Wide Baseline Stereo Matching Using Voting Schemas

David Xavier Aldavert Miró*, Arnau Ramisa Ayats× and Ricardo Toledo Morales+

* *Centre de Visió per Computador, Universitat Autònoma de Barcelona, 08193 Bellaterra, SPAIN*
E-mail:aldavert@cvc.uab.es

+ *Centre de Visió per Computador, Universitat Autònoma de Barcelona, 08193 Bellaterra, SPAIN*
E-mail:ricardo@cvc.uab.es

× *Institut d'Investigació en Intel.ligencia Artificial, Universitat Autònoma de Barcelona, 08193 Bellaterra, SPAIN*
E-mail:aramisa@iiia.csic.es

Abstract Local regions have proven to be very successful to put in correspondence images subject to point of view, scale or illumination changes. The correspondence accomplishment of local regions depends on the performance of the local region descriptor method. Depending on the region characteristics, a descriptor will be better than other, therefore combining different descriptors using a voting schema could improve the performance of the local region matching process.

Keywords: Pattern Recognition, affine covariant regions, local descriptors, voting schemas

1 Introduction

Local invariant features have become a powerful tool for finding correspondences between images or objects seen from a different point of view. Their local character gives them robustness against occlusions or dynamic background and invariance makes them resistant to point of view, scale or illumination changes. The detectors with a greater degree of invariance are the affine covariant region detectors that have recently appeared. Aside from the local regions detectors, a descriptor is needed in order to put in correspondence the same region seen from two different points of view. There is a large number of possible descriptors which emphasize different image properties like pixel intensities, colour, texture, edges, etc. In this work, we only focus on descriptors computed on grey-value images.

Depending on local region characteristics a descriptor may have a better performance than other. Therefore, the combination of several types of descriptors could increases the region matching efficiency. In this paper, we compare the performance of several descriptors used independently and combined using several voting schemas.

This article is structured as follows: In section 2 a short review of the state of art in invariant local features is presented. Section 3 develops our descriptor combination approach using voting schemas. Then, in section 4, there is a comparation of the performance of descriptors; individually and combined in voting schemas for several affine invariant region detectors. Finally, the discussion and conclusions of this work are presented in section 5.

2 Affine invariant features

In order to put two images in correspondence, detection and description of local regions is needed. In [1] the performance of the main affine covariant region detectors is evaluated. The evaluation criteria are the repeatability and accuracy measures, i.e. the ratio of matched regions and the overlapping surface between the same regions in two different views. In this work is shown that the best results are obtained by the Harris/Hessian-Affine[2] detectors and Maximal Stable Extrema Regions or MSER[3] detector. Considering this results, only Harris/Hessian-Affine and MSER are taken into account in this work. Harris-Affine detects Harris interest points at several scales and then employs the Lindeberg's schema for

scale selection and affine adaptation. Hessian-Affine is the same as Harris-Affine with the difference that interest points are detected using the determinant of the Hessian matrix. Finally, MSER detects regions which intensity values are higher or lower than all the pixels of their neighbourhood.

Once interest regions are extracted using invariant region detectors, these regions must be characterized in order to put them in correspondence. The simplest comparison between two regions can be done using a vector of image pixels as a descriptor and cross-correlation as a similarity measure between two descriptors. However, the higher is the dimensionality of a description the higher is the computational complexity for recognition. To reduce descriptor dimensionality, instead of using all region pixels, some property of the region is measured and used as a descriptor.

In [4] K. Mikolajczyk and C. Schmid evaluated the performance of many region descriptors based on image intensity values. Descriptors are tested under diverse image transformations such as changes in view points, illumination, zoom and rotation. The descriptors that have a best response are SIFT, GLOH, Shape Context, Steerable filters and Generalized Colour Moments.

The SIFT descriptor [5] divides the local region patch into 4x4 subregions and for each subregion builds a gradient orientation histogram which is quantized into eight orientations. The contribution of each gradient orientation in its histogram is weighted by its gradient magnitude and by a Gaussian weighting function with σ equal to one half the width of the local region patch. Weighting by a Gaussian function gives less significance to gradients which are far from the centre of the descriptor, as these gradients are most affected by misregistration errors. The GLOH descriptor [4] is an extension of the SIFT descriptor. The main difference between GLOH and SIFT is the number and distribution of histogram bins. The shape context descriptor [6] is another histogram descriptor similar to the SIFT descriptor, but it is based on Canny edges. Steerable filters [7] is a differential descriptor which computes the derivatives up to a given order in the centre of the normalized region patch to approximate its neighbourhood. And finally, the Generalized Colour Moments (GCM) [8] calculates powers of the image coordinates and of intensities of different colour channels. They yield a broader set of features to build the moment invariants and, as a result, these moment invariants are simpler and more robust than the classical ones.

Finally, a similarity measure is needed to put in correspondence two regions. The Euclidean distance between region descriptors can be used as a similarity measure, so that, two regions A and B are matched if the descriptor D_B is the nearest to the descriptor D_A. To reject possible false correspondences, the distances between the nearest neighbour and the second nearest neighbour are compared:

$$\frac{\|D_A - D_B\|}{\|D_A - D_C\|} < t \tag{1}$$

where D_A, D_B and D_C are the descriptors of the region A, region's A nearest neighbour and region's A second nearest neighbour, respectively. If the ratio is below the threshold t, then the putative matching is accepted, otherwise it is rejected.

3 Combining descriptors through voting schemas

As descriptors reflect different information of a region patch, information brought by one descriptor type could complement information brought by another descriptor type, so that, combining different descriptors could increase the overall matching performance. To combine the different descriptors the plurality, borda-count and condorcet voting schemas are used. First of all, for each descriptor type, the k-nearest corresponding regions are chosen. Then, depending on the voting schema used a different punctuation is assigned to chosen regions:

- **Plurality:** This is the simplest voting schema used by Matas et. al. in [3] for robust region matching. In this schema a vote is cast to a candidate region for each descriptor which selected it.

- **Borda-Count:** The Borda-count [9, 10, 11] is a consensus voting system where the selected

Figure 1: Examples of image pairs from data sets related only by an homography.

match is not the most voted but the broadly acceptable by all descriptor classifiers. Borda-count gives k votes for the region descriptor nearest neighbour, $k-1$ for the second nearest neighbour, $k-2$ for the third, and so on.

- **Condorcet Pair-wise Comparison:** In Condorcet Pair-wise Comparison method [11, 12], each candidate is matched one-to-one with each of the other candidates. Candidate regions gets 1 point for a head-to-head win (the region R_B^m wins R_B^n if it is nearer to R_A for descriptor D_i) and half a point for a tie.

After all descriptors have been voted, votes are summed and candidate regions which have a number of votes bellow a given threshold are rejected. We experimentally determined that $k = 10$ gives good results and a region needs to have a least, half of maximum possible score to be a putative matching.
For plurality, the maximum possible score depends only on the number of descriptors used. For Borda-count, the maximum possible scores is $k * d$ where d is the number of descriptor types. For Condorcet, the maximum possible scores is $(k-1)*d^2$. As we use 6 descriptors, a putative region needs at least 3 scores in plurality, 30 scores in Borda-count and 180 scores in Condorcet to be considered a putative matching.

Finally, after the voting, the GLOH descriptor is used in order to select the final putative matching from the remaining candidate regions. As in the previous section, the distance ratio between the nearest neighbour and the second nearest neighbour is verified in order to reject possible false region correspondences.

Figure 2: Examples of image pairs from data sets affected by occlusions, repetitive or poor texture and strong depth discontinuities.

4 Results

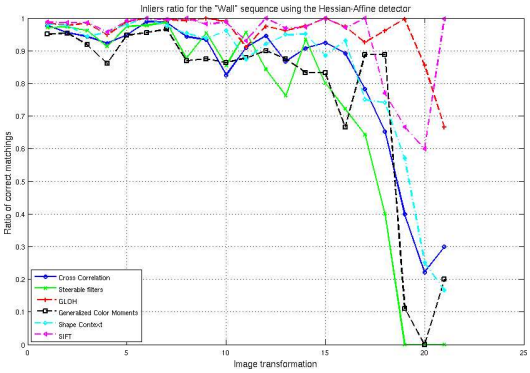
In [1, 4] Mikolajczyk has exhaustively studied the performance of different local region detectors and descriptors. The images of his dataset are related by different transformations such as illumination and point of view changes, zoom, rotation and/or image blurring. However, other distortions like partial occlusions, depth discontinuities and deformations due to objects non planar shape are also important. Thus, we defined another dataset and test the performance of the different descriptors for the Harris-Affine, Hessian-Affine and MSER. This dataset consist on five different image sets obtained with a digital camera, with a resolution of 1024 by 768 and each set have about 32 image pairs:

- **Wall set:** Images are mainly related by changes on the point of view. Regions are extracted from a high textured planar surface, so that this is the best situation for matching.
- **Illumination set:** Images are mainly related by changes on the illumination conditions. Unlike illumination of Mikolajczyk's data set, the images of our set are taken at different moments during the day.
- **Table set:** Images are mainly related by defo-

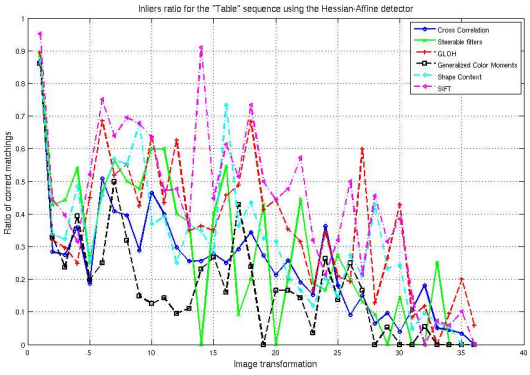
cusing distortion and changes on the point of view. In this set, the distance between objects and camera is small, so that objects cannot be assumed to be flat. Besides, table texture is very repetitive so the number of outliers would be great.

- **Room set:** Images are related by illumination and point of view changes. Images are taken from a room, so there are occlusions, zones with different levels of texture and depth discontinuities.

- **Train set:** Images are related by a change of the point of view. This set was acquired from a train in movement, so that the point of view change between two camera locations is very large.



(a)



(b)

Figure 3: Ratio of detected inliers respect all corresponding regions detected for: (a) wall set and (b) table set.

Then, for each detector and descriptor pair we determine the best nearest neighbour's threshold. We select the threshold that maximizes the number of image correspondences in the database. The criteria is based on epipolar geometry, so that two images related if at least 8 valid matching are detected and the amount of false correspondences not exceed the fifty percent of the total. Once, a threshold is determined we evaluate the performance of the different detector/descriptor pairs. We have compared the ratio of detected inliers against all corresponding regions estimated and the number correctly detected regions. To decide which is the best descriptor, we have compared the mean ratio of inliers against all detected correspondences and the mean number of inliers detected for all image pairs of the dataset. In tables 1(a) and 1(b), we show the results obtained for each descriptor/detector pair. The SIFT gives better results for Harris-Affine and MSER detectors and GLOH for Hessian-Affine detector.

(a) Mean of correct correspondences ratio:

	Harris	Hessian	MSER
SIFT	**0.67**	0.76	**0.75**
GLOH	0.67	**0.80**	0.76
Shape Context	0.63	0.73	0.72
Steerable fil.	0.42	0.51	0.55
GCM	0.53	0.67	0.69
Cross Corr.	0.54	0.69	0.60

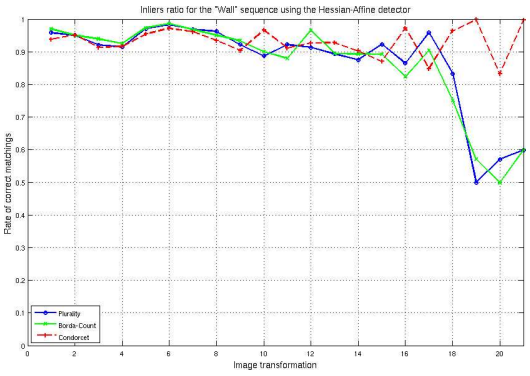
(b) Mean of number of correct correspondences:

	Harris	Hessian	MSER
SIFT	**11.5**	23	**13**
GLOH	11	**27.5**	11.5
Shape Context	9.5	19.5	12
Steerable fil.	5	8.5	8
GCM	2	7.5	5.5
Cross Corr.	9	20.5	11

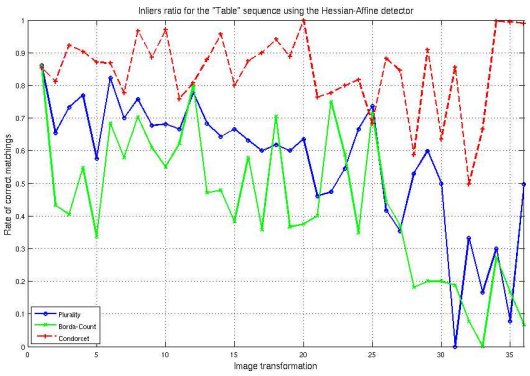
Table 1: Results obtained using descriptors

In figure 3 results obtained for "wall" and "table" sets, using the Hessian-Affine detector are shown. In the "wall" set, SIFT and GLOH have excellent results and the other descriptors have a good performance. In this set, regions are related by ideal conditions, i.e. regions are related only by an homography and objects have a high textured planar shape. However, in the set "table", images have a very repetitive texture and some objects have partial occlusions. In this set is difficult to decide which is the best descriptor because descriptors have a poor and "unstable" performance, i.e. the

number of inliers hardly exceed the fifty percent of all detected regions for all descriptors. Besides, a descriptor with a good performance in an image pair, in the following pair may have a poor performance. Considering this results we decided to combine the weakness and strengths of all descriptors in a voting schema.



(a)



(b)

Figure 4: Ratio of detected inliers using voting schemas for: (a) wall set and (b) table set.

In figure 4, the results obtained using the voting schemas are shown. For the set "wall", the results are slightly better however in "table" set, the ratio of detected inliers is clearly better that the results obtained by descriptors used individually. In this set, the Condorcet voting schema is clearly better than the others. This may be because in the presence of repetitive texture, some local regions of the image will have a very similar descriptor, so that, the nearest neighbours of a given region will vary between the different description measure. Then, depending on the voting schema, the number of votes received by a candidate region will vary more or less. In plurality, a candidate region could, at most, loss a vote for each descriptor type while in Borda-count and Condorcet, a candidate region could loss up to 10 votes for each descriptor type. However, in Condorcet if a region is not within the 10 nearest neighbours, a candidate region could loss a great amount of votes. For example, lets a voting process resulting in 30 candidate regions. For each candidate region, the difference of votes between being inside the result set or not, is one vote in Plurality and Borda-Count and 11 votes in Condorcet. Therefore, regions that are selected only by some descriptors have a lower number of votes in Condorcet than in Plurality and Borda-Count.

In tables 2(a) and 2(b) the mean ratio of inliers against all detected correspondence and the mean number of inliers detected for all image pairs of the dataset using voting schemas are shown. The Condorcet voting schema has the best results for each detector in the ratio of detected inliers as well as in the number of detected inliers. In addition, comparing this results with the results obtained in tables 1(a) and 1(b), we could see that the Condorcet voting schema have better results than SIFT and GLOH descriptors used individually.

(a) Mean of correct correspondences ratio:

	Harris	Hessian	MSER
Plurality	0.56	0.68	0.55
Borda-Count	0.58	0.71	0.61
Condorcet	**0.74**	**0.87**	**0.76**

(b) Mean of number of correct correspondences:

	Harris	Hessian	MSER
Plurality	12	24.5	10
Borda-Count	11.5	22	10
Condorcet	**12**	**29.5**	**13**

Table 2: Results obtained using voting schemas

5 Discussion and Conclusions

In this paper, we have presented an experimental evaluation of local region descriptors in the presence of image transformation such as point of view or illumination changes, occlusions and depth discontinuities. We also presented the results obtained combining all descriptors using voting schemas. The

goal was to compare the performance of descriptors using some of the affine covariant region detectors that have recently appeared.

When descriptors are used individually, GLOH and SIFT have the best results. Using Harris-Affine and MSER detectors, SIFT have slightly better results than GLOH, whereas using Hessian-Affine detector, GLOH is slightly better than SIFT. Using all descriptors combined in a voting schema, the matching performance is increased up to 10%. In our experiments, the Condorcet voting schema obtains better results than Borda-Count, Plurality and all descriptors used individually. Globally, the results seem to indicate that voting schemas increases the quality and quantity of matched regions.

As a future work we plan to include more scene categories and a greater number of image pairs for each category. In addition, the presented voting schemas shall be improved weighting the contribution of each descriptor.

References

[1] K. Mikolajczyk, *et al.*, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005.

[2] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[3] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions." in *Proceedings of the British Machine Vision Conference 2002, BMVC 2002, Cardiff, UK, 2-5 September 2002*. British Machine Vision Association, 2002.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," January 2001.

[7] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891–906, 1991.

[8] F. Mindru, T. Moons, and L. V. Gool, "Recognizing color patterns irrespective of viewpoint and illumination," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 01, p. 1368, 1999.

[9] J. de Borda, "Mémoire sur les élections au scrutin, historie de l'académie royale des sciences," Paris, 1781.

[10] T. K. Ho, J. J. Hull, and S. N. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp. 66–75, 1994.

[11] K. T. Leung and D. S. Parker, "Empirical comparisons of various voting methods in bagging," in *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM Press, 2003, pp. 595–600.

[12] M.-J. Condorcet, "Éssai sur l'application de l'analyse à la probabilité des décisions rendus à la pluralité des voix," Paris, 1785.