A revised evaluation framework for region of interest detectors using artificial 3d-scenes

Stephan Saalfeld

May 24, 2007

Contents

List of Figures

Abstract

In this project, we introduce an evaluation method for regions of interest detectors in images of general 3d-scenes addressing some issues of the framework by Mikolajczyk et al. (2005) and the extension to non-planar scenes by Fraundorfer and Bischof (2005). By generating more reliable ground truth and test images from detailed POV-Ray scenes, we were able to take the influence of *discontinuities* and *occlusion* into account. We re-evaluate the repeatability under viewpoint change of the Harris- and Hessian-Affine region detectors by Mikolajczyk and Schmid (2004), an edge based (EBR) and an intensity based (IBR) region detector by Tuytelaars and Van Gool (2004), the maximally stable extremal region (MSER) detector Matas et al. (2002) and the Difference of Gaussian (DoG) detector by Lowe (1999). In contrast to Fraundorfer and Bischof (2005), we regard the scale invariant DoG detections to be circular regions instead of keypoints. We discuss the influence of discontinuities in 3d-scenes in general, and evaluate each detector's tendency to extract regions on such discontinuities.

Chapter 1

Introduction

For various vision based tasks, it is a promising approach to preselect distinguished *points or regions of interest* in camera images at an early stage, in order to reduce the complexity of further processing. Such points or regions, subsumed as *low level visual features*, should represent meaningful properties of the image with respect to the task. In the cases of object recognition, wide baseline stereo, real time localisation and mapping, alignment of overlapping image patches, face recognition and related challenges, such a meaningful property is the correspondence of a point or region to a visible 3d-location, regardless of its content on a higher interpretation level.[1] That is, a detected feature should match to the same physical ground truth, not affected by photometric and geometric transformations applied through capturing the image. Furthermore it should be reliably detected regardless of the camera's pose and illumination conditions, thus providing high *repeatability* (Schmid et al., 2000).

In camera images, the most important geometric transformations come from changing the viewpoint, the camera's orientation and focal length. The most important photometric transformations are caused by changing illumination conditions and exposure time. While geometric transformations change the location and the shape of a feature in the image, photometric transformations affect its colour or intensity appearance. The detector should be both covariant to geometric and invariant to—or at least robust against—photometric transformations. Photometric invariance is a detector's ability to detect a feature reliably and unaffected by illumination or exposure changes. Geometrical covariance for an interest point detector

[1] For a widespread collection of applications, we refer to the introduction and references by Mikolajczyk et al. (2005).

means that the point will be detected at the transformed location. For a covariant region detector, in addition to the location, a subset of the applied transformations has to be extracted. That is, a point or region of interest should be re-detected reliably, well located and shaped whenever it is visible in the image.

Having a unique description vector for each of these features, it is possible to extract 3d-information about a moving scene or camera from pairwise association of the features in complementary camera images, that is images of the same scene taken from different viewpoints. Generally, the description vector is extracted from the image texture in the neighbourhood of the detected point location or inside the detected region. Normalisation of this texture pattern with respect to the set of transformations expressed by the detector guarantees the extracted descriptor to be invariant to these transformations. There is a tradeoff between invariance and distinctiveness of the descriptor. The geometrically normalized image pattern itself provides the highest possible distinctiveness but is very sensitive to even small localisation errors and noise. A descriptor being equal for each feature, would be perfectly invariant to every transformation, but could not be used for unique pairwise association of features.

In the computer vision literature, various techniques for extracting such points or regions of interest were published (see Mikolajczyk et al., 2005). The set of transformations a detector is covariant to allows a hierarchical classification of the proposed techniques. We focus on the most advanced family, region detectors being covariant to affine (Matas et al., 2002; Mikolajczyk and Schmid, 2004; Tuytelaars and Van Gool, 2004) or at least similarity transformations (Lowe, 1999).

This work addresses the problem of evaluating the performance of such region detectors under viewpoint change in general 3d-scenes. We refer to and use the generally accepted work by Mikolajczyk et al. (2005) and Fraundorfer and Bischof (2005) and compare our results with theirs. We introduce a new evaluation criterion, the *discontinuity ratio* and discuss its relevance for further research.

In Chapter 2, we briefly reference the related work. After introducing notational conventions and basic concepts, like the pinhole camera model and homogeneous transformations, we introduce the examined detectors in Chapter 3. In Chapter 4, we substantiate our decision to evaluate the Difference of Gaussian (DoG) detector as a region detector instead of a point detector. We illustrate our evaluation framework and method and compare it to the related work by Fraundorfer and Bischof (2005). In Chapter 5, we show and discuss the experimental results and, after all, present our conclusions.

Like Fraundorfer and Bischof (2005), we use the publicly available implementations of the MSER, the Harris- and Hessian-Affine and the IBR and EBR detectors by Mikolajczyk et al. (2005) for our experiments. The detailed design of the detectors is beyond the scope of this work. For further information, we refer to the excellent synopses of Mikolajczyk et al. (2005, Section 2) and Lowe (2004).

Chapter 2

Related Work

In 2000, Schmid et al. presented their evaluation framework for interest point detectors. They defined interest points as 2d-locations with a significantly changing signal in both dimensions and introduced the *repeatability score* as an evaluation criterion for the ‘geometrical stability of the detected interest points [...] under varying viewing conditions’ (Schmid et al., 2000, page 151). A high repeatability score signalises the capability of a detector to re-detect the same visible 3d-locations independently from mapping through changing viewpoints and illumination.

They evaluated each examined detector’s repeatability score under image noise, illumination variation, zoom (scale change), image rotation and viewpoint change. The evaluation is performed on images of two planar scenes. For each transformation, there is a sequence of images per scene with progressive transformation rate. For the affine transformations and viewpoint change, the ground truth projection of a location in image $\mathbf{I}_i$ to image $\mathbf{I}_j$ is defined by a priorly determined homography $\mathbf{H}_{ij}$. Furthermore, the authors discuss the repeatability score depending on *localisation error* ϵ.

An improved version of the *Harris* detector (Harris and Stephens, 1988) by the authors showed the best results compared to the inspected detectors, especially in presence of affine and viewpoint changes. Results are very good for a maximum localisation error $\epsilon \geq 1.5$px.

Mikolajczyk et al. (2005) extended this framework for the evaluation of affine covariant region detectors using a similar repeatability score. This work is strongly related to the ‘Performance evaluation of local descriptors’ by Mikolajczyk and Schmid (2003), where they reviewed state of the art descriptors for such regions regarding their distinctiveness, robustness and adequacy for the detector-specific image content. The *Scale Invariant Feature*

Transform (SIFT) descriptor by Lowe (2004) and an improved SIFT-based descriptor by the authors showed the best results independently from the detector.

They published an image data set similar to that of Schmid et al. (2000) for the following image changes: JPEG-compression, illumination variation, blur and zoom (scale change), image rotation and viewpoint change. The evaluation included the *Harris-* and *Hessian-Affine* region detector (Mikolajczyk and Schmid, 2004), an *intensity based* (IBR) and an *edge based* (EBR) region detector (Tuytelaars and Van Gool, 2004), a *maximally stable extremal region* (MSER) detector (Matas et al., 2002) and a *salient region* detector (Kadir et al., 2004). The maximally stable extremal region (MSER) detector showed the highest repeatability score in all test sequences except for blur and JPEG-compression.

Correspondences are identified by region overlapping instead of localisation accuracy, thus considering the regions shapes. The maximum overlap error for a correspondence was fixed to $\epsilon_O = 40\%$ in their experiments. Introducing a *matching score*, the authors rated the distinctiveness of the extracted regions. For each region, they extract the SIFT descriptor from an intensity pattern, geometrically normalised with respect to the transformation expressed by the detector. A true match is the nearest neighbour in SIFT descriptor space if the overlap error of both regions ϵ_O is lower than 40%.

While the MSER and IBR detectors showed very good matching scores but a low absolute number of matches, the Harris- and Hessian-Affine detectors provide a higher absolute number of matches, but also a higher rate of false positives. The region detectors were found to be complementary. Thus, the combination of different detectors would necessarily increase the absolute number of correct detections and matches as well as the required processing time.

Both, the frameworks of Schmid et al. (2000) and Mikolajczyk et al. (2005), restrict the evaluation in case of viewpoint change to planar scenes only. The authors argue this to be no limitation due to the fact that influences of occlusion, discontinuities and shadows are not modeled by the detectors (Schmid et al., 2000, see).

In practice, the discussed interest point and region detectors were successfully applied for solving general 3d-reconstruction and -recognition problems like object recognition and simultaneous localisation and mapping (e. g. Lowe, 2004; Se et al., 2001) including these non-planar effects. Thus motivated, Fraundorfer and Bischof (2005) proposed to take them into account, in order to investigate the influence of non-planarity to the specific detectors.

Instead of image sequences of planar scenes, they recorded two sequences, a set of textured boxes and an office interior, with a moving camera, such increasing the viewpoint angle from 0°up to 90°. The camera's parameters are known, the camera is assumed to be perfectly rectified. From the first two images of each scene, they created dense disparity maps through stereo matching. Such they were able to transfer a pixel location from the first two images of the sequence into each of the other images using the trifocal tensor (see Hartley and Zisserman, 2006, part 3), using the disparity matches as ground truth. They re-evaluated the Harris, Hessian and Difference of Gaussian detectors as interest points and the Harris- and Hessian-Affine (Mikolajczyk and Schmid, 2004), the IBR (Tuytelaars and Van Gool, 2004) and MSER (Matas et al., 2002) detectors as affine covariant regions for viewpoint changes comparable to Schmid et al. (2000) and Mikolajczyk et al. (2005). For the interest point detectors, they used the maximum localisation error $\epsilon \geq 1.5$px, introduced by Schmid et al.. For region detectors, they transferred each pixel location in a region detected in the first image to the other image. Then the overlapping rate is the ratio of intersection's over union's cardinalities. Regions overlapping more than 50% are counted as correspondences.

The MSER and the DoG detectors performed best in the box-scene. In the more complex office-scene, the IBR and the DoG detectors were best rated. Referencing Mikolajczyk et al., the authors evaluated the matching score of the detectors using the SIFT descriptor where the MSER detector performed best. Furthermore, they confirm the result of Mikolajczyk et al., that the detectors are complementary.

Chapter 3

Foundations

3.1 Notation

In the following, we specify notational conventions for a common formal representation of reference frames, vectors, matrices and transformations.

3.1.1 Reference frames

We distinguish three equally scaled *reference frames*: The *world coordinate frame* W and the *camera coordinate frame* C in Euclidean 3-space $\mathbb{R}^3$ and the *image coordinate frame* I in Euclidean 2-space $\mathbb{R}^2$. Projective n-spaces require an additional homogeneous component. As depicted in Figure 3.1, the coordinate frames are left-handed like in *POV-Ray* (see Cason et al., 2006). Rotations are also left-handed, looking into the positive direction of the rotation axis, the positive rotation direction is counter-clockwise.

For image processing, there is a *pixel reference frame* P derived from the image coordinate frame through scale and translation. In contrast to most current image processing applications, the left-handedness implies the origin of the frame to be in the lower left corner. In order to transfer a 2d-location in a pixel coordinate frame having its origin in the upper left corner into the left handed pixel coordinate frame used here, one has to flip the v-component.

$$\begin{pmatrix} u^{\mathsf{P_l}} \\ v^{\mathsf{P_l}} \end{pmatrix} = \begin{pmatrix} u^{\mathsf{P_r}} \\ v^{\mathsf{P}}_{max} - v^{\mathsf{P_r}} \end{pmatrix} \tag{3.1}$$

3.1.2 Vectors and matrices

All vectors in Euclidean n-space $\mathbb{R}^n$ are presented as vertical aligned scalar tuples or bold lower case letters, optionally superscripted by the referenced frame. We use x, y, z for the axes-components in 3-space and u, v in 2-space. Using these symbols, the axes of a reference frame A can be denoted as vectors of another reference frame B with superscript A and subscript B, e. g.

$$\mathbf{x}^{\mathsf{W}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \qquad \mathbf{x}^{\mathsf{W}} = (x, y, z)^{\mathsf{T}} \qquad \mathbf{x}^{\mathsf{I}} = (u, v)^{\mathsf{T}}$$

$$\mathbf{y}^{\mathsf{W}}_{\mathsf{W}} = (0, 1, 0)^{\mathsf{T}} \qquad \mathbf{z}^{\mathsf{W}}_{\mathsf{C}} = (a, b, c)^{\mathsf{T}} \qquad \mathbf{x}^{\mathsf{C}}_{\mathsf{W}} = \frac{2\mathbf{x}^{\mathsf{C}}_{\mathsf{C}} + \mathbf{y}^{\mathsf{C}}_{\mathsf{C}}}{|\mathbf{x}^{\mathsf{C}}_{\mathsf{W}}|}$$

Matrices are presented as rectangular scalar arrays or as bold uppercase letters. $\mathbf{I}$ denotes the identity matrix. Matrices can be written as combinations of matrices and vectors, e. g.

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \qquad \tilde{\mathbf{P}}^{\mathsf{CW}} = \left[\mathbf{R} | \mathbf{t}^{\mathsf{W}}\right]$$

Projective transformation necessitates the introduction of homogeneous coordinates, denoted by a superscript tilde, e. g. $\tilde{\mathbf{x}}^{\mathsf{C}} = \tilde{\mathbf{P}}^{\mathsf{CW}} \tilde{\mathbf{x}}^{\mathsf{W}}$. We assume, the reader is familiar with transformations using homogeneous coordinates (see Hartley and Zisserman, 2006, part 0), the most common transformation matrices are collected in Appendix A.

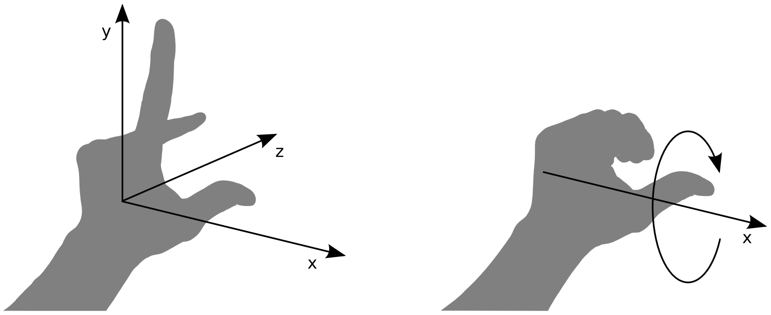


Figure 3.1: Positive rotation direction in left-handed coordinate frames is counter-clockwise.

3.2 The pinhole camera

The pinhole camera model simplifies the mapping of 3d-locations in the world coordinate frame onto a 2d-viewing plane consisting of pixels to a linear perspective projection $\mathbf{P}^{\mathsf{IW}}$ without distortion and other perturbation. We briefly introduce the most important concepts and the parameters used in this work. For a detailed description, we refer to Hartley and Zisserman (2006, part 1).

3.2.1 Inner parameters

The inner parameter set of a pinhole camera applies the transformation from a 3d-location $\mathbf{x}^{\mathsf{C}}$ in the camera coordinate frame to a homogeneous 2d-location $\tilde{\mathbf{x}}^{\mathsf{I}}$ in the image coordinate frame and—furthermore—scale and translation to a homogeneous 2d-location $\tilde{\mathbf{x}}^{\mathsf{P}}$ in the pixel coordinate frame.

The visible frame is defined by the *horizontal angle of view* μ and the vertical over horizontal frame size ratio $r_{v/u}$. Perspective mapping of a 3d-location $\mathbf{x}^{\mathsf{C}}$ from the camera into the image coordinate frame is performed by the perspective transformation matrix $\mathbf{K}^{\mathsf{IC}}$.

$$\tilde{\mathbf{x}}^{\mathsf{I}} = \mathbf{K}^{\mathsf{IC}}\mathbf{x}^{\mathsf{C}} \tag{3.2}$$

$$\mathbf{K}^{\mathsf{IC}} = \begin{bmatrix} \frac{1}{\tan\frac{\mu}{2}} & & \\ & \frac{1}{r_{v/u}\tan\frac{\mu}{2}} & \\ & & 1 \end{bmatrix} \tag{3.3}$$

The image-to-pixel transformation matrix $\mathbf{P}^{\mathsf{PI}}$ translates the origin to the lower left corner of the frame and scales with respect to the *horizontal* and *vertical pixel resolution* u_{max} and v_{max}. After applying $\mathbf{P}^{\mathsf{PI}}$ to $\tilde{\mathbf{x}}^{\mathsf{I}}$, the coordinates are in the pixel coordinate frame.

$$\tilde{\mathbf{x}}^{\mathsf{P}} = \mathbf{P}^{\mathsf{PI}}\tilde{\mathbf{x}}^{\mathsf{I}} \tag{3.4}$$

$$\mathbf{P}^{\mathsf{PI}} = \begin{bmatrix} u_{max} & & \\ & v_{max} & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \tan\frac{\mu}{2} \\ & 1 & r_{v/u}\tan\frac{\mu}{2} \\ & & 1 \end{bmatrix} \tag{3.5}$$

Both matrices are finally combined to the camera-to-pixel matrix $\mathbf{K}^{\mathsf{PC}}$.

$$\mathbf{K}^{\mathsf{PC}} = \mathbf{P}^{\mathsf{PI}}\mathbf{K}^{\mathsf{IC}} \tag{3.6}$$

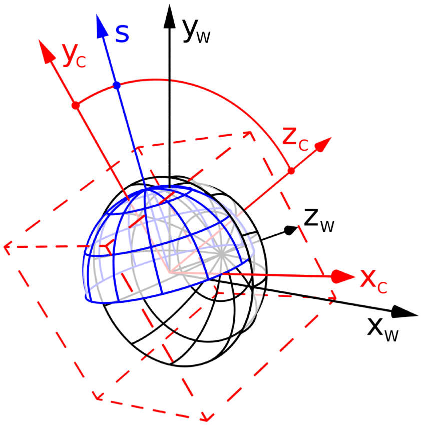


Figure 3.2: Camera orientation.

3.2.2 Outer parameters

The camera coordinate frame C preserves angles and scale of the world coordinate frame W. So C must result from *rotation* $\mathbf{R}^{\mathsf{WC}}$ and *translation* $\mathbf{t}^{\mathsf{WC}}$ of W. Both, rotation and translation, can be zero.

The translation vector $\mathbf{t}^{\mathsf{WC}}$ is equal to the cameras *location*. Its orientation is indirectly defined by a location $\mathbf{l}^{\mathsf{W}}$, the camera looks at, and a sky vector $\mathbf{s}^{\mathsf{W}}$ defining its virtual horizon. The *principal axis* $\mathbf{v}^{\mathsf{W}}$ (the viewing direction) of the camera is the difference of $\mathbf{l}^{\mathsf{W}}$ and $\mathbf{t}^{\mathsf{WC}}$.

The columns of the *rotation matrix* $\mathbf{R}^{\mathsf{WC}}$ are the axes of the cameras coordinate frame. Each axis length is 1. The axes can be derived from the vectors $\mathbf{s}^{\mathsf{W}}$ and $\mathbf{v}^{\mathsf{W}}$ due to orthogonality.

Let $\operatorname{norm} : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the normalization operator, scaling a vector to unit length.

$$\operatorname{norm}(\mathbf{x}) = \frac{\mathbf{x}}{|\mathbf{x}|}$$

Then the z-axis of C is the principal axis scaled to unit length.

$$\mathbf{z}_{\mathsf{C}}^{\mathsf{W}} = \operatorname{norm}\left(\mathbf{v}^{\mathsf{W}}\right) \tag{3.7}$$

The x-axis of C is per definition perpendicular to the z-axis and the sky vector $\mathbf{s}^{\mathsf{W}}$. Therefore it is the vector product of the sky vector $\mathbf{s}^{\mathsf{W}}$ and the

z-axis $\mathbf{z}_{\mathsf{C}}^{\mathsf{W}}$ scaled to unit length.

$$\mathbf{x}_{\mathsf{C}}^{\mathsf{W}} = \text{norm}\left(\mathbf{s}^{\mathsf{W}} \times \mathbf{z}_{\mathsf{C}}^{\mathsf{W}}\right) \tag{3.8}$$

The y-axis of C is per definition perpendicular to the z-axis $\mathbf{z}_{\mathsf{C}}^{\mathsf{W}}$ and the x-axis $\mathbf{x}_{\mathsf{C}}^{\mathsf{W}}$. Such it is the vector product of z-axis $\mathbf{z}_{\mathsf{C}}^{\mathsf{W}}$ and x-axis $\mathbf{x}_{\mathsf{C}}^{\mathsf{W}}$, already having unit length.

$$\mathbf{y}_{\mathsf{C}}^{\mathsf{W}} = \text{norm}\left(\mathbf{z}_{\mathsf{C}}^{\mathsf{W}} \times \mathbf{x}_{\mathsf{C}}^{\mathsf{W}}\right) \tag{3.9}$$

In order to get the coordinates of a fixed 3d-location in camera coordinates having world coordinates, the transformations defining the cameras location and orientation have to be applied inversely. For translation this is easily done by negation.

$$\mathbf{t}^{\mathsf{CW}} = -\mathbf{t}^{\mathsf{WC}} \tag{3.10}$$

For a rotation matrix the *orthogonality condition* guarantees that the inverse of the matrix is equal to its transpose.

$$\mathbf{R}^{\mathsf{CW}} = \left(\mathbf{R}^{\mathsf{WC}}\right)^{-1} = \left(\mathbf{R}^{\mathsf{WC}}\right)^{\mathsf{T}} \tag{3.11}$$

The 4×3 homogeneous matrix $\mathbf{P}^{\mathsf{PW}}$, transferring a homogeneous 4-vector in W onto a homogeneous 3-vector in P comes from consecutively applying the aforementioned transformations.

$$\tilde{\mathbf{x}}^{\mathsf{P}} = \mathbf{P}^{\mathsf{PW}}\tilde{\mathbf{x}}^{\mathsf{W}} \tag{3.12}$$

$$\mathbf{P}^{\mathsf{PW}} = \left[\mathbf{K}^{\mathsf{PC}}|\mathbf{0}\right]\begin{bmatrix}\mathbf{R}^{\mathsf{CW}} & \mathbf{R}^{\mathsf{CW}}\mathbf{t}^{\mathsf{CW}} \\ 0 & 1\end{bmatrix} \tag{3.13}$$

3.3 Region detectors

In the following, we give a short introduction in how the inspected region of interest detectors work, what kind of regions they preferably detect and what their covariance properties are. For further details, we reference the authors of each method.

3.3.1 Difference of Gaussian (DoG) detector

A finite scale space $L(\sigma, u, v)$ is built from an intensity image $I(u, v)$ by convolving it with Gaussian kernels $G(\sigma, u, v)$ with increasing σ. Neighbours over the finite scale domain σ are separated by a constant factor k.

$$\begin{aligned} L(\sigma, u, v) &= G(\sigma, u, v) * I(u, v) \\ G(\sigma, u, v) &= \frac{1}{2\pi\sigma^2} e^{\left(u^2+v^2\right)/2\sigma^2} \\ \sigma_{i+1} &= k\sigma_i \end{aligned} \tag{3.14}$$

The difference of two adjacent scales $D(\sigma, u, v) = L(k\sigma, u, v) - L(\sigma, u, v)$ is a constantly scaled approximation of the scale normalised Laplacian of Gaussian $\sigma^2 \nabla^2 G(\sigma, u, v)$ (Lindeberg, 1994). Extrema of the Laplacian—and so the Laplacian of Gaussian as well—denote locations, where the signal changes significantly in at least one direction. The characteristic size of a local structure is indicated by a local extremum over scale of the scale normalised Laplacian (Mikolajczyk and Schmid, 2004). Such a local extremum D at $\mathbf{x} = (\sigma, u, v)^\mathsf{T}$ is either greater or lower than each of its 26 neighbours $D(\sigma_{i+a}, u+b, v+c)$ with $a, b, c = \{1, 0, -1\}$. Each detected extremum $\hat{\mathbf{x}}$ is located with subpixel-accuracy by interpolation through fitting a 3d-quadric function to $D(\mathbf{x})$ and its local neighbourhood. The interpolated value $D(\hat{\mathbf{x}})$ is used to reject detections with low contrast.

The Difference of Gaussian—and so the Laplacian of Gaussian as well—have strong responses alongside edges. These detections are rejected by evaluating its *cornerness* using the approach by Harris and Stephens (1988). The eigenvalues α and β of the *Hessian matrix* $\mathbf{H}(\hat{\mathbf{x}})$ of $D(\hat{\mathbf{x}})$ are proportional to the *principal curvatures* of $D(\hat{\mathbf{x}})$. Homogeneous areas are indicated by two low principal curvatures, edges by one high and one low curvature and corners by two high curvatures. Because the detection process responds on edges and corners only, it is sufficient to estimate the ratio r of the two eigenvalues $\alpha = r\beta$, which is related to the trace $\mathrm{Tr}(\mathbf{H})$ and the determinant $\mathrm{Det}(\mathbf{H})$.

$$\mathbf{H} = \begin{bmatrix} D_{uu} & D_{uv} \\ D_{uv} & D_{vv} \end{bmatrix} \tag{3.15}$$

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} = \frac{(r+1)^2}{r} \tag{3.16}$$

The image gradients in a region around $\hat{\mathbf{x}}$ are gathered in an orientation histogram. Each value is weighted by the gradients amplitude and a Gaussian window with $\sigma_I = 1.5\hat{\sigma}$. The dominant orientation is then interpolated from the highest histogram bin and its neighbours. If there is more than one dominant orientation, the detection is duplicated for each of them.

The selected regions are blobs in characteristic scales. The detection process is covariant to translation, rotation and isotropic scale. Furthermore it is robust against significant arbitrary transformations and illumination changes. Due to the covariance restriction to similarity transformations, an extracted region descriptor for one of these detections can not be invariant to projective transformations introduced by large viewpoint changes.

For further details, see Lowe (2004).

3.3.2 Harris- and Hessian-Affine detectors

Both approaches are quite similar and differ in the initial keypoint selection technique only. The Harris-Affine detector uses the *second moment matrix* **M**, the Hessian-Affine detector the Hessian matrix **H**. Locations with a local maximum of both of the matrix' principal curvatures are detected over a finite differentiation scale space σ_D with σ_I being an integration scale slightly greater than σ_D for noise reduction.

$$\begin{aligned}\mathbf{M} &= M(\sigma_I, \sigma_D, u, v) \\ &= \sigma_D^2 G(\sigma_I, u, v) * \begin{bmatrix} I_u^2(\sigma_D, u, v) & I_u I_v(\sigma_D, u, v) \\ I_u I_v(\sigma_D, u, v) & I_v^2(\sigma_D, u, v) \end{bmatrix} \qquad (3.17)\\ \mathbf{H} &= H(\sigma_D, u, v) = \begin{bmatrix} I_{uu}(\sigma_D, u, v) & I_{uv}(\sigma_D, u, v) \\ I_{uv}(\sigma_D, u, v) & I_{vv}(\sigma_D, u, v) \end{bmatrix} \qquad (3.18)\end{aligned}$$

Both approaches detect one local structure over continuous scales. The characteristic scale of the structure is identified by an extremum of the Laplacian over scale. That is, for the Hessian-Affine detector, keypoint detection and scale selection are almost equivalent to those of the DoG detector in reversed order.

The affine shape of each keypoint is defined by a transformation, that projects the anisotropic pattern of the points neighbourhood to an isotropic one, having equal eigenvalues. 'This transformation is given by the square root of the second moment matrix $\mathbf{M}^{1/2}$' (Mikolajczyk et al., 2005, page 50), up to a rotation factor. The rotation factor can be identified with a method similar to the orientation assignment of the DoG detector.

The Harris-Affine detector detects regions on local structures being corners, junctions and spots at a characteristic scale. The Hessian-Affine detector detects blobs at characteristic scales. Both detectors are covariant to translation, rotation and anisotropic scale, thus being invariant to significant viewpoint changes. One disadvantage of the detection of regions around corners is the increased probability to handle regions covering discontinuities in 3d-space (see Section 4.2 for a detailed discussion). Especially the Harris-Affine detector's applicability is affected by this drawback.

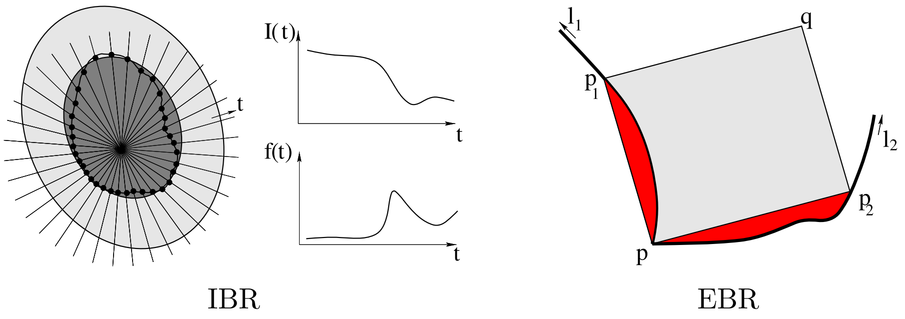
For further details, see Mikolajczyk and Schmid (2004).

3.3.3 Maximally Stable Extremal Regions (MSER)

Binary thresholding an intensity image means, that all pixels with an intensity below the threshold are marked as black and those above or equal as white. Black regions represent local minima, white regions local maxima, both are *extremal regions*. A *threshold space* is built by thresholding the image at all available intensities. For a typical 8-bit-grey-value image this is $S = \{0, \ldots, 255\}$. Depending on whether starting from the lowest or the highest threshold, black or white regions appear and grow when traversing through the space alongside the threshold axis. In each thresholded image, the extremal regions are marked as connected components, thus building a nested stack for each of the components. At some thresholds two or more components join to one. These thresholds and those where the shape of the component changes significantly are unstable and poorly located, especially in case of intensity changes and introduction of noise. The *maximally stable extremal regions* are extracted from the stacked components at a threshold representing a local minimum in the difference of the component's area at its predecessor and its successor in threshold space. Predecessor and successor do not have to be adjacent to the inspected threshold, the differentiation width is a parameter of the method. Increasing this parameter decreases the number of detected regions by rejecting less stable ones.

The selected regions are intensity maxima or minima with significant boundary contrast of arbitrary shape covering a wide variety of region size. They are invariant to affine intensity changes and covariant to adjacency preserving transformations, including rotation, scale change and perspective transformation. The algorithm necessarily fails on significantly smoothed images.

For further details, see Matas et al. (2002, Section 2).



IBR

EBR

Figure 3.3: Construction of the IBR and EBR detectors (Tuytelaars and Van Gool, 2004).

3.3.4 Intensity based region (IBR) detector

The detection method begins with detecting local intensity maxima and minima in an intensity image. To avoid the influence of noise, the image is priorly smoothed. Starting from each extremal point, a number of rays is emitted equally covering all directions $D = \{0°, \ldots, 360°\}$. Alongside each ray, a significant change of the pixel intensities is located by finding a local maximum evaluating the function

$$f(t) = \frac{\operatorname{abs}\left(I(t) - I(0)\right)}{\max\left(\int\limits_0^t \operatorname{abs}\left(I(t) - I(0)\right) dt/t, d\right)} \tag{3.19}$$

with $I(t)$ being the pixel intensity at position t, and d being a small number to prevent division by zero. The maxima of all rays are linked to each other and an ellipse is fitted to the resulting shape. Like depicted in Figure 3.3, the area of this ellipse is finally doubled.

The selected regions are intensity extrema comparable to those detected by the MSER detector. In contrast to those, the IBRs do not necessarily have a high contrast boundary. Thus, the IBR detector is less sensitive to smoothing.

It must fail, if scale change introduces additional local extrema inside the prior detection. Thus, the variety of detectable region sizes is significantly lower than those of the MSER, limiting the scale covariance. Nevertheless, the detector is invariant to illumination changes and covariant to affine transformations.

For further details, see Tuytelaars and Van Gool (2004, Section 5).

3.3.5 Edge based region (EBR) detector

This *geometry-based* method relies on the presence of corners and edges. Starting from a Harris corner point $\mathbf{p} = (u_p, v_p)^\mathsf{T}$ (Harris and Stephens, 1988), two points $\mathbf{p}_1$ and $\mathbf{p}_2$ are sent out along the two nearby Canny edges (Canny, 1986). Moving $\mathbf{p}_1$ with constant speed, at each time t its actual position on the edge is defined by an arbitrary curve-parameter s_1. For each $\mathbf{p}_1(s_1)$, $\mathbf{p}_2(s_2)$ is uniquely defined by equalizing the relative invariant parameters l_1 and l_2.

$$\left.\begin{aligned} l_i &= \int\limits_0^{s_i} \operatorname{abs}\left(|\mathbf{A}_i|\right) ds_i \\ \mathbf{A}_i &= \begin{bmatrix} \mathbf{p}'_i(s_i) & \mathbf{p} - \mathbf{p}_i(s_i) \end{bmatrix} \end{aligned}\right\} i = 1, 2 \tag{3.20}$$

That is, the area between the edges and the lines $\langle \mathbf{p}, \mathbf{p}_1(l_1 = l_2) \rangle$ and $\langle \mathbf{p}, \mathbf{p}_2(l_1 = l_2) \rangle$ are equal, like depicted in Figure 3.3. This criterion is invariant to affine transformations. The points $\mathbf{p}$, $\mathbf{p}_1(l_1 = l_2)$ and $\mathbf{p}_2(l_1 = l_2)$ define a parallelogram $\Omega(s_1, s_2) = \Omega(l_1 = l_2)$. In a finite range of $(l_1 = l_2)$, the photometric functions $f_1(\Omega)$ and $f_2(\Omega)$ are evaluated.[1] $\mathbf{p}_1$ and $\mathbf{p}_2$ are stopped at a point where one of these functions runs through an extremum.

$$f_1(\Omega) = \operatorname{abs}\left(\frac{|\mathbf{p}_1 - \mathbf{p}_g \quad \mathbf{p}_2 - \mathbf{p}_g|}{|\mathbf{p} - \mathbf{p}_1 \quad \mathbf{p} - \mathbf{p}_2|} \right) \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - \left(M_{00}^1\right)^2}} \tag{3.21}$$

$$f_2(\Omega) = \operatorname{abs}\left(\frac{|\mathbf{p} - \mathbf{p}_g \quad \mathbf{q} - \mathbf{p}_g|}{|\mathbf{p} - \mathbf{p}_1 \quad \mathbf{p} - \mathbf{p}_2|} \right) \frac{M_{00}^1}{\sqrt{M_{00}^2 M_{00}^0 - \left(M_{00}^1\right)^2}} \tag{3.22}$$

$$\text{with} \quad M_{pq}^n = \int_\Omega I^n(u, v) u^p v^q du dv \quad \text{and} \quad \mathbf{p}_g = \left(\frac{M_{10}^1}{M_{00}^1}, \frac{M_{10}^1}{M_{00}^1} \right)^\mathsf{T}$$

In the case of straight lines, s_1 and s_2 are decoupled because l_1 and l_2 are zero for each t. It is therefore necessary, to evaluate f_1 and f_2 in $\Omega(s_1, s_2)$ for two independent parameters instead of one. $\mathbf{p}_1$ and $\mathbf{p}_2$ are stopped, where both, $f_1(\Omega)$ and $f_2(\Omega)$, run through a minimum.

The detector selects parallelogram regions at corners and the adjacent edges. Ideally, these parallelograms are covariant to rotation, shear and anisotropic scale. Scale covariance is limited by the restriction of the search

[1] The authors of the method, Tuytelaars and Van Gool (2004), give an overview and detailed description of the used photometric functions.

space for s_1 and s_2 and local photometric properties in the parallelogram. The detector is invariant to changing intensity offset.

For further details, see Tuytelaars and Van Gool (2004, Section 5).

3.4 Repeatability

The repeatability score r_{ij} is the ratio of the number of re-detected points or regions (correspondences) to the minimal number of detections in image I_i or image I_j, being visible in both images. That is, the repeatability score is in the range $0 \leq r_{ij} \leq 1$. A repeatability score $r_{ij} = 0$ means, that there where no repeatable detections, $r_{ij} = 1$ means, that all possible detections were detected repeatably. The comparison to the minimal number of common visible detections is necessary, because coarser scaled images can not represent the fine local structures of finer scaled images.

For interest point detectors, a correspondence is determined by the locations of the points. An interest point $\mathbf{x}_i$ in image $\mathbf{I}_i$ is considered to be re-detected in image $\mathbf{I}_j$ if there is exactly one interest point $\mathbf{x}_j$ detected in image $\mathbf{I}_j$ within a distance $|\mathbf{x}_j - \mathbf{x}_i^j| < \epsilon$ from the transferred location $\mathbf{x}_i^j = \mathbf{x}_i \rightarrow \mathbf{I}_j$.

$$r_{ij}(\epsilon) = \frac{|C_{ij}(\epsilon)|}{\min\left(|P_i|, |P_j|\right)} \tag{3.23}$$

with $C_{ij}(\epsilon)$ being the set of correspondences with respect to localisation error ϵ and P_i and P_j being the sets of interest point detections in the commonly visible parts of images I_i and I_j.

For region detectors, the area covered by a region has to be taken into account. A region $\mathbf{r}_i$ in image $\mathbf{I}_i$ is considered to be re-detected in image $\mathbf{I}_j$ if there is exactly one region $\mathbf{r}_j$ detected in image $\mathbf{I}_j$ that overlaps more than a minimal *overlapping rate* $1 - \epsilon_O$ with the transferred region $\mathbf{r}_i \rightarrow \mathbf{I}_j$. Then ϵ_O itself is the *overlap error*.

$$r_{ij}(\epsilon_O) = \frac{|C_{ij}(\epsilon_O)|}{\min\left(|R_i|, |R_j|\right)} \tag{3.24}$$

with $C_{ij}(\epsilon_O)$ being the set of correspondences with respect to overlap error ϵ_O and R_i and R_j being the sets of interest point detections in the commonly visible parts of images I_i and I_j.

A high repeatability score is a necessary criterion for a detectors applicability in recognition tasks.

Chapter 4

Method

This work is focused on the evaluation of low level region of interest detectors for changing viewpoint in non-planar scenes. Mikolajczyk et al. (2005) estimated the repeatability score of affine-covariant region detectors for viewpoint changes on images of planar scenes. Fraundorfer and Bischof (2005) extended this analysis to non-planar scenes, implicitly assuming that there is no occlusion. We extend their evaluation method with regard to the influence of occlusion. We expect our results to be more accurately corresponding to what the implication of the repeatability evaluation is—the applicability of a detector for extracting region descriptors, which are invariant to a set of transformations albeit distinctive. We expect an inverse correlation of a detector's tendency to detect regions covering discontinuous edges and its repeatability score. Taking occlusion into account will decrease the potential overlapping of transferred regions. Compared to the results of Fraundorfer and Bischof (2005), we expect to estimate lower repeatability scores for detectors covering more discontinuous edges. We substantiate the correlation by estimating the *discontinuity ratio* as a measure for this tendency for each detector. We discuss this in Section 4.2.

Mikolajczyk et al. (2005) argue that larger region sizes increase the overlapping and thus the repeatability. They showed that this holds in planar scenes. We expect a contrary effect in non planar scenes due to the rising influence of occlusion and displacement alongside discontinuous edges with increasing region size. We discuss this in Section 4.5.

Like Fraundorfer and Bischof (2005), we use the implementations of the Harris-Affine, Hessian-Affine, IBR, EBR and MSER detectors provided by Mikolajczyk et al. (2005). As mentioned in Chapter 2, Fraundorfer and Bischof also evaluated a set of interest point detectors including the Difference of Gaussian (DoG) detector by Lowe (2004). Beeing a similar-covariant detector, the DoG provides scale invariance, and scale can not be expressed by a point. Thus—to deal with it as a point detector using a fixed maximal localisation error ϵ, violates the correlation of the evaluated property and the potentially extracted description vector for detected locations. While accepting two detections of completely different sizes at nearly the same location, good correspondences at larger scales, located very close with respect to its scale but too far with respect to pixel size, are rejected. Furthermore, adjacent locations in image $\mathbf{I}_i$ are not necessarily adjacent in image $\mathbf{I}_j$ in non-planar scenes due to the influence of depth discontinuities (see Section 4.2). Therefore, we handle the detector as a region detector similar to the affine-covariant detectors, testing the region overlap.

In their test framework, Fraundorfer and Bischof (2005) used sequences of photographs from known camera positions and acquire ground truth for pixel locations in the first image from stereo matching the first two images of a sequence. While stereo matching provides very good results for continuous textured surfaces, it introduces errors through mismatching at discontinuities and significant large areas without matches at homogeneous regions. In their evaluation, the authors reject those regions which cover such 'empty' areas by more than 50% of their size. Regions covering these areas by less than 50% are marked for manual inspection.

Instead of photos and disparity maps from stereo matching, we use artificially generated image sequences of three detailed POV-Ray scenes by the artists Tran and Piqueres[1]. Per pixel ground truth is acquired by ray-casting from the camera's origin through each pixel, thus acquiring the 3d-location and surface normal at the rays point of intersection with the closest object surface. The rendered results appear 'photo-realistic' to human observers, including direct and indirect illumination, shadows, planar and non-planar surfaces and textures, reflection and refraction. All three scenes show complex non-planar scenes of every day environments. The images are obtained using the pinhole camera model, such performing only linear transformations throughout the projection. Therefore, the rendered images are perfectly rectified without any geometrical distortions (like barrel, pincushion or wave distortion) and static photometric disturbance (like shading and chromatic aberration). In practice, a vision system can remove such disturbances from an image prior to the application of the feature detector. It is straightfor-

[1] The scene files are publicly available in the Internet at `http://www.oyonale.com/ressources/english/sources16.htm` and `http://www.ignorancia.org/en/index.php?page=Sources_map`

ward to rectify the images sufficiently through appropriate static calibration. Using a homography (see Schmid et al., 2000; Mikolajczyk et al., 2005) or the trifocal tensor (see Fraundorfer and Bischof, 2005) also implies the pinhole camera model by preserving linearity.

In fact, in the same way as the authors of all three evaluation frameworks do, we assume that the captured pixel intensity directly results from the closest rigid surface and, that all surfaces are Lambertian, such having constant radiance for all directions. This assumption does, of course, not hold for transparent, glossy or mirroring surfaces. So, the introduction of such materials, like in photographs of every day environments as well as in the used artificial scenes, will degrade the quality of intensity based low level feature detectors, like the examined.

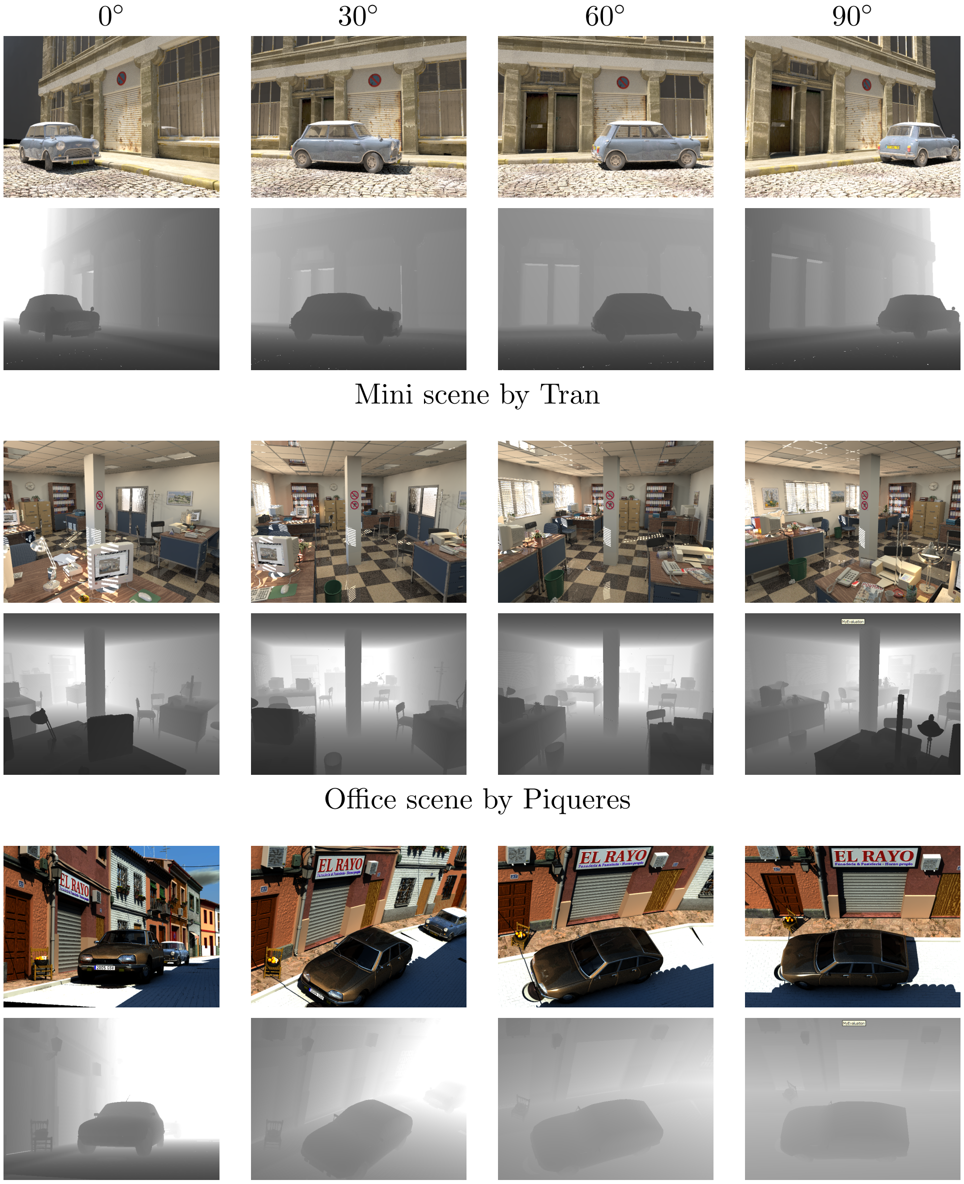
4.1 Data Set

The test framework consists of a sequence of 10 images per scene, each representing a progressive viewpoint change from 0° up to 90°. The camera is moved alongside a circle around an axis through the point the camera looks at in 10°-steps. In the Mini and Office scene this axis is parallel to $(0, 1, 0)^\mathsf{T}$, in the Town street scene it is a skew vector parallel to $(-1, 0.5, 0)^\mathsf{T}$ in world coordinates. The sky vectors in the Mini and Office scene are static and equal to the y-axis of the world coordinate frame. So the cameras orientation is changed only by panning and tilting. In the Town street scene, the sky vector progressively rotates about the worlds z-axis for 45°, such adding rolling to the orientation change.

For each image, there is a ground truth map file, that contains the unique camera definition and 3d-location and surface normal in world coordinates for each pixel of the image. Our images—and so the ground truth maps as well—have a size of 640×480px.

4.2 Discontinuities in non-planar scenes

The camera's location and orientation define a plane through the camera centre, containing the x- and y-axis of the camera coordinate frame and being perpendicular to the cameras principal axis—the camera's *infinity plane*. This plane divides the world 3-space into three parts, the part in front of the camera (positive z), the part behind the camera (negative z) and the plane itself ($z = 0$). Disregarding the finite window and resolution at the viewing plane, defined by the inner parameters of the camera, 3d-



Mini scene by Tran

Office scene by Piqueres

Town street scene by Piqueres

Figure 4.1: Data set. Images and ground truth maps. Each scene is rendered from a rotating camera changing the viewing direction from 0° up to 90°. All three sequences consist of 10 images. For each image, a ray-casted 3d-map provides ground truth per pixel.

locations in front of the camera are potentially visible—locations behind the camera or at the infinity plane are not.

As mentioned above, we consider continuous surfaces of rigid objects with Lambertian reflectance only and assume any other effects to be noise. Ideally, the intensity value of each surface-location is independent from the cameras location and orientation.

Each infinitesimal area of an object's surface in front of the camera is potentially visible, if the angle of its normal vector and the line connecting it with the cameras centre is smaller than 90°. Such a surface area is facing the camera.

Extracting locations or regions of interest from intensity patterns implies the assumption, that at least the topology of the pattern is preserved during projection. The intensity pattern on the viewing plane is therefore assumed to be the mapping of a 2d-manifold—a continuous surface—in the 3d-world, steadily facing the camera.

If the normal vectors of all infinitesimal areas of a surface are equal, the surface itself is planar. Then this projective mapping preserves lines to be lines and therefore the topology as well. The latter also holds for non planar surfaces, if—and only if—each of its infinitesimal areas faces the camera. If the surface partially does not face the camera, the continuity of the mapped intensity pattern is broken and therefore no longer topologically equivalent to the surface.

A general 3d-scene consists of a set of objects of arbitrary shape. This introduces another class of discontinuities—the transition from one object to another. Because each objects surface represents a 2d-manifold itself, the presence of more than one on the 2d-viewing plane necessarily requires discontinuities. Closer objects occlude distant ones.

The perspective mapping induces *parallaxes* when changing the viewpoint through shifting the camera. The relative movement of the mapped 3d-location directly depends on the inverse of its distance to the camera's infinity plane, the z-component in camera coordinates. Moving a camera C on its (x, y)-plane for $(a, b)^{\mathsf{T}}$ and assuming $\mu = 90°$ and $r_{v/u} = 1$, which only removes an additional constant scale factor, the homogeneous image-coordinates of a 3d-location in the cameras former coordinate frame are

$$\tilde{\mathbf{x}}^{\mathsf{I}'} = \begin{bmatrix} 1 & & & -a \\ & 1 & & -b \\ & & 1 & 0 \end{bmatrix} \tilde{\mathbf{x}}^{\mathsf{C}} = \begin{pmatrix} x - a \\ y - b \\ z \end{pmatrix} \tag{4.1}$$

Two adjacent pixels p and q at a discontinuous edge, p representing the foreground and q the background, are thus shifted for different distances

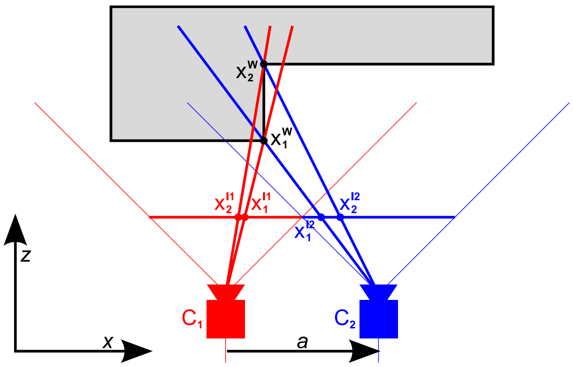


Figure 4.2: Parallaxes in non-planar scenes. The relative speed of the 3d-locations $\mathbf{x}_1^{\mathsf{W}}$ and $\mathbf{x}_2^{\mathsf{W}}$ mapped to a moving camera's image frame I depends on their distance to the camera. $\mathbf{x}_1^{\mathsf{W}}$ is occluded for camera location C_1 while being visible for C_2.

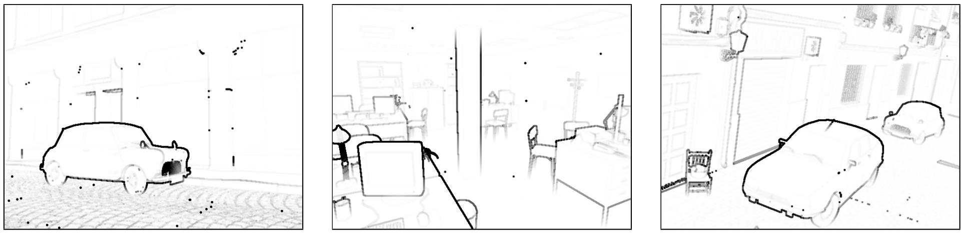
and can not preserve their neighbourhood. Possibly q disappears due to occlusion, like depicted in Figure 4.2 and Figure 4.4.

Discontinuities represent the outline of objects and/or break the mapping of the intensity pattern of an object's surface. Therefore they tend to be significant edges or corners in the image intensity pattern as well. A low level region detector based on intensities can not distinguish between edges or corners coming from surface patterns and those coming from discontinuities. Especially the detection of 2d-intensity-changing-speed extrema, like in the DoG, Harris-Affine and Hessian-Affine detectors, will lead to a significant amount of detections at discontinuities. The extraction of a description vector from an intensity pattern around these locations, possibly normalized to the expressed transformation (similarity or affinity), can not be invariant to viewpoint changes due to displacement of the pattern as a result of parallaxes.

In the pixel coordinate frame P, the transferred z-component—and so its inverse as well—is represented as a function of the pixel coordinates $\frac{1}{z} = f(u, v)$. For each pixel, a *discontinuity score* $d_{(u,v)^\intercal}$ is defined by the magnitude of the second gradient of f in u and v.

$$d_{(u,v)^\intercal} = \sqrt{\left(\frac{\partial^2 f}{\partial u^2}\right)^2 + \left(\frac{\partial^2 f}{\partial v^2}\right)^2} \tag{4.2}$$

The discontinuity score of a pixel correlates with the potential displacement of the intensity pattern in the pixel's neighbourhood when shifting the



Mini scene　　Office scene　　Town street scene

Figure 4.3: Second gradients of $\frac{1}{z} = f(u,v)$ represent the strength of discontinuities. Each in the 3rd image of the sequence, the magnitude of f is represented by grey-values, normalised to the images average, scaled by $i = 5$ and dilated for visualisation purposes. White represents $f = 0$, black $f \geq \frac{1}{5}$.

camera. That is, a region containing pixels with lower discontinuity scores is expected to be less influenced by potential displacement through camera movement than a region containing pixels with higher scores. A descriptor extracted from a region less influenced by displacement is expected to be more robust to viewpoint changes. Therefore we estimate a region's discontinuity score d_{P_i} being the average of the discontinuity scores of all pixels inside the region P_i. A detector's tendency to extract regions on discontinuous edges is expressed by the average of the discontinuity score of all detected regions $\{P_1, \ldots, P_n\}$ relative to the discontinuity score of the whole image, denoted as *discontinuity ratio* r_d.

$$r_d = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} d_{P_i}}{\frac{1}{u_{max}v_{max}}\sum\limits_{v=0}^{u_{max}-1}\sum\limits_{v=0}^{v_{max}-1} d_{(u,v)^\intercal}} \tag{4.3}$$

If $r_d > 1$, the regions are predominantly located on discontinuous edges, if $r_d < 1$, predominantly on continuous surfaces. We estimate r_d for each detector from all available images per scene for scale factors from $s = 0.25$ up to $s = 2$ applied to all region sizes. We assume r_d of region detectors preferring continuous surfaces to decrease for lower scale factors, while increasing for detectors preferring discontinuous edges. There exists a maximum value for each detector's curve at a detector specific scale. Beyond this scale value, r_d converges to 1 due to the correlation of the scaled region size to the image's size.

4.3 Visibility of 3d-locations

For the repeatability evaluation, it is necessary to check, if a 3d-location, corresponding to a detection in image $\mathbf{I}_i$, is potentially visible in image $\mathbf{I}_j$. Using a homography in case of planar scenes or the trifocal tensor for stereo pairs, one can find out, if the location is inside or outside the finite pixel frame of image $\mathbf{I}_j$. In a planar scene, a location inside the image is necessarily visible. In a non planar scene it can be visible as well or invisible due to occlusion. If ground truth is available for image $\mathbf{I}_i$ only, it is not possible to reliably state about occlusion in image $\mathbf{I}_j$. In our data set, there is ground truth for each pixel of each image, and such the detection of per pixel occlusion is possible, as follows.

1. Transfer the 3d-location $\mathbf{x}^{\mathsf{W}}$ corresponding to the pixel $\mathbf{x}^{\mathsf{P}_i}$ in image $\mathbf{I}_i$ to the pixel reference frame P_j of image $\mathbf{I}_j$, thus getting the homogeneous 3-vector $\mathbf{x}^{\mathsf{P}_j} = (u_j, v_j, z_j)^{\mathsf{T}} = \mathbf{P}^{\mathsf{P}_j\mathsf{W}}\mathbf{x}^{\mathsf{W}}$.

2. If $z_j \leq 0$ or $\frac{u_j}{z_j} < 0$ or $\frac{v_j}{z_j} < 0$ or $\frac{u_j}{z_j} \geq u_{max}$ or $\frac{u_j}{z_j} \geq v_{max}$, the location is outside the image or behind the camera and therefore not visible.

3. If $\mathbf{x}^{\mathsf{P}_j}$ is inside the image, estimate the minimal and maximal z-value z_{min} and z_{max} of the four pixels next to it in image $\mathbf{I}_j$. If $z_j \geq 2z_{max} - z_{min}$, the location is occluded.

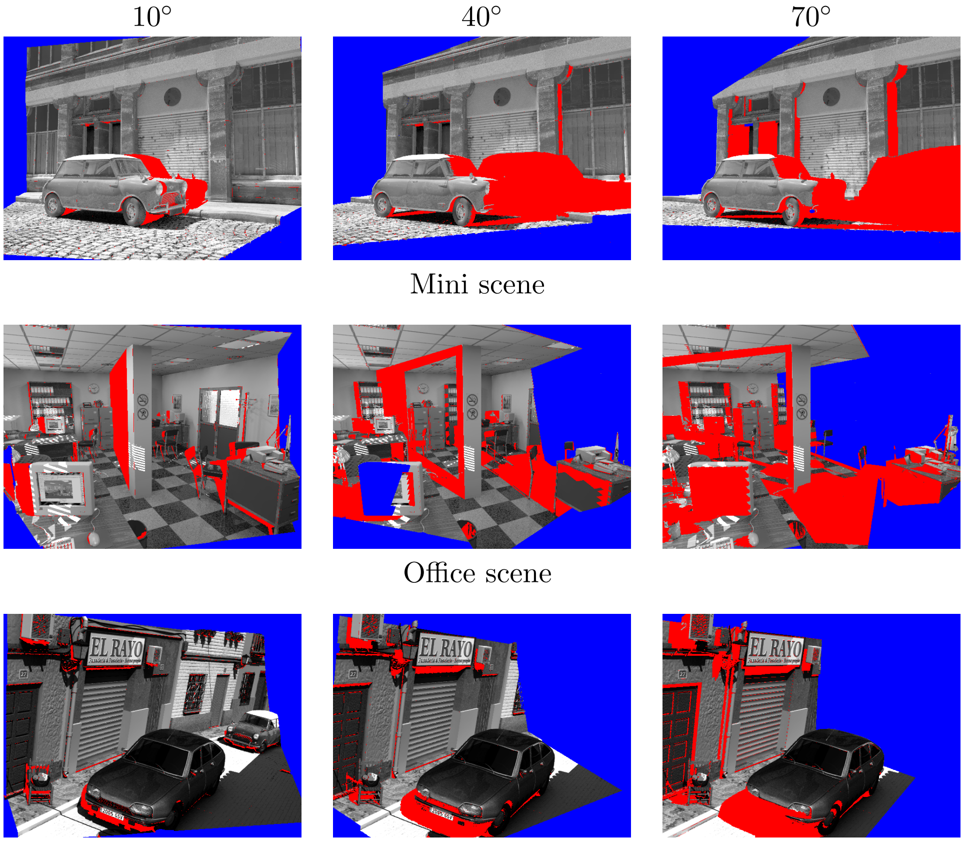
Inaccuracies from ray-casting the scene and coordinate transfer lead to false occluded-positives, especially at aslant surfaces, when checking the location for beeing behind z_{max} only. Therefore the slope-dependent tolerance is introduced.

4.4 Region overlap

Consistent with the work of Mikolajczyk et al. (2005), each detected region in an image's pixel frame is implicitly defined by an ellipse (u_0, v_0, a, b, c, s). A pixel coordinate (u, v) is inside the ellipse, iff

$$a(u - u_0)^2 + 2ab(u - u_0)(v - v_0) + c(v - v_0)^2 \leq s^2 \tag{4.4}$$

The 2d-pixel-location $(u_0, v_0)^{\mathsf{T}}$ is the center of the ellipse, a, b and c define its shape and size and s is an optional scaling factor. The regions detected by the DoG are exported as circular ellipses. We decided the circle's



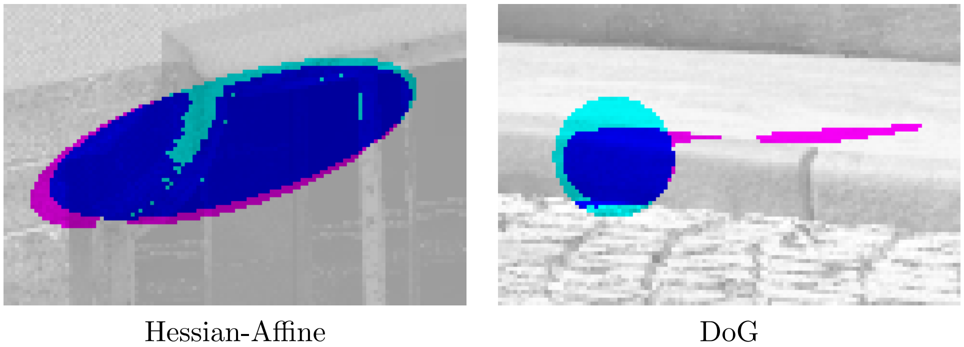
10° 40° 70°

Mini scene

Office scene

Town street scene

Figure 4.4: Visibility of transferred 3d-locations of the 3rd to the 4th, 7th and 10th camera of a sequence, such achieving a viewpoint change of 10°, 40° and 70°. Blue marks locations transferred to outside the camera's viewing frustrum, red marks locations occluded for the camera.



Hessian-Affine

DoG

Figure 4.5: Region transfer for arbitrary surfaces. Correspondence pairs for 40° viewpoint change. Cyan marks the ellipse present in this image, magenta the corresponding ellipse transferred from the other image.

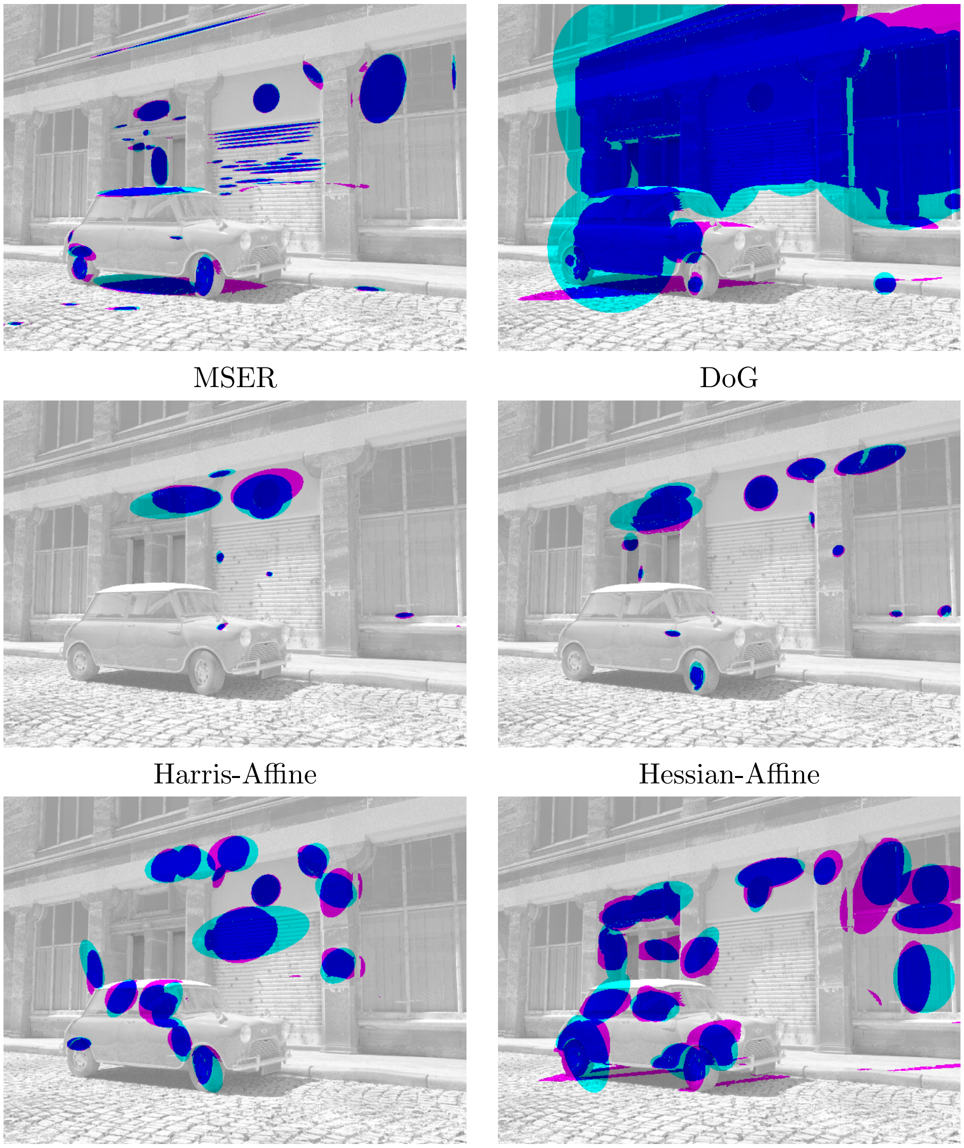
radius to be $r_{\sigma=1} = \frac{32}{\sqrt{\pi}}$px for the original scale. This corresponds to the area of the potentially extracted SIFT-descriptor window.

All regions, detected in image $\mathbf{I}_i$ are tested to be visible in image $\mathbf{I}_j$ and vice versa, thus getting the sets R_i and R_j, containing only the possible canditates for re-detection. This is done numerically by transferring each pixel's 3d-location covered by the region to the other image and checking its visibility. If more than 50% of the region's pixel coordinates are visible, the region is counted to be visible in both images.

For each pair of potentially corresponding regions, the overlap error ϵ_O is estimated from the finite set of pixel coordinates covering the regions, both seen in image $\mathbf{I}_i$.

$$\epsilon_O = 1 - \frac{|P_i \cap P'_j|}{|P_i \cup P'_j|} \tag{4.5}$$

where P_i is the set of all pixel locations inside the region detected in image $\mathbf{I}_i$. P'_j is the set of all pixel locations inside the transferred shape corresponding to the region detected in image $\mathbf{I}_j$. Transferring a visible elliptical region from image $\mathbf{I}_i$ to image $\mathbf{I}_j$ leads to arbitrary shapes, possibly wide spread small subregions all over the image (see Figure 4.5). Therefore all pixel locations in image $\mathbf{I}_i$ have to be tested. A location belongs to the region in image $\mathbf{I}_j$, if it is visible—inside the pixel frame and not occluded—and inside the ellipse. Like Fraundorfer and Bischof, we count a pair of regions to be a true correspondence, if their overlap error is smaller than 50% and if there is no other region overlapping one of both candidates more than 50%.



MSER

DoG

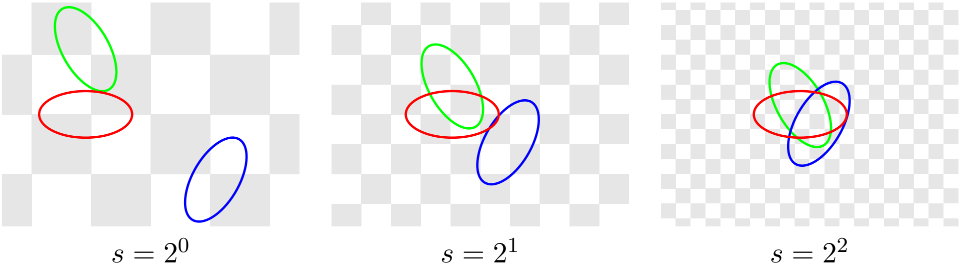
Harris-Affine

Hessian-Affine

IBR

EBR

Figure 4.6: Region correspondences in the Mini scene for 40° viewpoint change, detected by the different detectors. Cyan marks the ellipses present in this image, magenta the corresponding ellipses transferred from the other image. MSER provides very accurate re-detections. DoG detects a very large number of regions up to very large scales spread all over the image.



$s = 2^0$ $s = 2^1$ $s = 2^2$

Figure 4.7: In a planar scene, the overlap error for each region to each other decreases with increasing scale factor s. The grid visualises the corresponding reference frame.

4.5 The effect of region size

Covariance to scale requires the detection of regions of arbitrary size. The projected size of a local structure changes continuously, depending on the angle of view of the camera or its distance to the structure. The output of a scale covariant detector should therefore show a continuous histogram over a wide range of region sizes, theoretically bounded by the images inner and outer scale. The inner scale of an image is its pixel size, its outer scale the image's finite dimensions (see Lindeberg, 1994, Section 4.2).

Mikolajczyk et al. (2005) discuss the correlation of the size of regions and their tendency to overlap each other in planar scenes. They argue larger regions to overlap better than smaller ones, due to the decreasing relative localisation error for increasing region size. Increasing the size of the regions increases the overlapping rate of all regions to each other by decreasing the influence of localisation errors (see Figure 4.7). Towards infinitely large size, the relative localisation error tends to zero and the overlapping test would only compare the shapes of the ellipses. The estimated repeatabillity score, plotted as a function of region scale will therefore ascend up to a maximum and then drop due to getting more than one correspondence per region. In the ascending part, a detector would benefit from detecting larger regions.

The authors avoid this by normalizing the region detected in image $\mathbf{I}_i$ to a fixed radius and scaling the transferred region detected in image $\mathbf{I}_j$ with the same scale factor. Such the size of both regions relative to each other is preserved while balancing the influence of localisation.

In non planar scenes the situation is slightly different. As a result of discontinuities, larger regions do not necessarily overlap better than smaller. Quite contrary, larger areas tend to contain more discontinuous edges than smaller ones, thus introducing more displacement (see Section 4.2 and Figure 4.5). We therefore expect, that increasing the region size will lead to lower repeatability scores.

Non planarity breaks the direct correlation of overlapping rate and scale of two candidates. Therefore a normalization to a fixed radius would violate the regions correspondence to the underlying scene and thus adulterate the results in a not intended manner. That is, we check the overlapping of two regions using the original dimensions. We verify our assumption through performing the experiments with all regions scaled by a factor $s_i = 2^i$ with $i \in \{-2, -1, 0, 1\}$.

4.6 Evaluation procedure

4.6.1 Region size

We plot the sizes of the regions detected by a detector in a 197-bin-histogram with logarithmic scale. Like Mikolajczyk et al. (2005), we estimate the size of a region being the geometric mean of the ellipse's both half-axes. This corresponds to the radius of a circle with the same area. For each histogram, the detected regions in all 30 images of the data set are used.

4.6.2 Repeatability

We estimate each detector's repeatability score as a function of viewpoint change by comparing the set of detected regions in each image to that in each other image of a sequence. The repeatability score for viewpoint change of a specific angle is then estimated by arithmetically averaging the scores of all image pairs expressing the same viewpoint change. Having more than one image pairs for a viewpoint change, besides averaging, we estimate the standard deviation of the population. Each scene is evaluated separately. The repeatability score is plotted as a function of viewpoint change.

To investigate the influence of scaling the region size (see Section 4.5), we perform the repeatability evaluation for all regions scaled by a factor $s_i = 2^i$ with $i \in \{-2, -1, 0, 1\}$. The repeatability score for representative viewpoint changes is plotted as a function of region scale.

4.6.3 Discontinuity ratio

We estimate each detector's discontinuity ratio r_d as a measure for potential displacement of detected regions under viewpoint change being its discontinuity score d_P in all images of the sequence normalised relative to the images discontinuity score d_I (see Section 4.2). We arithmetically average the discontinuity ratios from all images and estimate the standard deviation. Each scene is evaluated separately.

We verify our assumption regarding the influence of scaling all regions by repeating the discontinuity evaluation for all regions scaled by a factor $s_i = 2^i$ with $i \in \{-2, -\frac{3}{2}, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$. The discontinuity ratio is plotted as a function of region scale.

Chapter 5

Results

5.1 Region size

All detectors detect regions of arbitrary sizes. This is a necessary criterion for scale covariance. The range of detected sizes and its distribution differs due to the design of the detectors. The estimated histograms of detected region sizes are depicted in Figure 5.1.

The MSER detector shows a continuous distribution over a wide range of region sizes, from very small up to very large regions. There are more small regions detected than larger ones. This is a necessary consequence of the property, that the maximally stable extremal regions can not intersect. A finite image contains more small than large regions in general.

The DoG detector shows a wide range of detected region sizes as well. The frequency of the discrete scale space is clearly visible, showing a local maximum at the scale samples. Due to interpolation of the detections over image space and scale, the sizes are not limited to the discrete scale samples. A detected region's size depends on the initial scale of detection and consecutive interpolation only, the affine shape of the structure is not considered. Downscaling an image or its derivatives reduces the absolute number of extrema, therefore less large regions are detected than smaller ones.

The Harris- and Hessian-Affine detectors also rely on initial extrema detection over discrete scales, therefore more small than large regions are detected. For both detectors, the scale frequency is visible as well. In contrast to the DoG detector, the size of a detection is not interpolated over scale, but searched over discrete scale samples. The plot is thus not continuous.

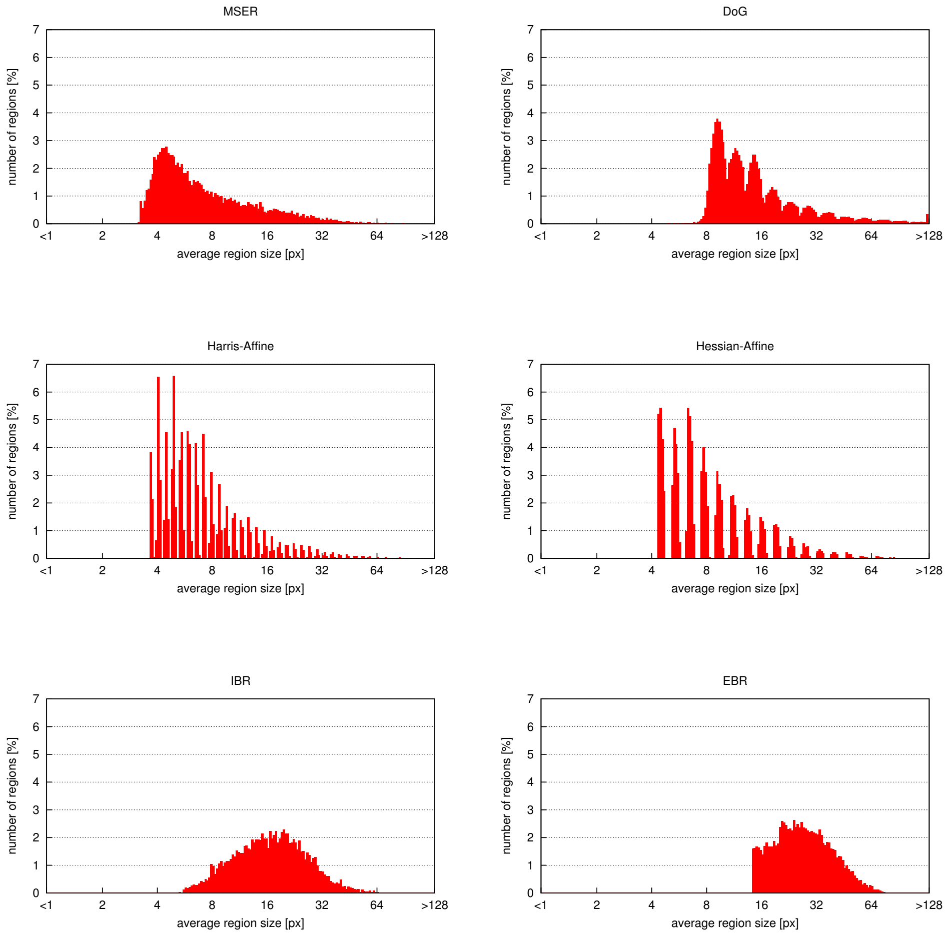


Figure 5.1: Histograms of the region sizes detected by the inspected detectors, averaged from all images of all scenes. The MSER and DoG detectors detect regions over a wide range of sizes. The Harris- and Hessian-Affine detectors prefer smaller regions while the IBR and EBR detectors detect regions of medium size within a limited range. Note the visible sampling frequency of the discrete scale space used by the scale space related detectors (DoG, Harris-Affine and Hessian-Affine).

The IBR detector relies on initial intensity extrema detection at a fixed scale. That is, the presence of finer structures is decreased, while coarser structures might not be found due to clutter. Furthermore, the boundaries of each detection are detected within a limited range of possible sizes. Nevertheless, the approach is robust and shows a continuous range of region sizes, thus providing scale covariance within this range. Note, that the range covers regions with a radius of 5px $\leq r \leq$ 64px, which implies a maximal scale factor $s \approx 12$ being quite sufficient for recognition tasks in camera images.

The EBR detector relies on the initial detection of corner points and edges detected at a fixed scale. Like for the IBR detector, the possible size of a region is limited by a fixed restriction of the search space. The plot is continuous within a range including a scale factor $s \leq 4$.

5.2 Repeatability

We evaluate the repeatability score of each detector for viewpoint change in the interval of $0°$ up to $90°$. We repeat the evaluation with scaled region sizes up to a factor $s_i = 2^i$ with $i \in \{-2, -1, 0, 1\}$.

5.2.1 Repeatability under viewpoint change

The number of correspondences decreases with increasing viewpoint change for all detectors. This is a consequence of two factors, firstly the decreasing absolute number of valid detections due to the smaller commonly visible area, and secondly the decreasing repeatability due to the increasing geometric and photometric influence of projective transformation.

On planar Lambertian scenes, the repeatability score is expected to decrease continuously with increasing viewpoint change through insufficient covariance to the applied geometric transformations (Mikolajczyk and Schmid, 2003; Mikolajczyk et al., 2005). Non-planarity (Fraundorfer and Bischof, 2005) and non-Lambertian reflectance decrease the repeatability as well. The influence of non-planarity and non-Lambertian reflectance is expected to be different for changing viewpoint and different scenes, thus affecting the results non-continuously.

We see, that the repeatability score of all detectors decreases like expected for increasing viewpoint change. However, it does not necessarily decrease continuously (see Figure 5.2, the plot of the MSER-detector in the Mini scene).

Consent with Mikolajczyk et al. (2005) and contrary to the results of Fraundorfer and Bischof (2005), the MSER detector shows outstanding re-

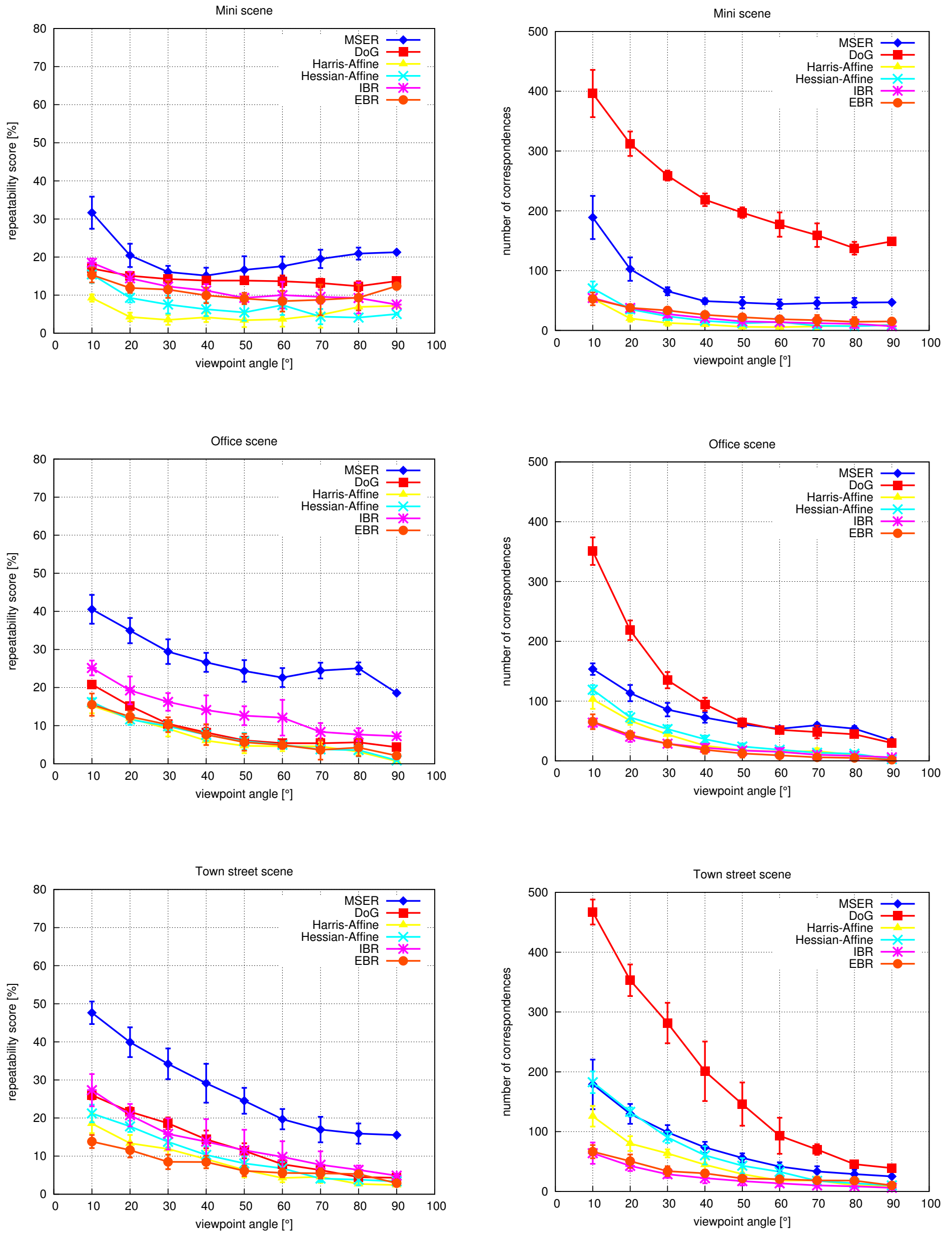


Figure 5.2: Repeatability score and absolute number of correspondences, both as a function of viewpoint change.

peatability scores compared to the other detectors for all scenes and for all viewpoint changes. It provides more than 150 correspondences for small viewpoint changes and still more than 30 correspondences for viewpoint changes of more than 50° in all scenes.

The repeatability score of the DoG detector evaluated from region overlapping instead of localisation error is significantly lower than in the evaluation of Fraundorfer and Bischof (2005), but still comparable to the affine covariant detectors. The DoG detector performs in most cases comparable to the IBR detector, which showed better results than the EBR, Hessian- and Harris-Affine detectors. It provides the highest absolute number of correspondences, from more than 300 for small viewpoint changes to more than 50 for viewpoint changes larger than 50° in all scenes. Only in the Office scene, it was slightly outperformed by the MSER detector for viewpoint changes of more than 70°.

The absolute number of correspondences detected by the EBR and IBR detectors drops below 20 for viewpoint changes larger than 50° in all scenes.

The Harris- and Hessian-Affine detectors show different results in different scenes. While both detectors perform bad in the Mini scene, the results for the Town street scene are better than for the EBR detector. Especially the absolute number of correspondences detected by the Hessian-Affine detector in this scene is comparable to the MSER detector. This is all the more surprising, because both scenes show a comparable environment and indeed the same car.

Our experiments show significantly lower repeatability scores than those of Fraundorfer and Bischof (2005) for the IBR and the Harris- and Hessian-Affine detectors. Due to the reliable ground truth, consideration of occlusion, reverse pixel-location transfer and inclusion of all image pairs representing the same viewpoint change, our results are more accurate. However—for the MSER detector, we estimated comparable results. We assume three drawbacks of their framework being liable for this discrepancy. Firstly, stereo matching implies discontinuous edges to be continuous while rejecting continuous homogeneous regions. Secondly, the transfer of pixel-locations from one image to the other leads to spotted shapes by integer approximation. Thirdly, the influence of occlusion is negated. All three points compromise the accuracy of the estimated overlapping rate.

5.2.2 The effect of region size

Especially for large viewpoint changes, all affine covariant detectors perform best for the originally detected size. Downscaling the region size increases

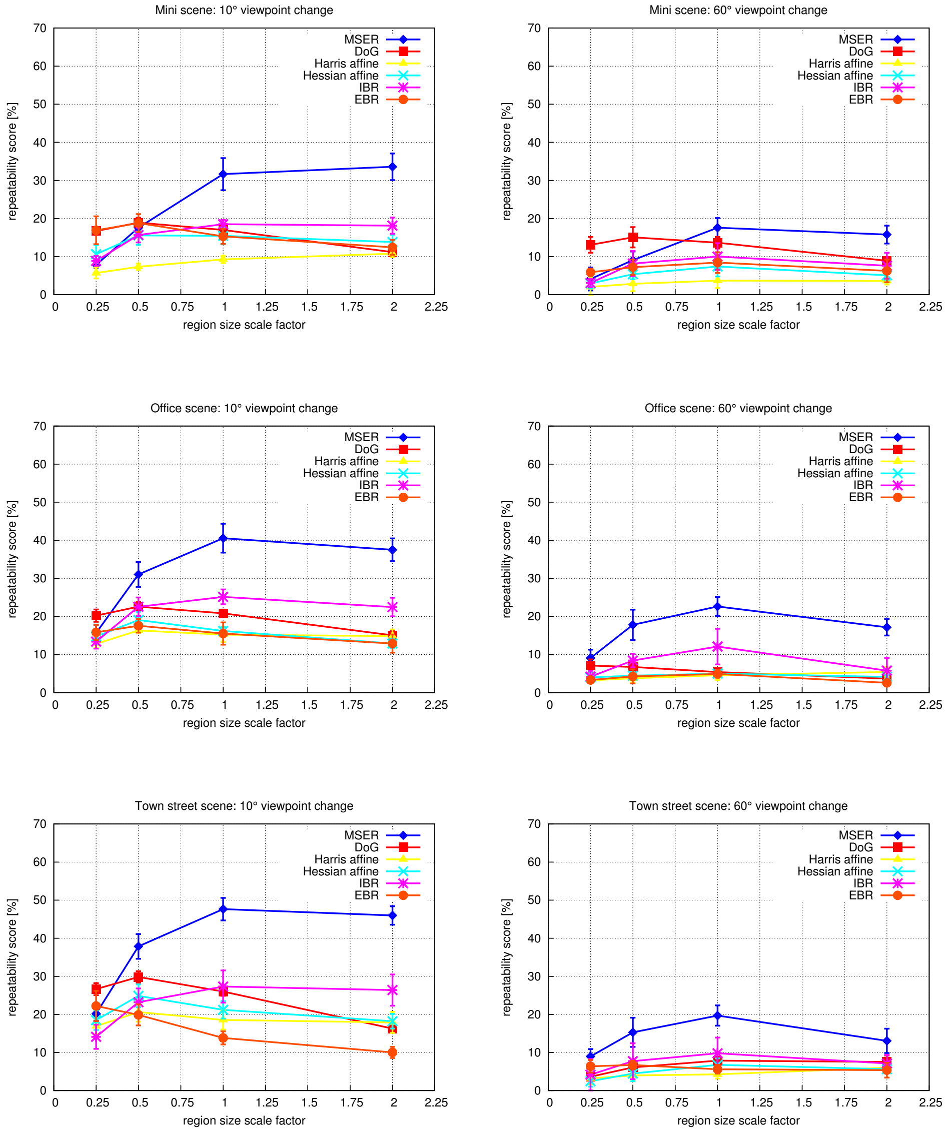


Figure 5.3: Repeatability score as a function of region scale, averaged from all image pairs with a viewpoint change of 10°and 60°.

the influence of potential localisation errors, while upscaling the size increases the potential influence of discontinuities and ambiguous correspondences. The intensity based detectors, MSER and IBR, show similar results for small viewpoint changes, while the corner based Harris-Affine detector seems to benefit from increasing the region size, but on a very low performance level.

The EBR detector shows better results for lower region scales and small viewpoint changes. The detected regions are in the neighbourhood of corners, that can result from discontinuities. Thus downscaling the size decreases the probability to cover these discontinuities at least partially.

The DoG detector shows a maximum at lower sizes, this is due to the relatively large initial size, we fixed for our experiments. Also the larger number of detections will lead to ambiguous overlapping for larger sizes. However, below the half of the original size the repeatability score decreases due to the influence of localisation errors comparable to the affine covariant detectors.

These results confirm our expectations. Decreasing the region size increases the influence of localisation error and decreases the repeatability score. In contrast to planar scenes, increasing the region size does not improve the repeatability automatically. The originally detected regions size shows the best results for most detectors, especially the intensity based detectors.

5.3 Discontinuity ratio

The discontinuity ratio highly depends on scene content. All detectors benefit from presence of continuous textured or structured surfaces with Lambertian reflectance.

The corner based Harris-Affine detector shows a significantly higher tendency to detect regions covering discontinuous edges than the other detectors. Downscaling the region size increases its discontinuity ratio, that is, a significant number of detections are centered on discontinuous edges. This observation matches well with the low repeatability scores of the Harris-Affine detector from our prior experiments.

Not surprisingly, the EBR detector shows very low rates. Tuytelaars and Van Gool (2004) designed it to address the problem of detections over discontinuities by selecting a region adjacent to a corner and two edges instead of the corner itself. Their approach seems to accomplish the task very well. Increasing the region size includes the corners and edges and

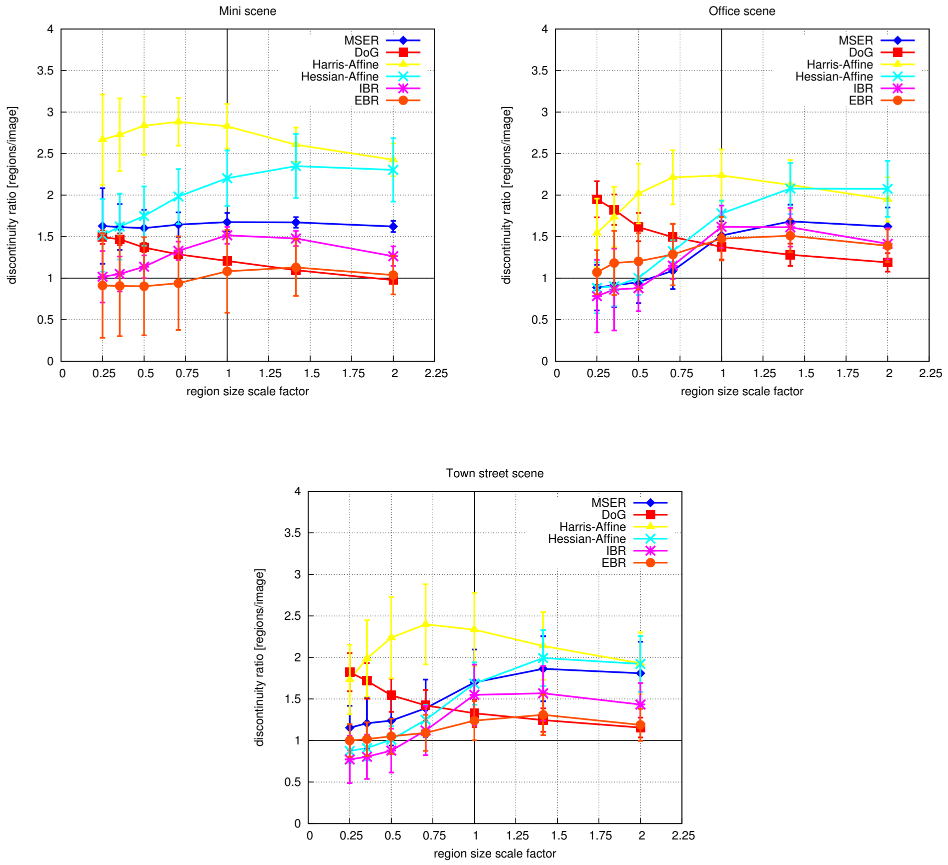


Figure 5.4: Discontinuity ratio as a function of region scale.

thus increases the discontinuity ratio as well. That is, the low repeatability score must be originated by unreliable initial detection or inaccurate shape estimation. We assume the latter, because the initial detection relies on the Harris corner point detector (Harris and Stephens, 1988), which was rated to be very reliable by Schmid et al. (2000).

The IBR detector shows lower rates than the MSER detector, which shows lower rates than the Hessian-Affine detector. All three detectors show comparable results in case of region scaling. Decreasing the size significantly decreases their discontinuity ratio, that is, Hessian blobs as well as intensity extrema are more probably located at continuous surfaces than on discontinuities.

The DoG detector benefits from providing a very large number of detections spread all over the image. It is due to the relatively large number and the relatively large size of the detected regions, that the plot shows only the descending part of the predicted progression (see Section 4.2).

All these observations clearly fulfill our expectations. Detections over discontinuities degrade the overlapping rate and thus the repeatability of a region detector. This matches well to the probability of extracting invariant region descriptors from the intensity pattern inside a region. The corner based Harris-Affine detector tends by design to detections on discontinuities. Scale space blobs (Dog and Hessian-Affine detector) and intensity extrema (IBR and MSER detector) might be originated by discontinuous corners, although these phenomena are much more probably located on continuous surfaces or at a specific distance to the discontinuity. The EBR detector is succesfully designed to avoid detections on discontinuities.

Chapter 6

Conclusion

We presented a revised evaluation framework for state of the art low level affine covariant region detectors under viewpoint change in non-planar scenes. By using artificially generated scenes and reliable ground truth, we addressed the drawbacks of the framework by Fraundorfer and Bischof (2005), namely the non-observance of depth-discontinuities and occlusion as well as inaccurate region overlapping estimation. Furthermore, we evaluated the similarity covariant DoG detector as a circular region detector, thus considering its scale invariance and the adjacency-dissolving influence of depth discontinuities in non-planar scenes.

All evaluated region detectors, except the MSER detector, showed significantly lower repeatability scores in our experiments, especially for large viewpoint changes. The MSER detector shows by far the best results, the IBR and DoG detectors performed better than the Harris- and Hessian-Affine and the EBR detector.

We showed, that the correlation of region size and repeatability score in planar scenes does not hold in non-planar scenes.

Confirming the observation of Fraundorfer and Bischof (2005), we were able to show experimentally, that the number of detections covering depth discontinuities differs for different detectors. We introduced the discontinuity ratio, being a measure for the tendency to detect regions covering depth continuities. A high discontinuity ratio decreases a detectors repeatability score by compromising the overlapping rate of the detections.

Bibliography

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 8(6), November 1986.

Christopher Cason, Thorsten Froehlich, Nathan Kopp, and Ron Parker. POV-Ray – the persistance of vision raytracer [web page], accessed August 8th 2006. URL `http://www.povray.org`.

Beman Dawes, David Abrahams, and Rene Rivera. Boost C++ libraries [web page], accessed January 29th 1999–2006. URL `http://www.boost.org`.

Friedrich Fraundorfer and Horst Bischof. A novel performance evaluation method of local detectors on non-planar scenes. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, Washington, DC, USA, 2005. IEEE Computer Society.

Chris Harris and Mike Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, Manchester, 1988.

Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, second edition, 2006. ISBN 0521 54051 8.

Timor Kadir, Andrew Zisserman, and Michael Brady. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision*, pages 404–416, 2004.

Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.

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

David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99: Proceedings of the International Conference on Computer Vision*, volume 2, pages 1150–1157, Washington, DC, USA, 1999. IEEE Computer Society.

Jiří Matas, Ondřej Chum, Martin Urban, and Tomáš Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In Paul L. Rosin and David Marshall, editors, *Proceedings of the British Machine Vision Conference*, volume 1, pages 384–393, London, UK, September 2002. BMVA.

Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, June 2003.

Krystian Mikolajczyk and Cordelia Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1): 63–86, 2004.

Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiří Matas, F. Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

Jaime Vives Piqueres. The persistance of ignorance [web page], accessed November 20th 2006. URL `http://www.ignorancia.org`.

Cordelia Schmid, Roger Mohr, and Christian Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37 (2):151–172, 2000. ISSN 0920-5691.

Stephen Se, David G. Lowe, and James J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *ICRA*, pages 2051–2058. IEEE, 2001. ISBN 0-7803-6578-X.

Gilles Tran. Oyonale [web page], accessed November 20th 2006. URL `http://www.oyonale.com`.

Tinne Tuytelaars and Luc Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59 (1):61–85, 2004.

Appendix A

Typical Transformations

Each linear transformation in Euclidean 3-space $\mathbb{R}^3$ can be expressed using homogeneous coordinates. The following matrices perform the basic transformations in a left-handed coordinate frame:

1. Scale:

$$\mathbf{S} = \begin{bmatrix} s_x & & \\ & s_y & \\ & & s_z \end{bmatrix} \tag{A.1}$$

2. Translate:

$$\tilde{\mathbf{T}} = \begin{bmatrix} 1 & & & t_x \\ & 1 & & t_y \\ & & 1 & t_z \end{bmatrix} \tag{A.2}$$

3. Rotate around axis x (tilt) by α:

$$\mathbf{R}_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \tag{A.3}$$

4. Rotate around axis y (pan) by β:

$$\mathbf{R}_y = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \tag{A.4}$$

5. Rotate around axis z (roll) by γ:

$$\mathbf{R}_z = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.5}$$

6. Perspective transformation from the coordinate frames origin with $(0, 0, 1)^{\mathsf{T}}$ being the *principal axis* and f beeing the *focal length*:

$$\tilde{\mathbf{P}} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix} \tag{A.6}$$

7. Perspective transformation from the coordinate frames origin with $(0, 0, 1)^{\mathsf{T}}$ being the *principal axis* and μ and ν beeing the *horizontal* and *vertical angle of view*:

$$\tilde{\mathbf{R}} = \begin{bmatrix} \frac{1}{\tan\frac{\mu}{2}} & & 0 \\ & \frac{1}{\tan\frac{\nu}{2}} & 0 \\ & & 1 \end{bmatrix} \tag{A.7}$$

Appendix B

File formats

Using the vector-to-string and vice versa conversion of the lexical_cast-library (see Dawes et al., 1999–2006), the ground truth maps are thus presented in the following ASCII-file format:

[3](16,1,-6)	location of the camera $\mathbf{t}^{\mathsf{WC}}$
[3](8.7,1.8,0)	look at $\mathbf{l}^{\mathsf{W}}$
1.0	
0.75	vertical over horizontal frame size ratio $r_{v/u}$
[3](0,1,0)	sky vector $\mathbf{s}^{\mathsf{W}}$
55	horizontal angle of view μ
640	horizontal pixel resolution u_{max}
480	vertical pixel resolution v_{max}
[3](12.39,0.024,-5.19)	pixel 3d-location $\mathbf{p}^{\mathsf{W}}_{(0,0)}$
[3](-0.15,0.97,-0.20)	pixel surface normal $\mathbf{n}^{\mathsf{W}}_{(0,0)}$
...	from the lower left $(0,0)$ to the upper right $(u_{max}-1, v_{max}-1)$ pixel of the image

The implementations of the affine region detectors provided by Mikolajczyk et al. (2005), and so our implementation of the DoG detector as well, generate a file containing the ellipse parameters in the following format:

1.0	scale factor s
723	number of detected regions $\|R\|$
22.7␣313.6␣0.001␣0.0004␣0.014	u_0, v_0, a, b, c
...	

Note, that these parameters imply a right handed pixel reference frame, having its origin in the upper left corner. We therefore have to flip the coordinates during import.

$$v_0^{\mathsf{P}} = v_{max}^{\mathsf{P}} - v^{\mathsf{P_r}} \qquad\qquad b = -b_r \tag{B.1}$$