COMPARISON OF AFFINE-INVARIANT LOCAL DETECTORS AND DESCRIPTORS

Krystian Mikolajczyk and Cordelia Schmid

INRIA Rhône-Alpes GRAVIR-CNRS
655 av. de l'Europe, 38330 Montbonnot, France
email: Name.Surname@inrialpes.fr
web: www.inrialpes.fr/lear

ABSTRACT

In this paper we summarize recent progress on local photometric invariants. The photometric invariants can be used to find correspondences in the presence of significant viewpoint changes. We evaluate the performance of region detectors and descriptors. We compare several methods for detecting affine regions [4, 9, 11, 18, 17]. We evaluate the repeatability of the detected regions, the accuracy of the detectors and the invariance to geometric as well as photometric image transformations. Furthermore, we compare several local descriptors [3, 5, 8, 14, 19]. The local descriptors are evaluated in terms of two properties: robustness and distinctiveness. The evaluation is carried out for different image transformations and scene types. We observe that the ranking of the detectors and descriptors remains the same regardless the scene type or image transformation.

1. INTRODUCTION

Local photometric invariants have shown to be very successful in various applications including: wide baseline matching for stereo pairs [1, 9, 18], reconstructing cameras for sets of disparate views [14], image retrieval from large databases [15], model based recognition [8, 13] and texture classification [6]. They are robust to occlusion and clutter, distinctive as well as invariant to image transformations. In this paper we present an overview of local photometric invariants –affine-invariant regions and their description– and evaluate their performance.

There is a number of methods proposed in the literature for detecting local features in images. We require that a detector finds the same regions regardless the transformation of the image, that is the detection method is invariant to these transformations. This property can be measured with a relative number of repeated (corresponding) regions in two images. Figure 1(a) shows an example of images with a viewpoint change. Figure 1(b) displays affine regions detected with the approach proposed in [11]. The regions are represented by ellipses and the corresponding ellipses capture the same image part. We can use different techniques to describe the regions (see section 2.2). We require a descriptor to be robust and distinctive, that is to be insensitive to noise and to differ from descriptors computed on other regions. The descriptors can be also invariant to a class of transformations. Given the invariant descriptors and a similarity measure we can determine the corresponding regions by finding the most similar pairs of descriptors (matches).

The comparison of region detectors and descriptors is based on the number of corresponding regions and the number of correct matches between two images representing the same scene. To compute these numbers we use a test data with ground truth transformations. We use images of planar scenes, which are related by a homography. Thus, region-to-region correspondences can be easily found. The test data includes structured and textured scenes as well as different types of transformations: viewpoint changes, scale changes, illumination changes, blur and jpeg compression.

In this paper we address the questions: which region type and which descriptor leads to the best results, how the performance depends on the transformation between images and on the scene type. Previous evaluations [2, 11, 12, 16] of invariant detectors/descriptors were done independently and used different criteria and different test data, therefore the results were not comparable. Here we use a unified framework for the comparison.

2. INVARIANT LOCAL FEATURES

In this section we describe different affine-invariant local photometric descriptors. The first step is the extraction of affine-invariant regions, see section 2.1. Each of these regions is then characterized by a descriptor, see section 2.2.

2.1 Affine-invariant region detectors

There are numerous approaches for feature detection and the features extracted by these algorithms differ in localization, scale and structure (corners, blobs, multi-junctions). In general, the objective is to develop a detector invariant to a class of geometric and photometric transformations introduced by different viewing conditions. In the following we briefly describe the work related to the detection of affine-invariant regions. All the regions are affine invariant up to rotation, which can be recovered by other methods [8, 11].

Tuytelaars and Van Gool [17, 18] proposed two approaches for detecting affine invariant regions. The first is *geometry-based* and combines Harris points with nearby edges. An affine-invariant parallelogram region is determined by a Harris corner and two points on the nearby edges. The second method is purely *intensity-based* and is initialized with local intensity extrema. For each extremum the algorithm finds significant changes in the intensity profiles along rays going out from the extremum. An ellipse is fitted to the region defined by the locations of these changes. Matas et al. [9] proposed to use the water-shed algorithm to find intensity regions. It searches for regions that have either higher or lower intensity than the pixels on its outer boundary. These features are called Maximaly Stable Extremal Regions (*MSER*).

The second moment matrix is an efficient way to determine an affine-invariant region. It describes the first order signal changes in a neighborhood of a point. Distinctive fea-

(a) (b)

Figure 1: (a) Example of test images with viewpoint change of 60°. (b) Corresponding affine-invariant regions detected with Harris-Affine detector.

tures are localized on strong signal changes. Lindeberg and Garding [7] developed a method for finding blob-like affine features. The authors first extract maxima of the normalized Laplacian in scale-space and then iteratively modify the scale and shape of the regions based on the second moment matrix. Affine shape estimation was used for matching and recognition by Baumberg [1]. He extracts Harris interest points at several scales and then adapts the shape of the point neighborhood to the local image structure using the iterative procedure proposed by Lindeberg. The affine shape is estimated for a fixed scale and fixed location. Note that there are many points repeated at neighboring scale levels, which increases the probability of false matches as well as the complexity. Recently, Mikolajczyk and Schmid [11] proposed affine invariant *Harris-Affine* and *Hessian-Affine* detectors. They extended the scale invariant detectors [10] by the affine normalization. The location and scale of points are given by the scale-invariant Harris and Hessian detector.

A different approach for detection of scale and affine invariant features was proposed in [4]. They search for *Salient regions* which locally maximize the entropy of pixel intensity distribution.

2.2 Region descriptors

Many different techniques for describing image regions have been recently developed. The simplest descriptor is a vector of image pixels. However, it is not invariant to different geometric and photometric transformations. Moreover, the high dimensionality of such a description increases the computational complexity. Therefore, this technique is mainly used for finding point-to-point correspondences between two images. To reduce the dimensionality the distribution of the pixel intensities can be represented by a histogram. However, the spatial relations between pixels are lost and the distinctiveness of such descriptor is low. Another possibility is to use a distribution of gradient orientations weighted by the gradient values within a region. Lowe [8] developed a 3D histogram, called *SIFT*, where the dimensions are: gradient angle quantized to 8 principal orientations and 4x4 location grid on the region. Generalized *moment invariants* [19] have been introduced to describe the multi-spectral nature of the data. The moments characterize the shape and the intensity distribution in a region.

A family of descriptors is based on Gaussian derivatives and can be computed to represent a point neighborhood [3, 5]. The derivatives can be computed up to a given order and normalized to be invariant to pixel intensity changes. *Differential invariants* and *steerable filters* were successfully used for image retrieval [10, 15]. *Complex filters* [1, 14] were designed to obtain rotation invariance and are similar to the Gaussian derivatives. In the context of texture classification Gabor filters and wavelets are frequently used to describe the frequency content of an image.

3. EXPERIMENTAL SETUP

In this section we describe the evaluation framework. Sections 3.1 and 3.2 present the evaluation criteria for detectors and descriptors, respectively. In section 3.3 we discuss our test data.

3.1 Detector evaluation criteria

The important properties characterizing a feature detector are: the repeatability as well as the accuracy of localization and region estimation under different geometric and photometric transformations. We follow the approach proposed in [11] to evaluate the detectors. The repeatability score for a given pair of images is computed as the ratio between the number of region-to-region correspondences and the smaller number of regions detected in one of the images. We take into account only the points located in the part of the scene present in both images. Given the ground truth transformation we can find the corresponding regions. Moreover, we can project the regions from one image on the other and verify how closely the regions overlap. The overlap error between corresponding regions is the ratio $(1 - \texttt{intersection}/\texttt{union})$ of the elliptic regions and it is analytically computed using the ground truth transformation. The repeatability score depends on the arbitrary set overlap error. In this evaluation we compute the repeatability for different overlap error.

We evaluate six different detectors described in section 2.1: Harris-Affine, Hessian-Affine [11], MSER [9], Intensity based regions [18], Geometry based regions [17] and Salient regions [4].

3.2 Descriptor evaluation criteria

The regions should be repeatable, but it also very important in the matching process that the region descriptors are distinctive. Distinctiveness of the descriptor is measured with the Receiver Operating Characteristics (ROC) of detection rate versus false positive rate. The detection rate is the number of correct matches with respect to the number of corresponding regions. Two points are matched if the distance between their descriptors is below an arbitrary threshold. This threshold is varied to obtain the ROC curves. We use the ground truth homography to verify if the match is correct. The false positive rate is the actual number of false matches in a database of descriptors with respect to the number of all possible false matches. The images in the database are different from the query images, therefore all the matches in the

database are incorrect.

We compare six methods for computing region descriptors introduced in section 2.2: SIFT descriptors [8], steerable filters [3], differential invariants [5], complex filters [14], moment invariants [19], and cross-correlation.

3.3 Test data

We evaluate the descriptors on real images [1] with different geometric and photometric transformations. Different changes in imaging conditions are evaluated: gradual viewpoint changes, scale changes, image blur, JPEG compression and illumination. We use planar scenes such that the homography can be used to determine the correctness of a match. The images contain different structured (e.g. graffiti, buildings), or textured (e.g. trees, walls) scenes. To evaluate the false positive rate, that reflects the distinctiveness of the descriptors, we use a database of 1000 images extracted from a video.

4. COMPARISON RESULTS

In the following we present the evaluation results. In section 4.1 we discuss the results for detectors and section 4.2 for the descriptors.

4.1 Detectors

Figure 2 presents the results for the repeatability score for Graffiti images (cf. Figure 1). Figure 2(a) shows how the repeatability depends on the perspective transformation that the image undergoes. For a given overlap error of 50% (detection accuracy) we increase the transformation between the reference image and the query image and measure the relative number of corresponding regions. We can observe that the performance decreases for large viewpoint angles. Figure 2(b) displays the actual number of corresponding features with the overlap error of 50%.

We require a detector to have a high repeatability score and a large number of correspondences. For most transformations and scene types the MSER, Harris-Affine and Hessian-Affine regions obtain the best repeatability score. Harris and Hessian detector provide several times more corresponding regions than the other detectors but this number decreases for larger transformations between images.

Figure 2(c) shows the repeatability score computed for a pair of images with a viewpoint change of 50 degrees. We compute the repeatability score while varying the overlap error. The threshold rejects the corresponding regions detected with larger overlap error (lower accuracy). A high score for a small overlap error indicates a high accuracy of a detector. In all test MSER and Intensity based regions obtain higher repeatability score than other detectors for a small overlap error. The number of corresponding regions detected with Harris-Affine and Hessian-Affine significantly increases when larger overlap error is allowed.

4.2 Descriptors

In the following we discuss the results for descriptors evaluation. Figure 3(a) shows the detection rate with respect to false positive rate.

Figure 3(a) shows the results for a viewpoint change of the Graffiti sequence. Examples for other image transformations can be found in [12]. In all tests, except for light changes, SIFT descriptors obtain better results than the other descriptors. This shows the robustness and the distinctive character of the SIFT descriptors computed on image patches normalized to affine photometric and geometric transformations. It was designed to tolerate small shift error and affine deformations while keeping high distinctiveness. The second best descriptors are the steerable filters. Cross correlation gives unstable results. Its performance depends on the accuracy of region detection, which decreases for significant geometric transformations. Differential invariants give significantly worse results than the steerable filters, which is surprising as they are based on the same basic components (Gaussian derivatives). The multiplication of derivatives necessary to obtain the rotation invariance increases the instability of the descriptors.

The distinctiveness of a descriptor also depends on the type of detected regions. Figure 3(b) shows the matching score obtained with SIFT descriptors computed on different region types. It displays the relative number of correct matches for increasing viewpoint change between images. The results are consistent with those of the repeatability test. It means that the distinctiveness is similar for different region types. We can observe that the best results are obtained for MSER regions. This is because of the high accuracy of the MSER detector. Harris-Affine and Hessian-Affine come second. Note that the Harris based features contain strong signal changes in the center of the regions. The other detectors find blob like structures where the essential signal changes are on the region boundaries. Therefore, the region size have to be larger to capture the signal variations.

5. DISCUSSION AND CONCLUSIONS

In this paper we have presented an experimental evaluation of local affine invariant features for sequences with real geometric and photometric transformations. In all tests the best repeatability score was obtained by MSER as well as by the regions provided by Harris-Affine and Hessian-Affine detectors. MSER detector is also more accurate than other detectors. Harris affine detector provides several times more regions than other methods. We observe that the detectors provide complementary regions and their performance as well as the number of detected regions depends on the scene type. Harris and Hessian based approaches give stable results regardless the scene type, while the performance of other detectors is low for scenes with irregular texture.

In the descriptor evaluation, SIFT obtains the best results. Steerable filters can be considered as an alternative given the low dimensionality of this descriptor. Cross-correlation are not robust to small geometric deformations.

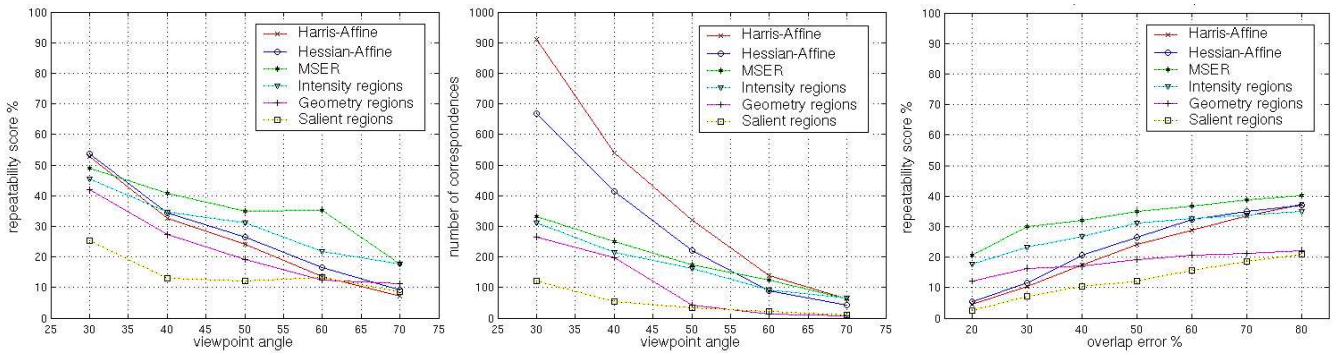
In the future work we plan to include in this comparison other affine detectors and local descriptors. It would be also of interest to evaluate the impact of different sources of error which can occur in the estimation of region parameters.

[1] http://www.inrialpes.fr/movi/Mikolajczyk/Database

REFERENCES

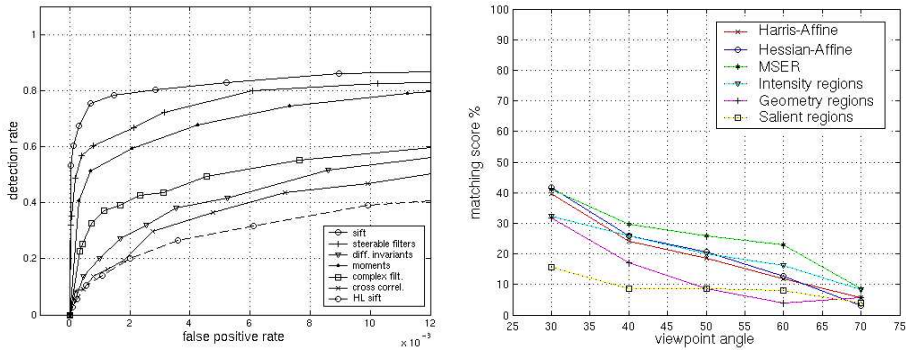
[1] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 774–781, 2000.

[2] G. Carneiro and A. D. Jepson. Phase-based local features. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume I, pages 282–296, 2002.



(a) (b) (c)

Figure 2: Evaluation of the affine detectors. (a) Repeatability score for an increasing viewpoint change. (b) Number of corresponding regions in the images. (c) Repeatability score for increasing overlap error.



(a) (b)

Figure 3: Descriptor evaluation. (a) ROC for different local descriptors computed on Harris-Affine regions. (b) Matching score for SIFT descriptor computed on different region types with respect to viewpoint changes.

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

[4] T. Kadir and A. Zisserman. An affine invariant detector. In *Proceedings of the 8th European Conference on Computer Vision, Pague, Tcheque Republic*, 2004.

[5] J. Koenderink and A. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

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

[7] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

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

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the 13th British Machine Vision Conference, Cardiff, England*, pages 384–393, 2002.

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

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

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

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

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

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

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

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

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

[19] L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *Proceedings of the 4th European Conference on Computer Vision, Cambridge, England*, pages 642–651, 1996.