MSCC: Maximally Stable Corner Clusters[*]

Friedrich Fraundorfer, Martin Winter, and Horst Bischof

Institute for Computer Graphics and Vision,
Graz University of Technology,
Inffeldgasse 16/2, A-8010 Graz, Austria
{fraunfri, winter, bischof}@icg.tu-graz.ac.at
http://www.icg.tu-graz.ac.at

Abstract. A novel distinguished region detector, complementary to existing approaches like Harris-corner detectors, Difference of Gaussian detectors (DoG) or Maximally Stable Extremal Regions (MSER) is proposed. The basic idea is to find distinguished regions by clusters of interest points. In order to determine the number of clusters we use the concept of maximal stableness across scale. Therefore, the detected regions are called: Maximally Stable Corner Clusters (MSCC). In addition to the detector, we propose a novel joint orientation histogram (JOH) descriptor ideally suited for regions detected by the MSCC detector. The descriptor is based on the 2D joint occurrence histograms of orientations. We perform a comparative detector and descriptor analysis based on the recently proposed framework of Mikolajczyk and Schmid, we present evaluation results on additional non-planar scenes and we evaluate the benefits of combining different detectors.

1 Introduction

Recently, there has been a considerable interest in using local image region detectors and descriptors for wide base-line stereo [1, 2], video retrieval and indexing [3, 4], object recognition [5], and categorization tasks [6, 7]. There exist two main categories of distinguished region detectors. Corner based detectors like Harris [8], Harris-Laplace [4], Harris-Affine [9] etc. and region based detectors such as Maximally Stable Extremal Regions (MSER) [1], Difference of Gaussian points (DoG) [5] or scale space blobs [10]. In addition Brown et al. proposed to use groups of interest points [11]. Corner based detectors locate points of interest at regions which contain a considerable amount of image structure, but they fail at uniform regions and regions with smooth transitions. Region based detectors deliver blob like structures of uniform regions but highly structured regions are

[*] This work is funded by the Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04, by a grant from Federal Ministry for Education, Science and Culture of Austria under the CONEX program and partially supported by the European Union Network of Excellence, MUSCLE under contract FP6-507752.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 45–54, 2005.
© Springer-Verlag Berlin Heidelberg 2005

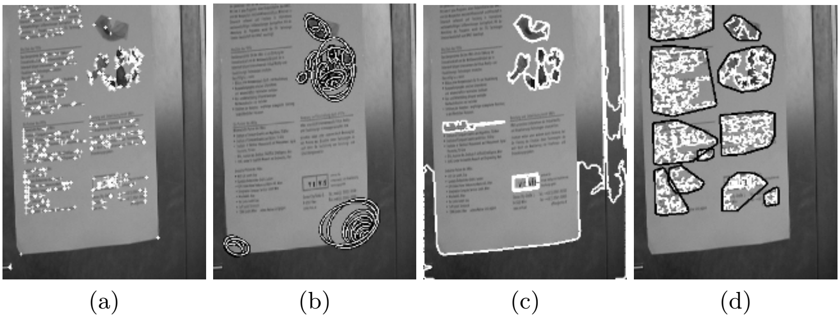


Fig. 1. (a) Harris corner detector. (b) Hessian-Affine detector. (c) MSER detector. (d) MSCC detector (one scale only)

not detected. As the two categories act quite complementary it is no surprise that people have started to use detector combinations (e.g. Video Google [3]). The main benefit in combining detectors with complementary properties is the increasing number of detected regions and thus possible matches. A second benefit is that the different detectors will fire in different regions of the image. Thus the image will be more uniformly covered with detected regions, which in turn improves the accuracy of e.g. the wide base-line stereo. In general, it is expected that the robustness of most algorithms will improve by combining different detectors. The accuracy in geometry estimation will be improved if more matches can be used and if the matches are distributed over the whole scene. In object recognition a better coverage of the object increases the robustness against partial occlusions. The available corner and region based detectors cover already a broad variety of image content. However, there are images not sufficiently covered by neither class of detectors. For example, Fig. 1 shows an image from a database we are using for studying visual robot localization tasks. The image shows a door with an attached poster. The poster contains a lot of text. Fig. 1(a) shows as an example detected Harris corners. The text in the image results in a high number of Harris corners with a very similar appearance which will result in a lot of mismatches. Fig. 1(b) shows the results of an Hessian-Affine detector. The detector selects strong corners and constructs local affine frames around the corners on multiple scales. This leads to unfortunate detections as one can see in the lower right part of the image. The MSER detector (see Fig. 1(c)) perfectly detects the homogeneous regions but ignores the parts of the image containing the text. If the resolution of the image would be higher, the detector would detect the individual letters, which in this case would be useless for matching because there are lots of similar letters. This simple example demonstrates that neither of the available detectors delivers satisfactory results.

The dense but locally distinct patterns of detected corners in textured image parts resulting from the Harris detector (Fig. 1(a)) suggests a new distinguished region detector based on characteristic groups of individual detections. In particular we propose to cluster the responses of individual detectors. The thus clustered regions are our distinguished regions. Fig. 1(d) shows some of the obtained regions (only at a single scale). It is clearly seen that we detect also regions where other detectors have problems. The remainder of the paper is organized as follows: In Section 2 we present the MSCC detector. Section 3 is devoted to the novel joint orientation histogram descriptor. The detector and descriptor evaluations are presented in Section 4. A discussion and outlook concludes the paper.

2 The MSCC Detector

The MSCC detector aims to benefit from the high repeatability of simple interest point detectors (as shown in [12]). The main idea is to use point constellations instead of single interest points. Point constellations are more robust against viewpoint changes than single interest points because a few missing single points will not affect the detection of the constellation itself. Point constellations are detected by clustering of the interest points for multiple scales. Selecting only those clusters which fulfill a stability criteria leads to robust and highly repeatable detections.

In particular the MSCC algorithm proceeds along the following three steps:

1. Detect simple, single interest points all over the image, e.g. Harris-corners.
2. Cluster the interest points by graph-based point clustering using a minimal spanning tree (MST) for multiple scales.
3. Select clusters which are stable across several scales.

It should be noted, that the steps 2 and 3 of the algorithm can be implemented very efficiently as it is possible to cluster and perform the cluster selection already during the MST construction.

2.1 Interest Point Detection and MST Calculation

To detect the interest points acting as cluster primitives we make use of the structure tensor [13]. For every image pixel we evaluate the structure tensor and calculate a Harris-corner strength measure (cornerness) as given by Harris and Stephens [8]. We select a large number of corners (all local maxima above the noise level) as our corner primitives. This ensures that we are not dependent on a cornerness threshold. The interest points itself represent the nodes of an undirected weighted graph in 2D. The MST computation is performed by applying Kruskal's iterative growing algorithm on the Delaunay-triangulation of the nodes [14]. The weight for the edge between two graph nodes is their geometric distance to which we will also refer to as edge length.

2.2 Multi Scale Clustering

Since we do not know the number of clusters we have to use a non-parametric clustering method. The method we use is inspired by the MSER detector. Given a threshold T on the edge length we can get a subdivision of the MST into subtrees

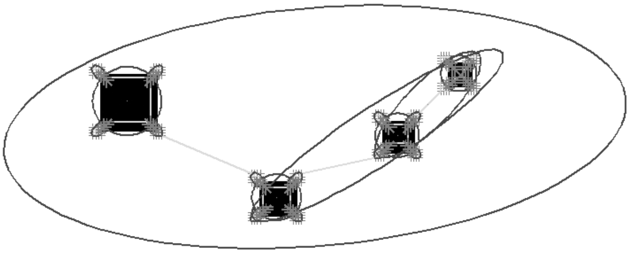


Fig. 2. Example of the MSCC detector on a synthetic test image (clustered interest points are indicated by ellipses around them)

by removing all edges with an edge length higher than this threshold. Every subtree corresponds to a cluster and gives rise to an image region. Different values for T produce different subdivisions of the MST, i.e. different point clusters. To create a multi scale clustering we compute subdivisions of the MST for a certain number of thresholds $T_1 ... T_p$ between the minimal and maximal edge length occurring in the MST.

We are now interested in clusters which are stable over several scales, i.e. have the same interest points. Fig. 2 illustrates the method on a synthetic test image. The image shows 4 differently sized squares. The Harris corner detection step produces several responses on the corners of the squares. Connecting the single points with the MST reveals a structure where one can easily see that clustering can be done by removing the larger edges. Clusters of interest points are indicated by ellipses around them. The test image shows the capability of detecting stable clusters at multiple scales, starting from very small clusters at the corners of the squares itself up to the cluster containing all detected interest points.

Unlike many other detectors the MSCC-clusters show arbitrary shapes, an approximative delineation may be obtained by convex hull construction or fitting ellipses. Fig. 4(c) shows examples for the convex hull and the fitted ellipses of detected regions in a 3D plot. One can see, that ellipse fitting is only a poor estimation of region delineation and will also introduce some errors for area and overlap calculation. However we will propose a descriptor without the need for region delineation which uses the clustered points directly.

3 The Joint Orientation Histogram Descriptor

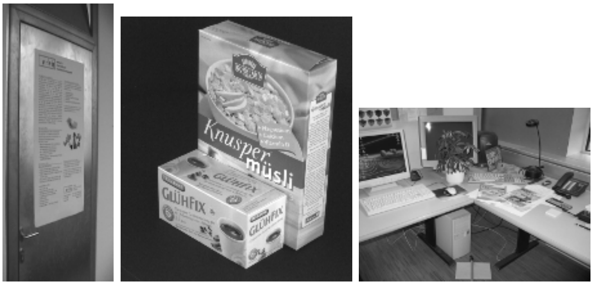
Orientation information has already been successfully used by Lowe in his SIFT-descriptor [5]. To make an optimal use of the MSCC detector results we have designed a new descriptor also based on orientation information, the so called joint orientation histograms (JOH).

As an estimate of the local gradient orientation we use Gaussian derivatives. In order to obtain rotational invariance the mean corner orientation within a region is computed. All corner orientations are normalized with respect to this mean orientation. The basic idea is that for each corner c within a MSCC region the joint orientation occurrence to its n local neighbors c_t weighted by the gradient magnitude is entered in a 2D histogram. That is, the gradient magnitude of c is added to the bin defined by the orientation of c and c_t. The histogram is smoothed in order to avoid boundary effects and normalized to sum to one. Best results are obtained for a typical histogram size of 8×8 and $n = 40$ nearest neighbors resulting in a 64-dimensional descriptor. The histogram takes into account the local structure within the region (e.g. are there many similar orientations or are the orientations uniformly distributed). In contrast to most other descriptors we do not depend on the definition of a center point or a region (ellipse) where the descriptor is calculated from. Therefore, even if the detected clusters differ in shape, this does not affect the descriptor as long as a sufficiently high number of points are re-detected. Matching (see 4.4) is done by nearest neighbor search using the Bhattacharyya distance [15].

4 Experimental Results

4.1 Detector Evaluation: Repeatability and Matching Score

To compare the performance of the MSCC detector with other approaches we use the publicly available evaluation framework of Mikolajczyk [16]. The evaluation framework gives two performance measures, a repeatability score and a matching score. The repeatability score is the relative number of repetitive detected interest regions. The matching score is the relative number of correctly matched regions compared to the number of detected regions. The correctness of the matches is verified automatically using a ground truth transformation. A homography is calculated between two (planar) images which allows to transfer the position of an interest region from the first to the second image. For the evaluation we use 10 images from a robot localization experiment. Fig. 3(a) shows an image of the image set and Fig. 4(a) shows an image with detected MSCC regions. To comply with the evaluation framework ellipses are fitted to the MSCC regions, i.e the ellipse parameters are calculated from the covariance



(a) (b) (c)

Fig. 3. (a) Test scene "doors". (b) Test scene "group". (c) Test scene "room"

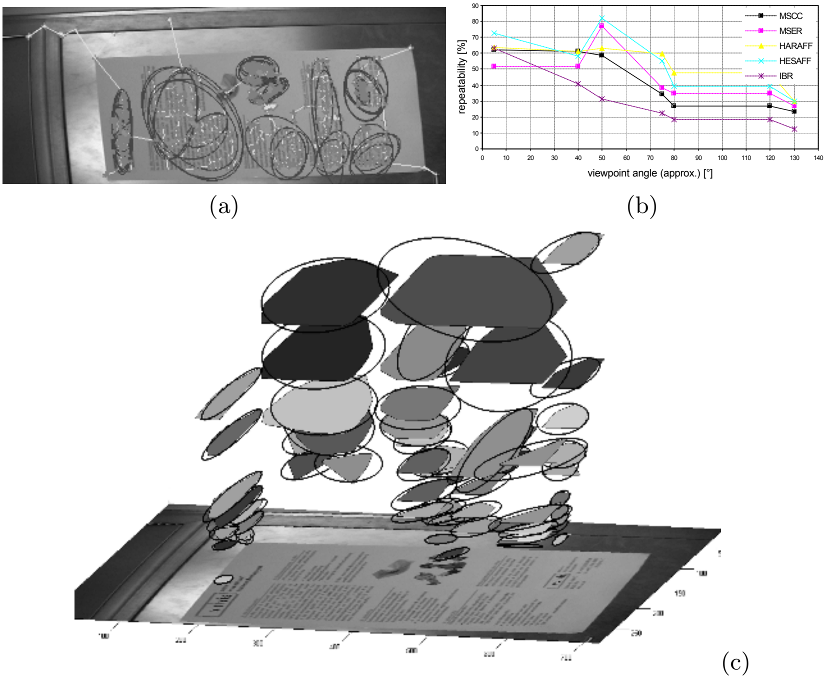
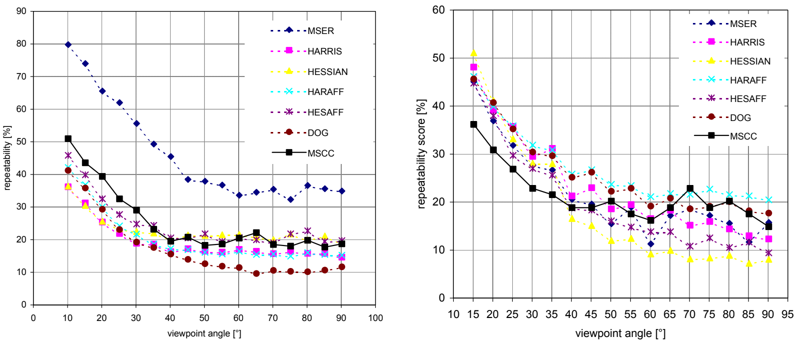


Fig. 4. (a) Example for detected MSCC regions. (b) Repeatability score for "doors" scene. (c) Convex hulls and fitted ellipses for detected MSCC regions

matrix of the interest points belonging to the region. We compare the repeatability score and the matching score of our MSCC detector to 4 other detectors on increasing viewpoint change up to 130°. For the matching score the SIFT descriptor is used. Fig. 4(b) shows the results of the MSCC detector compared to the Maximally Stable Extremal Regions (MSER) [1], the Hessian-Affine regions (HESAFF) [9], the Harris-Affine regions (HARAFF) [9] and the intensity based regions (IBR) [2]. The experiment reveals a competitive performance of our novel detector when compared to other approaches. The regions detected by our approach are consistently different from those of other detectors (see also 4.3).

4.2 Detector Evaluation on Non-planar Scenes

Non-planar scenes can not be evaluated with the method proposed by Mikolajczyk. We use the method proposed in [17] to evaluate the MSCC detector on non-planar scenes. We compare the results to 6 other detectors on increasing viewpoint angle additionally including the Difference of Gaussian keypoints (DOG) [5] and simple interest point detectors like Harris corners (HARRIS) and Hessian corners (HESSIAN) [8]. The compared value is the repeatability score. We use the publicly available implementation from Mikolajczyk for the other de-



(a) (b)

Fig. 5. (a) Repeatability score for ”group” scene. (b) Repeatability score for ”room” scene

tectors. We evaluated the detectors on 2 different complex scenes (see Fig. 3(b) and (c)). The test scene ”group” shows two boxes acquired with a turntable. The second test scene ”room” shows a part of an office and is of higher complexity than the first one. Both image sequences consist of 19 images and the viewpoint varies from 0° to 90°. Fig. 5(a) shows the repeatability score for the ”group” scene. The best performance is obtained by the MSER detector. The proposed MSCC detector comes second and shows a repeatability score noticeable higher than the other detectors. Fig. 5(b) shows the evaluation results for the ”room” scene. In this scene the performance of the different detectors is very similar and no one shows a really outstanding performance. The evaluations demonstrate that our MSCC detector is competitive to the other established detectors.

4.3 Combining Local Detectors

This experiment evaluates the complementarity of the MSCC detector. This is done by counting the non-overlapping correct matching regions from different detectors. Regions from different detectors are counted as non-overlapping if they do not overlap more than 40%. Matching is done using SIFT descriptors and nearest neighbor search (as implemented in Mikolajczyks evaluation framework). Fig. 6(a) shows the absolute number of matched MSER regions, MSER regions combined with HESAFF regions, combination of MSER, HESAFF and HARAFF, combination of MSER, HESAFF, HARAFF and IBR and combination of the previous detectors with the MSCC detector. Fig. 6(b)-(e) show the region numbers for combining the MSCC detector with each of the other detectors. The graphs show that our MSCC detector is able to add a significant amount of new matches to the ones of the other detectors. Fig. 6(f) and (g) show an example for 120° viewpoint change. The dashed dark ellipses mark the matches from the combination of MSER, HESAFF, HARAFF and IBR. The bright ellipses mark the additional matches obtained from the MSCC detector.

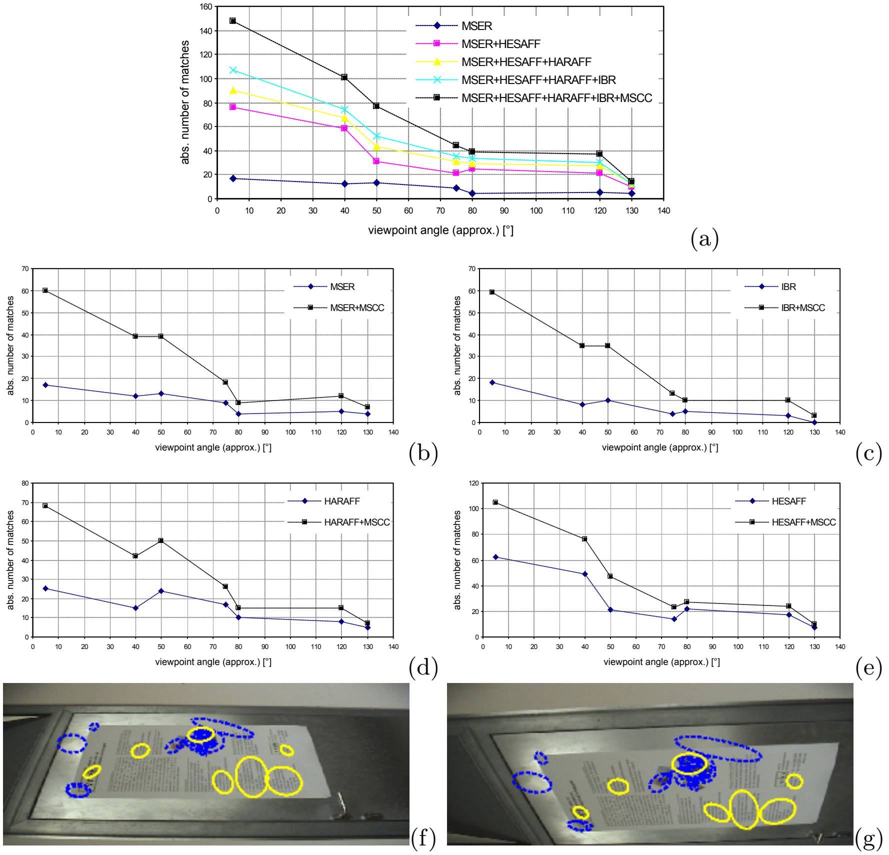
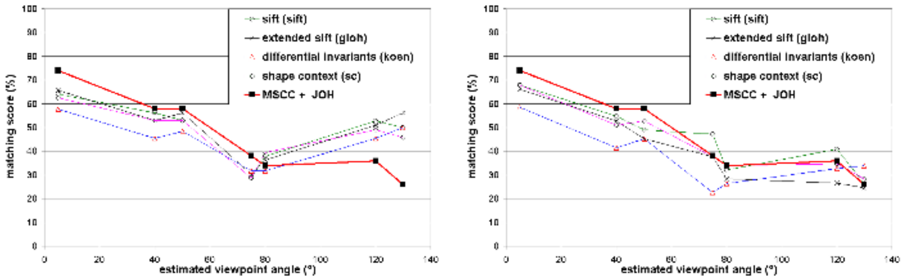


Fig. 6. Absolute numbers of non-overlapping matched regions. (a) Combining all detectors. (b) Combining MSER and MSCC. (c) Combining IBR and MSCC. (d) Combining HARAFF and MSCC. (e) Combining HESAFF and MSCC. (f),(g) Matches for combination of all detectors at 120° viewpoint change. The bright ellipses mark the additional matches obtained from the MSCC detector

4.4 Descriptor Evaluation

To compare our joint orientation histogram (JOH) descriptor against others we use Mikolajczyk's evaluation framework [16]. We show the results of SIFT-keys, extended SIFT-keys, differential invariants and shape context (see Fig. 7). All the other descriptors of the framework give similar results. For MSCC detector and JOH descriptor we use the convex hull of the MSCC region points instead of ellipse fitting for region overlap calculation. Fig. 7(a) shows the matching scores (on the "doors" scene) of our JOH descriptor on MSCC regions compared to different descriptors on Hessian-Affine regions, as we found them to give best results. In contrast to our approach the descriptors for the Hessian-Affine regions are calculated on affine normalized patches. Fig. 7(b) depicts the results for



(a)

(b)

Fig. 7. Matching scores for images from "doors" dataset on Hessian-Affine regions (a) and MSCC regions (b) for different viewpoints

the same scene but all descriptors are calculated on MSCC regions. For MSCC regions all detectors show similar performance. The same behavior is observed for Hessian-Affine regions.

5 Summary and Conclusions

We have presented a novel method for the detection of distinguished regions by clustering feature primitives - the so called Maximally Stable Corner Clusters (MSCC). We have developed a novel local descriptor based on 2D joint orientation (JOH) histograms ideally suited for the properties of the detector. We have evaluated the repeatability of the MSCC under changing viewpoints and compared the performance to other established detectors on planar and non-planar scenes. The results show a competitive performance of the MSCC. Further evaluations on the combination of different detectors have shown, that our detector consistently detects regions different from those of other detectors. Finally, we evaluated the performance of our JOH descriptor against others and obtained comparable results. The results indicate, that the detector successfully enriches the variety and power of the current available set of local detectors.

References

1. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. 13th British Machine Vision Conference, Cardiff, UK. (2002) 384–393
2. Tuytelaars, T., Gool, L.V.: Matching widely separated views based on affine invariant regions. International Journal of Computer Vision **1** (2004) 61–85
3. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision. (2003)

4. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proc. 8th IEEE International Conference on Computer Vision, Vancouver, Canada. (2001) I: 525–531
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
6. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2003)
7. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Proc. 7th European Conference on Computer Vision, Prague, Czech Republic. (2004) Vol I: 228–241
8. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference. (1988)
9. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark. (2002) I: 128 ff.
10. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision **30** (1998) 79–116
11. Brown, M., Lowe, D.: Invariant features from interest point groups. In: Proc. 13th British Machine Vision Conference, Cardiff, UK. (2002) Poster Session
12. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. International Journal of Computer Vision **37** (2000) 151–172
13. Bigün, J., Granlund, G.H.: Optimal orientation detection of linear symmetry. In: Proceedings of the IEEE First International Conference on Computer Vision, London, Great Britain (1987) 433–438
14. Cormen, T., Leiserson, C., Rivest, R.: Introduction to Algorithms. MIT Press, Cambridge MA (1990)
15. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press Professional (1990)
16. Mikolajczyk, K., Schmid, C.: Comparison of affine-invariant local detectors and descriptors. In: Proc. 12th European Signal Processing Conference, Vienna, Austria. (2004)
17. Fraundorfer, F., Bischof, H.: Evaluation of local detectors on non-planar scenes. In: Proc. 28th Workshop of the Austrian Association for Pattern Recognition, Hagenberg, Austria. (2004) 125–132