Keypoint extraction and selection for object recognition

Maja Rudinac Boris Lenseigne Pieter Jonker
Delft Biorobotics Laboratory, Department of BioMechanical Engineering
Delft University of Technology, 2628 CD, Delft, The Netherlands
{ m.rudinac, b.a.j.lenseigne, p.p.jonker}@tudelft.nl

Abstract

In order to improve the performance of affine invariant detectors, approaches that combine different keypoint extraction methods can be found in literature. However, such a combining has two major drawbacks: a high computational cost in matching similar objects, and a large number of false positives because only a relatively small subset of those keypoints is really discriminative. In this paper we propose a method to overcome these limitations: First we combine different keypoint extractors in order to obtain a large set of possible interest points for a given object. Then, a two step filtering approach is applied: First, a reduction using a spatial criterion to reject points that are close in a specified neighborhood, and second, filtering based on the information entropy in order to select only a small subset of keypoints that offer the highest information content. A qualitative analysis of this method is presented.

1. Introduction

In cluttered real world scenes, object recognition is a demanding task and its success depends on the algorithm's invariance to partial occlusions, illumination changes and main object variations. For these situations, local invariant features seem to provide the most promising results [1, 2] since they are robust to occlusions, background clutter and content changes [3]. Variations of these features are successfully used in many applications. They are used to describe the object appearance in order to determine the object class in bag of words models [4], where the information about feature location is neglected. Or they are applied in applications where the information about the spatial feature distribution is crucial, such as in localization of autonomous mobile robots [5]. Besides local invariant features several other methods proved to be very successful for object detection in real world situations. Ekvall uses receptive field cooccurrence histogram for vision guided robot grasping [18] while Bicego proposes Hidden Markov Models in combination with wavelets for appearance based 3D object recognition [17].

In our research, recognition is used for the purpose of object localization and grasping with various robotic arms, so both information about the object class and its current location must be provided in real-time. The conditions present while creating object models differ a lot from the situation when an object should be recognized and grasped. Moreover, our recognition framework should work in industrial as well as in lab environments. For these reasons, experimenting with local invariant features is a logical step. In this paper we present part of our work, the method for selecting the most representative points in the scene. In order to gain as much information as possible, we decided to combine different keypoint extraction methods for detecting the object and then to reduce the number of found keypoints using an independent measure for information content. This reduction is performed for two reasons: to keep the most discriminative points, and to speed up the matching. For creating the object models, the most representative keypoints are then described using SIFT [1] and GLOH [6] descriptors. This paper is organized as follows: chapter 2 gives a short overview of related work. The detailed explanation of our approach and the test results are presented in chapters 3-5. The final conclusions are drawn in the chapter 6.

2. Related work

Robust and affine invariant keypoint extraction is a well known problem and recently intensive research in this area has been done. A very detailed evaluation of affine region detectors made by Tuytelaars et al. [6] gives a framework for testing future detectors as well as the state of the art and their performance. Analysis showed that the detectors extract regions with different properties and the overlap of these regions is so small, if not empty, that one detector can outperform others only in one type of scenes or one type of transformation. In order to obtain the best performance, several detectors should be used simultaneously. This observation inspired us to experiment with different keypoint extraction methods.

The best overall results were obtained using MSER [7] followed by the Hessian Affine detector [6]. Apart from these two several other evaluations of detectors were published, e.g. detectors for 3D by Morales [8] and local features for object class recognition by Mikolajczyk [9].Experiments showed that the Hessian-Laplace in combination with GLOH gives the best overall result. Stark [10] confirmed this and concluded that the choice of the detectors is much more important for the overall performance of the recognition than the choice of descriptors. For this reason we limited our descriptor set to just two that proved to be the best: SIFT and GLOH.

Several authors tried to combine local invariant features but without significant results [11, 12]. Experiments showed that combinations of detectors perform better than one detector alone, if they produce keypoints in different parts of the image. However, the main problem they encountered is the matching speed of the detected keypoints and a high number of false matches due to the fact that only a small number of points is really discriminative. The conclusion that rises is that a reduction must be applied. In this paper we proposed a method to overcome the mentioned limitations.

3. Keypoints combining

Several detectors and descriptors were used as building blocks in our combined algorithm. A short description of every one of them follows below.

3.1. Building blocks

a) Hessian Affine: It spatially localizes and selects the scale and affine invariant points detected at multiple scales using the Harris corner measure on the second-moment matrix. On each individual scale, interest points are chosen based on the Hessian matrix at that point [3, 16].
b) Harris Affine: This relies on the combination of corner points detected through Harris corner detection, multi-scale analysis through Gaussian scale-space, and affine normalization using an iterative affine shape adaptation algorithm. It makes it possible to identify similar regions between images that are related through affine transformations and which have different illumination [3, 16].
c) Hessian Laplace: A method that responds to blob like structures. It searches for local maxima of the Hessian determinant and selects a characteristic scale where the Laplacian attains an extremum in scale-space [16].
d) MSER: A method for blob detection in images which denotes a set of distinguished regions which are defined by an extremal property of its intensity function in the region and on its outer boundary. It was originally used to find correspondences between image elements from two images with different viewpoints [7].

3.2. Syntheses

In our approach we decided to combine different detectors in order to extract a large set of points which offer as different information about the object as possible. Combinations of either two or three different detectors were applied simultaneously on the image and all extracted keypoints were saved together in the same subset. We combined mostly the detectors a), b) and d) while the other combinations were used for comparison.

Since the number of keypoints is extremely large we consequently apply a reduction method, so that only the *n* most representative points for every extracted combination are selected. For this reduced set of keypoints, a SIFT or GLOH descriptor is calculated, forming the feature matrix for a given image.

In total, we experimented with 7 different combinations, and simulation results for all of them can be found in the section 5. A schematic overview of the proposed method is shown in the figure 1.

4. Method for keypoint reduction

We propose to use a two step algorithm for keypoint reduction: First, we apply reduction using a spatial criteria to reject points that are close in a specified neighborhood, and then we filter based on the information entropy, in order to select only a small subset of the most representative keypoints offering the highest information content.

4.1. Reduction using spatial criteria

As keypoints close to each other represent redundant information, these points are first filtered using a spatial criterion. Every keypoint is represented by an ellipse that defines the affine region. For every pair of keypoints we evaluate whether the absolute distance between the centers of their ellipses lies within a certain threshold. If this is so, it means that those points lie in the neighborhood of each other and only one point from the pair is kept. A threshold is determined by manual tuning and in the end we established the 9 neighborhood of a point as a measure of closeness for our application. For a more restrictive reduction higher thresholds can be chosen.

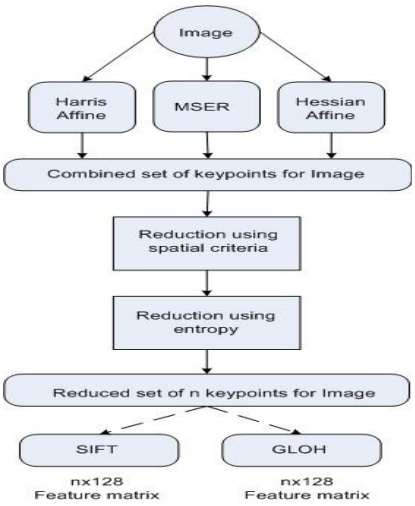

Figure 1: Scheme of the proposed method

4.2. Reduction using entropy

Since the set of extracted keypoints is computed using different techniques, an independent measure of keypoint relevance must be applied. It is shown that the probability of correct matching increases with increasing information content [13, 14]. That inspired us to use the entropy of local regions for distinctive keypoint selection. We propose the following algorithm:

1. For every keypoint a region of interest is set being the 9 neighborhood around it.
2. If the keypoint is on the edge of the image and its region of interest is out of the image boundary, we clone the pixel values from the existing part and fill in the missing values of the 9 neighborhood (see figure 2).
3. Calculate the local entropy using (1) for every pixel within the region of interest. In this formula P_i is the probability of a pixel *i* within the region of interest.

$$H = -\sum_i P_i \log_2 P_i \quad (1)$$

4. The Entropy of the region of interest is now estimated as the Euclidean norm of entropy values calculated for every pixel in the previous step.
5. Repeat steps 1 till 4 for every keypoint.
6. Sort keypoints in descending order, according to the entropy values of the region of interest around them.
7. Select only the first *n* ranked keypoints with the highest entropy values.

8. Calculate the SIFT or GLOH descriptor only for those most representative keypoints.

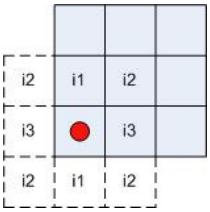

Figure 2: Cloning the missing pixels

In testing we used $n = 200$ as a threshold value but this number depends on the application and it is really difficult to predict how many keypoints are really necessary for an efficient recognition. One should also bear in mind that a higher number of extracted keypoints leads to a larger number of false positives. Obviously, a tradeoff must be made between this high threshold and a small number of features that allow fast matching and a low threshold and a higher number of features which will provide more information about the image content.

5. Testing and analysis

In order to determine the quality of selected keypoints, we tested our algorithm using a standard framework for detector performance evaluation proposed by Mikolajczyk et al. [16]. The detector performance is characterized using the repeatability defined as the average number of corresponding regions detected in images under different transformations. The repeatability score is calculated using ground truth data for three different types of scenes that represent the main transformations. Results are shown for the scene with a viewpoint change, a zoomed and rotated scene and a scene with varying lighting conditions. Since we work with object recognition in real world situations with a constant change of environmental conditions, good results under these transformations are of crucial importance. We tested the following detectors and their combinations, labelled: hesaff – Hessian Affine; mshesa – MSER + Hessian Affine; harhes – Harris affine + Hessian Affine; mseraf – MSER; kombaf – MSER + Harris affine + hessian affine; heslap – Hessian Laplace. For the purpose of testing we reduced the number of keypoints approximately 10 times compared to its original size and the repeatability is calculated separately for the reduced set and for the original one. The results are shown in figures 3-8.

The overall conclusion can be drawn that if the original set is reduced even 10 times in size, the repeatability score will decrease no more than 10% for all three types of scenes, while the speedup in matching is significant. We also tried higher reduction thresholds and noticed a linear decrease in repeatability, meaning that depending on the application a different number of keypoints can be selected. The best results in the repeatability tests were achieved by MSER, which underpinned the conclusions from literature. Since MSER shows significant drawbacks in clustering and localization [11], we looked at combined detectors as an alternative solution. Our results justified our hypothesis, since the second best is kombaf – a combination of three different detectors, which also gives a more discriminative representation of the image.

In order to localize the objects in the scene, all keypoints were described with a 128 dimensional vector of either SIFT or GLOH features. For matching we deploy a simple keypoint voting using the Euclidean distance as the measure of the closeness of points. The computational complexity of such a matching is proportional to the squared number of keypoints, so a reduction of 10 times gains a significant speedup.

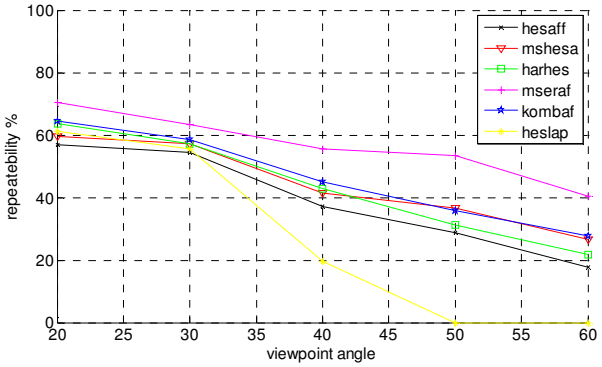

Figure 3: No reduction, scene with viewpoint change

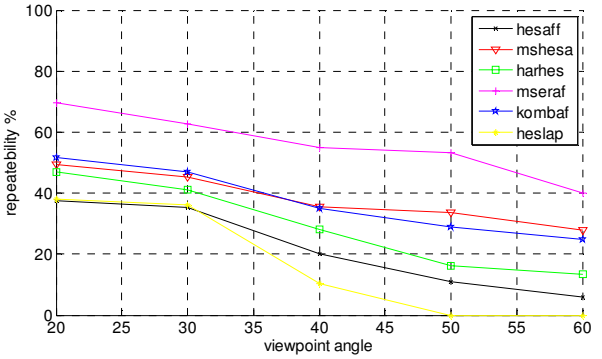

Figure 4: Reduction, scene with viewpoint change

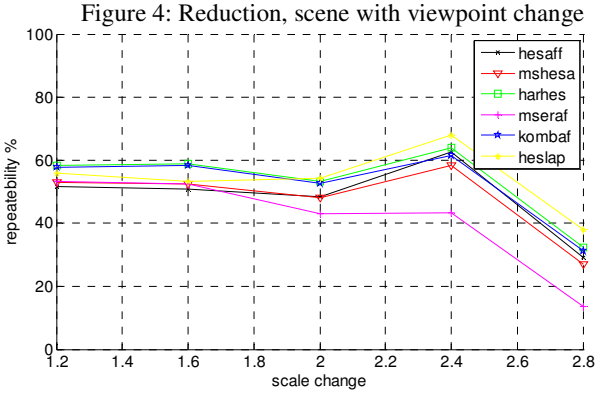

Figure 5: No reduction, zoomed and rotated scene

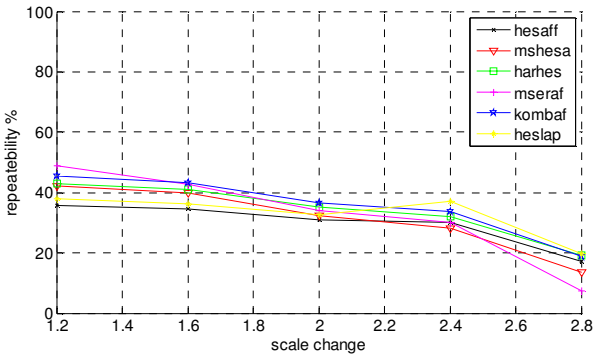

Figure 6: Reduction, zoomed and rotated scene

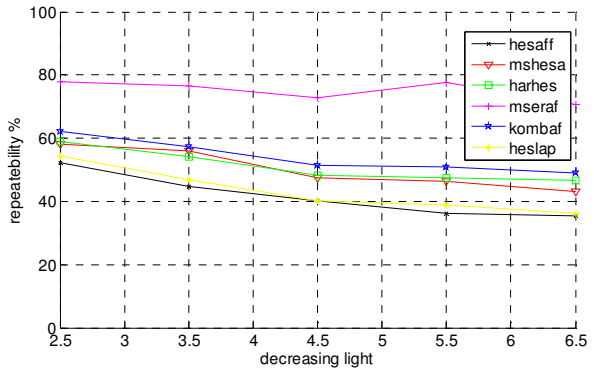

Figure 7: No reduction, scene with light change

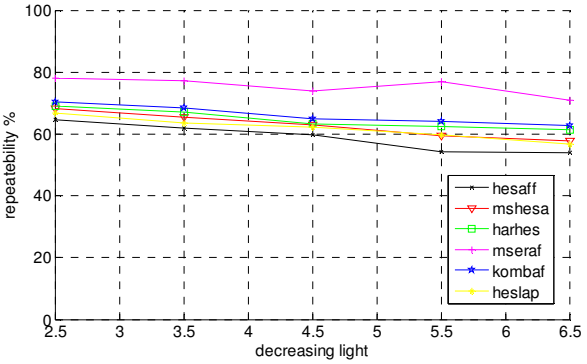

Figure 8: Reduction, scene with light change

6. Conclusion

In this paper we proposed an algorithm for the reduction of a large set of keypoints which we collected using different keypoint extraction methods and their combinations. Our approach consists of a two step filtering: spatial filtering to reduce close points as a first step and a selection of the most discriminative points with the highest information content as the second step. The overall performance of the method was tested using a standard framework for testing the quality of detectors. Our results showed that reducing the set of keypoints to only 10% of its original size leads to a less then 10% decrease in the repeatability score, while the matching speed is significantly improved.

In our future work we would like to expand our feature set with more global descriptors such as shape context and different color and texture descriptors, and to try to combine that information with the one gained from the keypoints. We hope that using such a versatile approach, more precise information about the object appearance as well as its location in the scene could be obtained.

Acknowledgments

This work has been carried out as part of the FALCON project under the responsibility of the Embedded Systems Institute. This project is partially supported by the Netherlands Ministry of Economic Affairs under the Embedded Systems Institute (BSIK03021) program.

References

[1] D. G. Lowe: “Distinctive image features from scale invariant key points”, IJCV, 60(2):91–110, 2004

[2] H.Bay, A. Ess, T. Tuytelaar, and L. Van Gool: “Speeded-Up Robust Features (SURF)”, Computer Vision and Image Understanding 110(3): 346-359, 2008

[3] K. Mikolajczyk, C. Schmid: "Scale & Affine Invariant Interest Point Detectors", International Journal of Computer Vision 60(1): 63-86, 2004

[4] R. Fergus, P. Perona, A. Zisserman: "Object Class Recognition by Unsupervised Scale-Invariant Learning", Conference on Computer Vision and Pattern Recognition (2) 264-271, 2003

[5] T. Goedemé, M. Nuttin, T. Tuytelaars, L. Van Gool: "Omnidirectional Vision Based Topological Navigation", International Journal of Computer Vision 74(3): 219-236, 2007

[6] T. Tuytelaars, K. Mikolajczyk: "Local Invariant Feature Detectors: A Survey", Foundations and Trends in Computer Graphics and Vision 3(3): 177-280, 2007

[7] J. Matas, O. Chum, M. Urban, T. Pajdla: "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions", British machine vision conference, 384-393, 2002

[8] P. Moreels and P. Perona: "Evaluation of Features Detectors and Descriptors based on 3D Objects", International Journal on Computer Vision 73(3), 263-284, 2007

[9] K. Mikolajczyk, B. Leibe, B. Schiele: "Local Features for Object Class Recognition", International conference on computer vision, (2), 1792-1799, 2005

[10] M. Stark and B. Schiele, "How good are local features for classes of geometric objects", International Conference on Computer Vision, 1-8, 2007

[11] A. Ramisa, M. Lopez, A. Ramon, R. Toledo: “Comparing Combinations of Feature Regions for Panoramic VSLAM”, International Conference on Informatics in Control, Automation and Robotics, 292-297, 2007

[12] F. Fraundorfer, H. Bischof: "A novel performance evaluation method of local detectors on non-planar scenes", International conference Computer Vision and Pattern Recognition - Workshops, 33-33, 2005

[13] J.L. Starck and F. Murtagh: “Multiscale entropy filtering. Signal Processing”, 76(2), 147-165, 1999

[14] G. Fritz, L. Paletta, H. Bischof: "Object Recognition Using Local Information Content", International Conference on Pattern Recognition (2), 15-18, 2004

[15] T. Kadir, M. Brady: "Saliency, Scale and Image Description", International Journal of Computer Vision 45(2), 83-105, 2001

[16] K.Mikolajczyk, T.Tuytelaars, C.Schmid, A.Zisserman, J.Matas, F.Schaffalitzky, T.Kadir, L.Van Gool: "A Comparison of Affine Region Detectors", International Journal of Computer Vision 65(1-2), 43-72, 2005

[17] J. M. Bicego, U. Castellani, and V. Murino: "A hidden Markov model approach for appearance-based 3D object recognition", Pattern Recognition Letters. 26, 2588-2599, 2005

[18] S. Ekvall, D. Kragic and P. Jensfelt: "Object detection and mapping for service robot tasks", Robotica 25(2) 175-187, 2007