors is evaluated from a rather theoretical point of view, focusing on the overlap error and repeatability. In this section, we follow a more practical approach. In a practical application, regions need to be

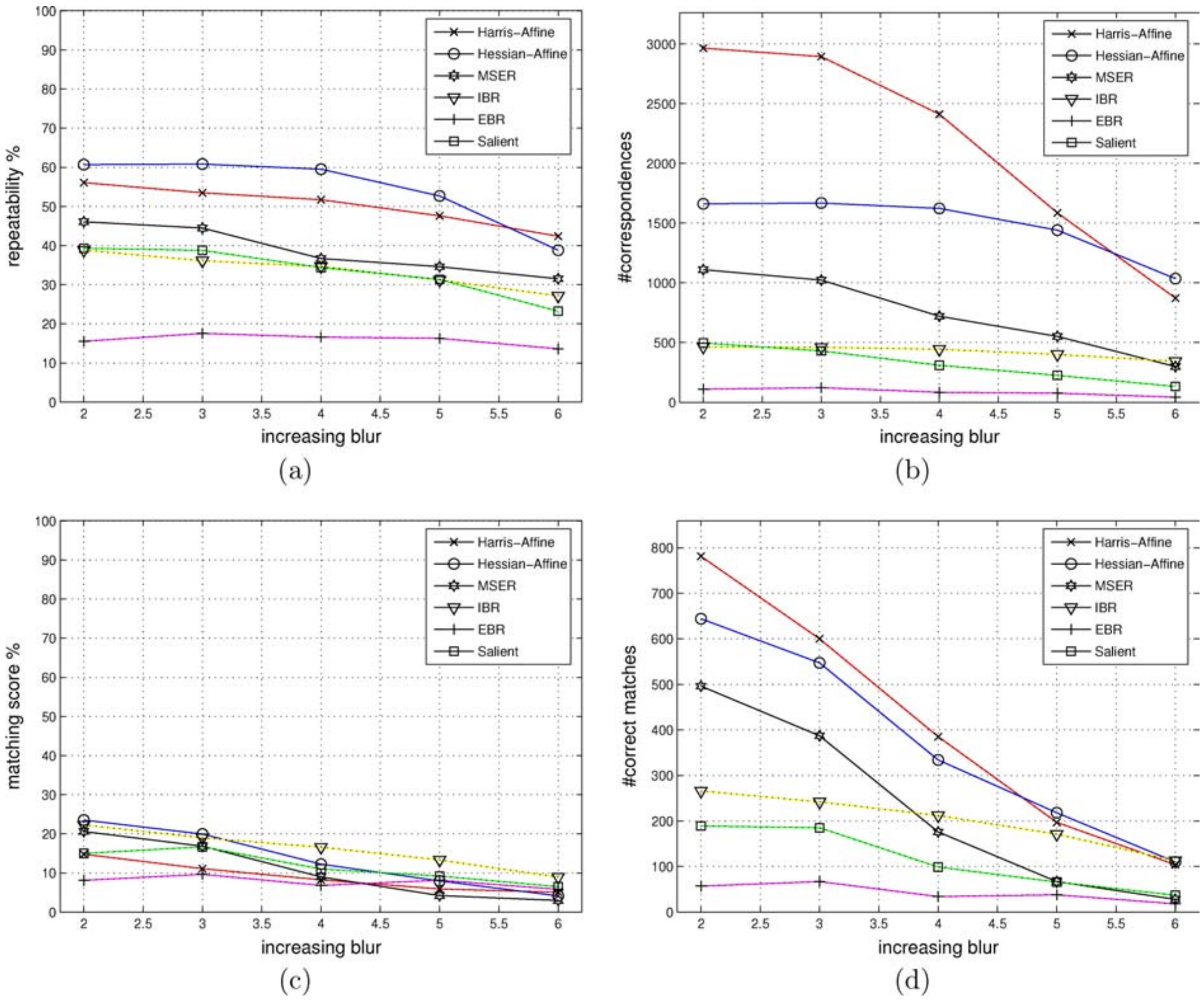


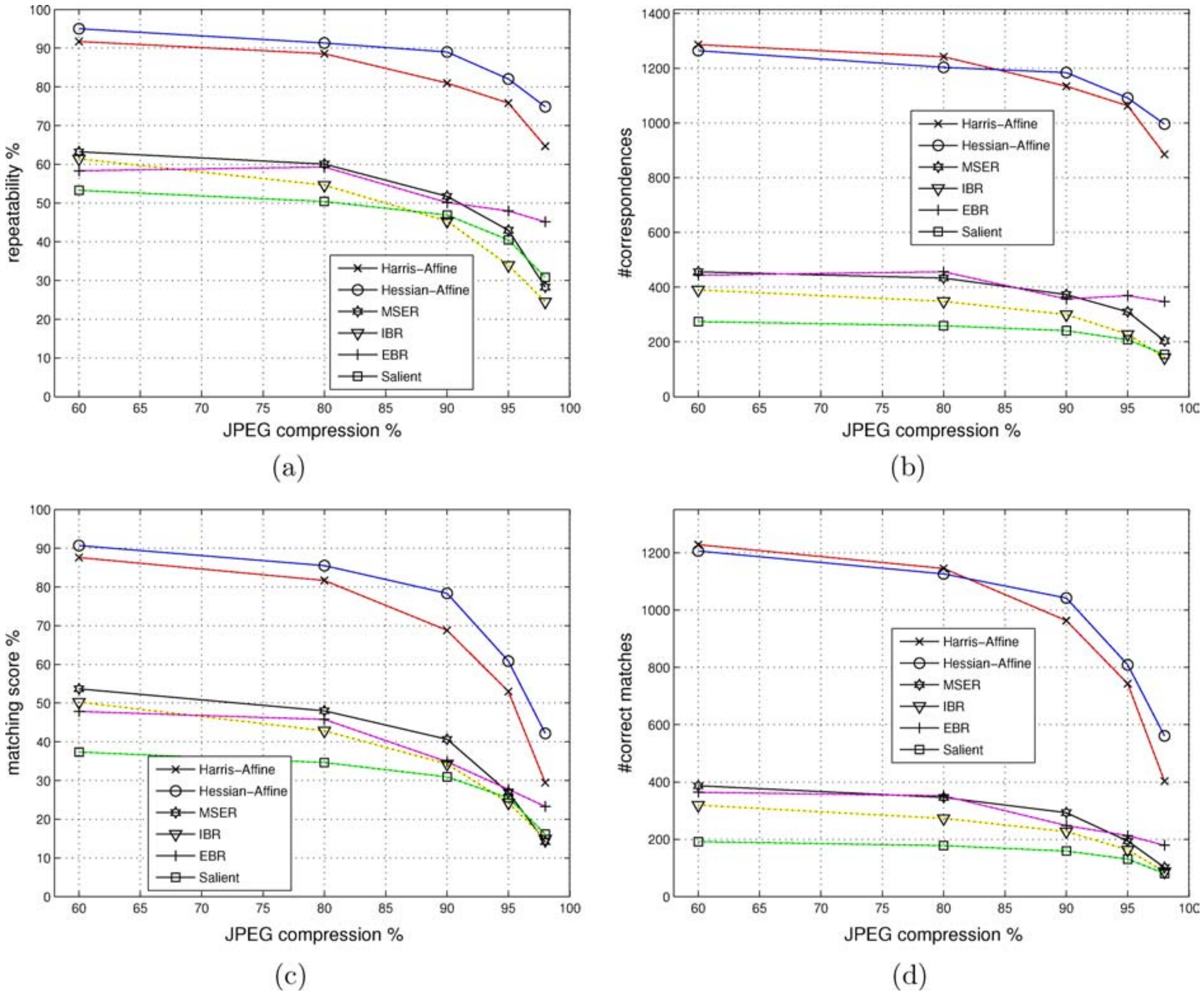
Figure 18. Blur for the textured scene (Trees sequence Fig. 9(f)). (a) Repeatability score for blur change (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

matched or clustered, and apart from the accuracy and repeatability of the detection, also the distinctiveness of the region is important. We test how well the regions can be matched, looking at the number of matches found as well as the ratio between correct matches and mismatches.

To this end, we compute a descriptor for the regions, and then check to what extent matching with the descriptor gives the correct region match. Here we use the SIFT descriptor of Lowe (1999). This descriptor gave the best matching results in an evaluation of different descriptors computed on scale and affine invariant regions (Mikolajczyk and Schmid, 2003, 2005). The descriptor is a 128 dimensional vector computed from the spatial distribution of image gradients over a circular region. To this end, each elliptical region is first mapped to a circular region of 30 × 30 pixels, and rotated based on the dominant gradient orientation, to compensate for the affine geometric deformations, as shown in Fig. 2(e). Note that unlike in Section 4, this mapping concerns descriptors; the region size is coincidentally the same (30 Pixels).

5.1. *Matching Score*

Again the measure is computed between a reference image and the other images in a set. The matching score is computed in two steps.



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

Figure 19. JPEG compression (UBC sequence Fig. 9(g)). (a) Repeatability score for different JPEG compression (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

1. A region match is deemed correct if the overlap error defined in the previous section is minimal and less than 40%, i.e., $\epsilon_O \leq 0.4$. This provides the ground truth for correct matches. Only a single match is allowed for each region.
2. The *matching score* is computed as the ratio between the number of correct matches and the smaller number of detected regions in the pair of images. A match is the nearest neighbour in the descriptor space. The descriptors are compared with the Euclidean distance.

This test gives an idea on the distinctiveness of features. The results are rather indicative than quantitative. If the matching results do not follow those of the repeatability test for a particular feature type that means that the distinctiveness of these features differs from the distinctiveness of other detectors.

The Effect of Rescaling the Regions. Here, the issue arises on what scale to compute the descriptor for a given region. Indeed, rather than taking the original distinguished region, one might also rescale the region first, which typically leads to more discriminative power–certainly for the small regions. Figure 22(c) shows how the matching score for the different detectors varies for different scale factors. Typically, the curves go slightly up for larger measurement regions, except for EBR and salient regions which attain their maximum score for scale factor of 2 and 3 respectively.

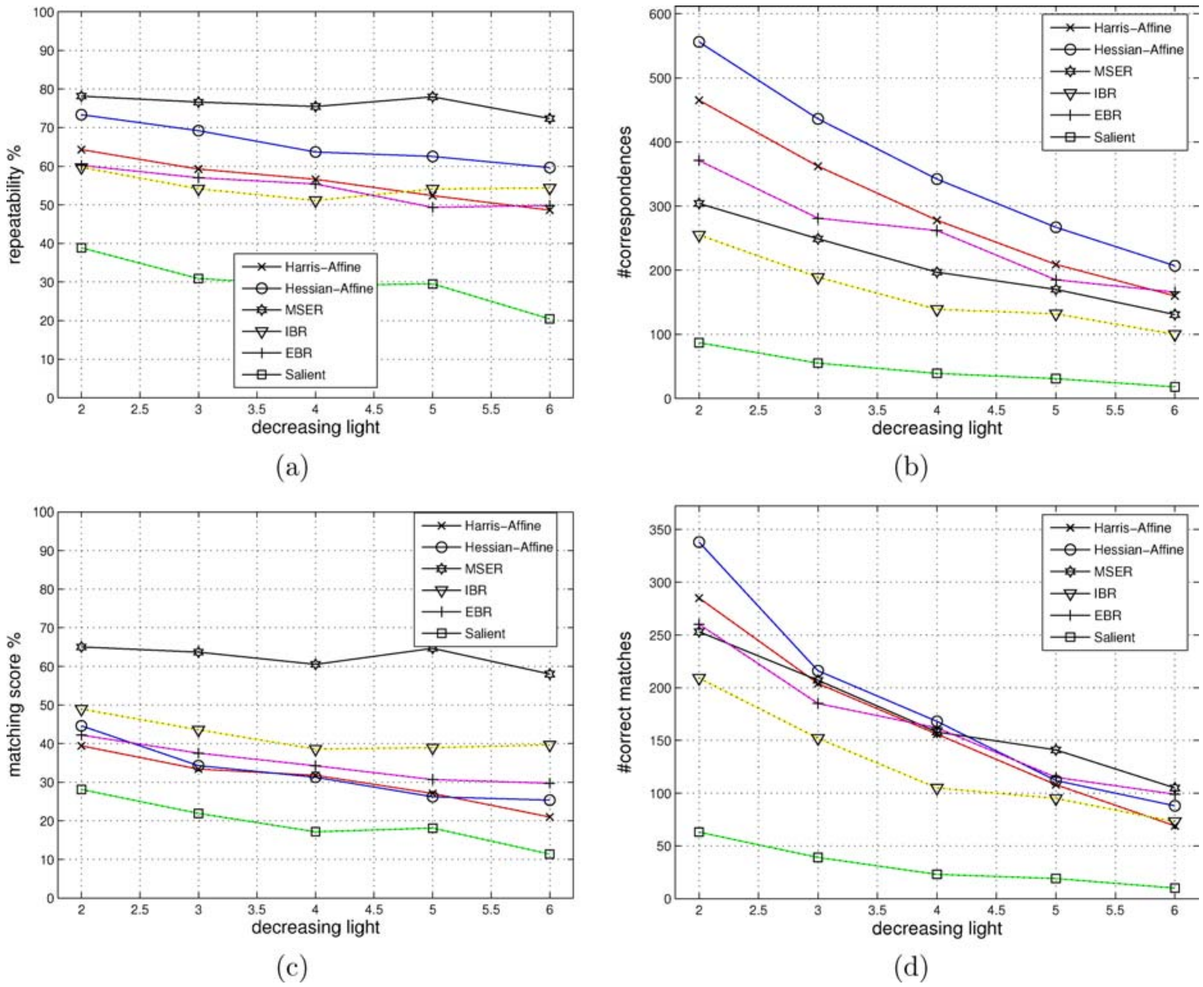


Figure 20. Illumination change (Leuven sequence Fig. 9(h)). (a) Repeatability score for different illumination (default settings). (b) Number of corresponding regions. (c) Matching score. (d) Number of correct nearest neighbour matches.

However, except for EBR the relative ordering of the different detectors remains unaltered. For all our matching experiments, we selected a scale factor of 3.

It should be noted though that in a practical application a large scale factor can be more detrimental, due to the higher risk of occlusions or non-planarities. Since in our experimental setup all images are related by homographies, these effects do not occur.

5.2. *Matching Under Various Transformations*

Figures 13 –20(c) and (d) give the results of the matching experiment for the different types of transformations. These are basically the same plots as given in Figs. 13–20(a) and (b) but now focusing on regions that have actually been matched, rather than just corresponding regions.

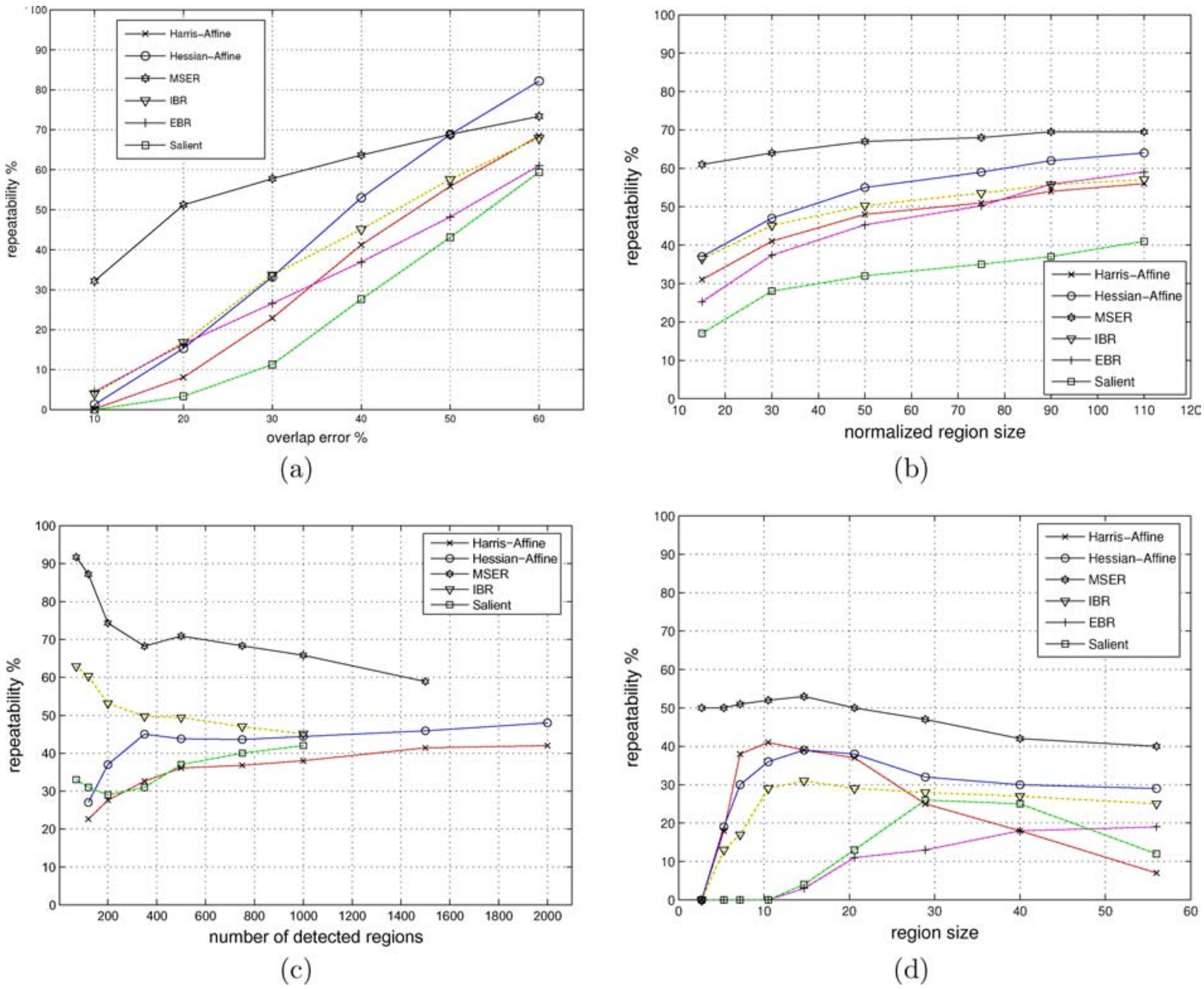
For most transformations, the plots look indeed very similar to (albeit a bit lower than) the results obtained with the overlap error test. This indicates that the regions typically have sufficient distinctiveness to be matched automatically. One should be careful though to generalize these results because these might be statistically unreliable, e.g. for much larger numbers of features in database retrieval.

Sometimes, the score or relative ordering of the detectors differs significantly from the overlap error tests of the previous section (e.g. Figs. 17 and 20). This means that the regions found by some detectors are not distinctive and many mismatches occur.

The ranking of the detectors changes in Fig. 20(c) comparing to Fig. 20(a) which means the Harris-Affine and Hessian-Affine are less distinctive. These detectors find several slightly different regions containing the same local structure all of which have a small overlap error. Thus, the matched regions might have the overlap smaller than 40% but the minimum overlap error is for a slightly different region. In this way the matched regions are counted as incorrect. The same change in ranking for Harris-Affine and Hessian-Affine can be observed on the results for other transformations. However the rank of the fig. (d) showing the number of matched regions do not change with respect to the number of corresponding regions on figures (b).

The curves for Fig. 18(c) and (d) give the results for the textured scene shown in Fig. 9(f). For this case, the matching scores are significantly lower than the repeatability scores obtained earlier. This can be explained by the fact that the scene contains many similar local structures, that can hardly be distinguished.

Ratio Between Correct and False Matches. So far, we investigated the matching capability of corresponding regions. In a typical matching application, what matters is the ratio between correct matches and false matches, i.e., are the regions within a correct match more similar to each other than two regions that do not correspond but accidentally look more or less



(a)

(b)

(c)

(d)

Figure 21. Viewpoint change (Graffiti image pair - 1st and 3rd column in Fig. 9(a)). (a) Repeatability score for different overlap error for one pair (normalized size = 30 pixels). (b) Repeatability score for different normalized region size (overlap error <40%). (c) Repeatability score for different number of detected regions (overlap error = 40%, normalized size = 30 pixels). (d) Repeatability score as a function of region size.

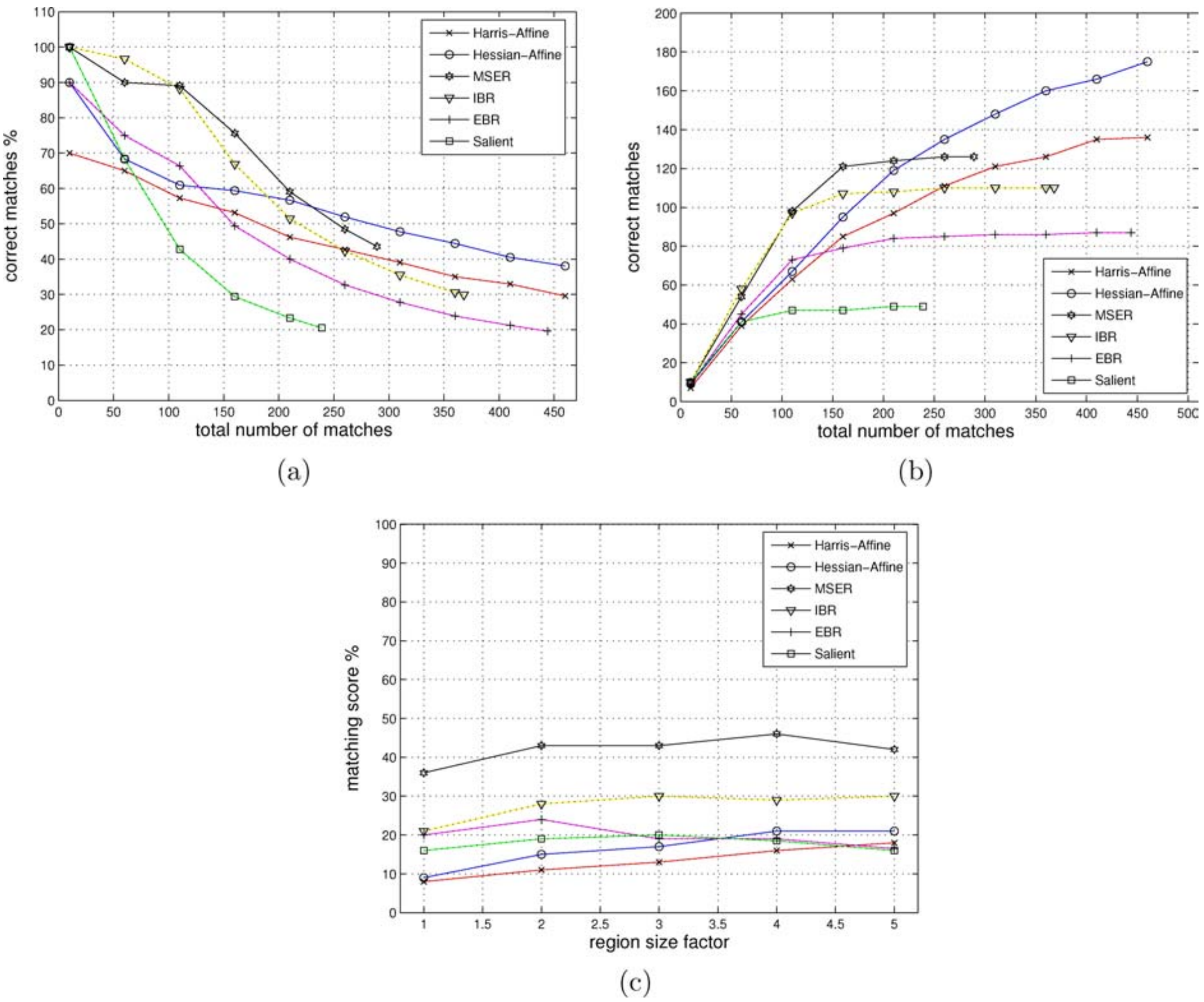


Figure 22. Viewpoint change (Graffiti image pair - 1st and 3rd column in Fig. 9**(a)**). (a) Percentage of correct matches versus total number of nearest neighbour matches. (b) Number of correct matches versus total number of nearest neighbour matches. (c) Matching score for different size of measurement region. Region size factor is the ratio measurement/detected region size.

similar ? Here, the accuracy of the region detection plays a role, as does the variability of the intensity patterns for all regions found by a detector, i.e., the distinctiveness. Figure 22(a) shows the percentage of correct matches as a function of the number of matches. A match is the nearest neighbour in the SIFT feature space. These curves were obtained by ranking the matches based on the distance between the nearest neighbours. To obtain the same number of matches for different detectors the threshold was individually changed for each region type.

As the threshold, therefore the number of matches increases (Fig. 22(a)), the number of correct as well as false matches also increases, but the number of false matches increases faster, hence the percentage of correct matches drops. For a good detector, a small threshold results in almost exclusively correct matches. Figure 22(b) shows the absolute number of correct matches with respect to the total number of matches. We observe that MSER and IBR provide a large number of correct matches for a small descriptor threshold. Up to 100 matches more than 90% are correct. This means one does not have to rely so heavily on semi-local or global consistency checks to remove the false matches afterwords. Harris-Affine and Hessian-Affine obtain low score but improve when the distance is larger.

Depending on the application, the number of matches a user is interested in may vary. If only a very small number of matches is needed (e.g. for computing epipolar geometry), the MSER or IBR detector is the best choice for this type of scene. Above 200 matches, Hessian-Affine and Harris-Affine perform better–albeit at the cost of a large false positive rate.

6. Conclusions

In this paper we have presented the state of the art on affine covariant region detectors and have compared their performance. The comparison has shown that the performance of all presented detectors declines slowly, with similar rates, as the change of viewpoint increases. There does not exist one detector which outperforms the other detectors for all scene types and all types of transformations. In many cases the highest score is obtained by the MSER detector, followed by Hessian-Affine. MSER performs well on images containing homogeneous regions with distinctive boundaries. This also holds for IBR since both methods are designed for similar region types. Hessian-Affine and Harris-Affine provide more regions than the other detectors, which is useful in matching scenes with occlusion and clutter. EBR is suitable for scenes containing intersections of edges. Salient regions obtained low scores in this evaluation but performed well in the context of object class recognition (Kadir et al., 2004).

The detectors are complementary, i.e., they extract regions with different properties and the overlap of these regions is small if not empty. Several detectors should be used simultaneously to obtain the best performance. The output of different detectors can be combined by concatenating the respective matches. This increases the number of matches and therefore the robustness to occlusion, at the expense of processing time. The choice of the optimal subset depends on the context, for example on the required number of extracted regions and processing time. In general, matching on descriptors alone is not sufficient (as some are mismatched), and further steps are required to disambiguate matches (Ferrari et al., 2004; Rothganger et al., 2003; Schaffalitzky and Zisserman, 2002; Sivic and Zisserman, 2003). These steps depend on the application, but generally use methods of geometric filtering based on the local spatial arrangement of the regions, or on multiple view geometric relations.

Another contribution of the paper is the carefully designed test protocol which is available on the Internet together with the test data. This allows the evaluation of future detectors and their comparison with those studied in the paper. Note that the criteria, as defined here, are only valid for planar scenes or in the case of camera rotation or zoom. Only in these cases is the geometric relation between two images defined by a homography. However, many 3D objects are composed of smooth surfaces, which are planar in the small–that is, sufficiently small patches can be treated as being comprised of coplanar points. Naturally, regions are also detected at depth and surface orientation discontinuities of 3D scenes. Evaluating the repeatability of such regions is beyond the scope of this paper.

Research on covariant regions and their description is now well advanced–they are the building blocks for general recognition systems–but more remains to be done. One direct generalization is to apply the detectors to representations of the image other than intensity, for example ordering functions such as 'saturation' or 'projection on the red-blue direction in RGB space' could be used. Furthermore, affine detectors for shape (object boundaries) or completely unstructured textures should be developed. Finally, an important issue is how to design detectors for images of an object class, where there is within-class variation in addition to affine viewpoint changes, and how to measure their repeatability (Kadir et al., 2004).

Acknowledgments

This research was initiated and supported by the European FET Fifth Framework project VIBES. K. Mikolajczyk was funded by an INRIA postdoctoral fellowship, T. Tuytelaars by the Fund for Scientific Research-Flanders, J. Matas by the Czech Academy of Sciences project 1ET101210407 and T. Kadir by the European project CogViSys. F. Schaffalitzky is a fellow of Balliol College, Oxford.

References

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

Brown, M. and Lowe, D. 2003. Recognizing panoramas. In *Proceedings of the International Conference on Computer Vision*, Nice, France, pp. 1218–1225.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8: 679–698.

Csurka, G., Dance, C., Bray, C., and Fan, L. 2004. Visual categorization with bags of keypoints. In *Proceedings Workshop on Statistical Learning in Computer Vision*.

Dorko, G. and Schmid, C. 2003. Selection of scale invariant neighborhoods for object class recognition. In *Proceedings International Conference on Computer Vision*, Nice, France, pp. 634–640.

Fergus, R., Perona, P., and Zisserman, A. 2003. Object class recognition by unsupervised scale-invariant learning. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA.

Ferrari, V., Tuytelaars, T., and Van Gool, L. 2001. Simultaneous object recognition and segmentation by image exploration. In *Proceedings European Conference on Computer Vision*, Prague, Czech Republic, pp. 40–54.

Ferrari, V., Tuytelaars, T., and Van Gool, L. 2005. Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, to appear.

Goedeme, T., Tuytelaars, T., and Van Gool, L. 2004. Fast wide baseline matching for visual navigation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 24–29.

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

Hartley, R.I. and Zisserman, A. 2004. *Multiple View Geometry in Computer Vision*, 2nd edition, Cambridge University Press, ISBN: 0521540518.

Kadir, T., Zisserman, A., and Brady, M. 2004. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision*, Prague, Czech Republic, pp. 345–457.

Lazebnik, S., Schmid, C., and Ponce, J. 2003a. A sparse texture representation using affine-invariant regions. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, pp. 319–324.

Lazebnik, S., Schmid, C., and Ponce, J. 2003b. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proceedings of the International Conference on Computer Vision*, Nice, France, pp. 649–655.

Lazebnik, S., Schmid, C., and Ponce, J. 2005. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8):1265–1278.

Lindeberg, T. and Gårding, 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.

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

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

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision* 60(2):91–110.

Matas, J., Burianek, J., and Kittler, J. 2000. Object Recognition using the Invariant Pixel-Set Signature. In *Proceedings of the British Machine Vision Conference*, London, UK, pp. 606–615.

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

Matas, J., Chum, O., Urban, M., and Pajdla, T. 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing* 22(10):761–767.

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.

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

Mikolajczyk, K. and Schmid, C. 2003. A performance evaluation of local descriptors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA.

Mikolajczyk, K., Zisserman, A., and Schmid, C. 2003. Shape recognition with edge-based features. In *Proceedings of the British Machine Vision Conference*, Norwich, UK.

Mikolajczyk, K. and Schmid, C. 2004. Scale & affine invariant interest point detectors. *International Journal on Computer Vision* 60(1):63–86.

Mikolajczyk, K. and Schmid, C. 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10):1615–1630.

Obdržálek, Ŝ. and Matas, J. 2002. Object recognition using local affine frames on distinguished regions. In *Proceedings of the British Machine Vision Conference*, Cardiff, UK, pp. 113–122.

Opelt, A., Fussenegger, M., Pinz, A., and Auer, P. 2004. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of European Conference on Computer Vision*, Prague, Czech Republic, pp. 71–84.

Pritchett, P. and Zisserman, A. 1998. Wide baseline stereo matching. In *Proceedings of the 6th International Conference on Computer Vision*, Bombay, India, 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 IEEE Conference on Computer Vision and Pattern Recognition*, Madison, Wisconsin, USA, pp. 272–277.

Rothganger, F., Lazebnik, S., Schmid, C., and Ponce, J. 2005. Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial consraints. *International Journal of Computer Vision*, to appear.

Schaffalitzky, F., and Zisserman, A. 2002. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark, pp. 414–431.

Schaffalitzky, F. and Zisserman, A. 2003. Automated Location matching in movies. *Computer Vision and Image Understanding*, 92(2):236–264.

Schmid, C. and Mohr, R. 1997. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(5):530–535.

Se, S., Lowe, D., and Little, J. 2002. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research* 21(8):735–758.

Sedgewick, R. 1988. *Algorithms*, 2nd edition. Addison-Wesley.

Sivic, J., and Zisserman, A. 2003. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, Nice, France.

Sivic, J., Schaffalitzky, F., and Zisserman, A. 2004. Object level grouping for video shots. In *Proceedings of the 8th European*

Conference on Computer Vision, Prague, Czech Republic, pp. 724–734.

Sivic, J., and Zisserman, A. 2004. Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, pp. 488–495.

Tell, D. and Carlsson, S. 2000. Wide baseline point matching using affine invariants computed from intensity profiles. In *Proceedings of the 6th European Conference on Computer Vision*, Dublin, Ireland, pp. 814–828.

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. 68–81.

Turina, A., Tuytelaars, T., and Van Gool, L. 2001. Efficient Grouping under perspective skew. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, USA, pp. 247–254.

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

Tuytelaars, T., Van Gool, L., D'haene, L., and Koch, R. 1999. Matching of affinely invariant regions for visual servoing. In *Int. Conference Robotics and Automation ICRA 99*.

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

Tuytelaars, T. and Van Gool, L. 2004. Matching Widely Separated Views based on Affine Invariant Regions. *International Journal on Computer Vision* 59(1):61–85.