COMPARING COMBINATIONS OF FEATURE REGIONS FOR PANORAMIC VSLAM

Arnau Ramisa, Ramón López de Mántaras
Artificial Intelligence Research Institute, UAB Campus, 08193, Bellaterra, SPAIN
aramisa@iiia.csic.es, mantaras@iiia.csic.es

David Aldavert, Ricardo Toledo
Computer Vision Center, UAB Campus, 08193, Bellaterra, SPAIN
aldavert@cvc.uab.es, ricardo.toledo@cvc.uab.es

Keywords: Affine covariant regions, local descriptors, interest points, matching, robot navigation, panoramic images.

Abstract: Invariant (or covariant) image feature region detectors and descriptors are useful in visual robot navigation because they provide a fast and reliable way to extract relevant and discriminative information from an image and, at the same time, avoid the problems of changes in illumination or in point of view. Furthermore, complementary types of image features can be used simultaneously to extract even more information. However, this advantage always entails the cost of more processing time and sometimes, if not used wisely, the performance can be even worse. In this paper we present the results of a comparison between various combinations of region detectors and descriptors. The test performed consists in computing the essential matrix between panoramic images using correspondences established with these methods. Different combinations of region detectors and descriptors are evaluated and validated using ground truth data. The results will help us to find the best combination to use it in an autonomous robot navigation system.

1 INTRODUCTION

Autonomous robot navigation is one of the most challenging problems of mobile robotics and, although it has been widely studied, it is far from being solved completely. To date, the most successful approaches are a set of techniques known as SLAM (Simultaneous Localization And Mapping) (Thrun, 2002). These methods consist in iteratively searching for an optimal solution to both problems: self localization and map building. SLAM methods can be classified amongst three main categories: metric SLAM, for methods that give an accurate localization and mapping (Thrun, 2002; Castellanos and Tardos, 2000); topologic SLAM, where the environment is usually represented as a graph of “places” and the connectivity information amongst them (Tapus and Siegwart, 2006); and finally hybrid approaches, which try to combine the advantages of both techniques and reduce their drawbacks (Tomatis et al., 2002).

To correct the accumulative errors of odometry, SLAM techniques use additional information from the environment acquired with sensors like sonars, laser range-scanners, etc. When a camera is used to obtain such data, the method is known as Visual SLAM. Recently, this approach has gained strength thanks to the development of new computer vision algorithms that extract discriminative and meaningful information from the images. One promising line of research consists in using invariant visual features extracted from images to construct a map of the environment and locate the robot in it (Se et al., 2001; Booij et al., 2006; Ramisa et al., 2006). The core of these methods consist in finding corresponding features between two or more images acquired by the robot and set up relations between the places where this images were taken. One particularly interesting subset of invariant features are the affine covariant regions, which can be correctly detected in a wide range of acquisition conditions (Mikolajczyk et al., 2005). Equally important are the local descriptors such as SIFT (Lowe, 2004), which make the matching of local regions acquired in different conditions possible.

In (Ramisa et al., 2006), the authors developed a topological localization method which uses constellations of affine covariant regions from a panoramic image to describe a place (for example a room). When a new panorama of features is acquired, it is compared to all the stored panoramas of the map, and the most similar one is selected as the location of the robot. Using different types of feature detectors and descriptors simultaneously increases the probability of finding good correspondences, but at the same time can cause other problems, such as more processing time and more false correspondences. As means to improve the results of their approach, in this article various of these covariant region detectors and descriptors are compared. Our objective is to evaluate the performance of different combinations of these methods in order to find the best one for visual navigation of an autonomous robot. The results of this comparison will reflect the performance of these detectors and descriptors under severe changes in the point of view in a real office environment. With the results of the comparison, we intend to find the combination of detectors and descriptors that gives better results with widely separated views.

The remainder of the paper is organized as follows. Section 2 provides some background information in affine covariant region detectors and descriptors. Section 3 explains the experimental setup used in the comparison and section 4 presents the results obtained. Finally, in section 5 we close the paper with the conclusions.

2 DETECTORS AND DESCRIPTORS

Affine covariant regions can be defined as sets of pixels with high information content, which usually correspond to local extrema of a function over the image. A requirement for these type of regions is that they should be covariant with transformations introduced by changes in the point of view, which makes them well suited for tasks where corresponding points between different views of a scene have to be found. In addition, its local nature makes them resistant to partial occlusion and background clutter.

Various affine covariant region detectors have been developed recently. Furthermore, different methods detect different types of features, for example Harris-Affine detects corners while Hessian-Affine detects blobs. In consequence, multiple region detectors can be used simultaneously to increase the number of detected features and thus of potential matches.

However, using various region detectors can also introduce new problems. In applications such as VSLAM, storing an arbitrary number of different affine covariant region types can increase considerably the size of the map and the computational time needed to manage it. Another problem may arise if one of the region detectors or descriptors gives rise to a high amount of false matches, as the mismatches can confuse the model fitting method and a worse estimation could be obtained.

Recently Mikolajczyk et al. (Mikolajczyk et al., 2005) reviewed the state of the art of affine covariant region detectors individually. Based on Mikolajczyk et al. work, we have chosen three types of affine covariant region detectors for our evaluation of combinations: Harris-Affine, Hessian-Affine and MSER (Maximally Stable Extremal Regions). These three region detectors have a good repeatability rate and a reasonable computational cost.

Harris-Affine first detects Harris corners in the scale-space using the approach proposed by Lindeberg (Lindeberg, 1998). Then the parameters of an elliptical region are estimated minimizing the difference between the eigenvalues of the second order moment matrix of the selected region. This iterative procedure finds an isotropic region, which is covariant under affine transformations.

The Hessian-Affine is similar to the Harris-Affine, but the detected regions are blobs instead of corners. Local maximums of the determinant of the Hessian matrix are used as base points, and the remainder of the procedure is the same as the Harris-Affine.

The Maximally Stable Extremal region detector proposed by Matas et al. (Matas et al., 2002) detects connected components where the intensity of the pixels is several levels higher or lower than all the neighboring pixels of the region.

Matching local features between different views implicitly involves the use of local descriptors. Many descriptors with wide-ranging degrees of complexity exist in the literature. The most simplest descriptor is the region pixels alone, but it is very sensitive to noise and illumination changes. More sophisticated descriptors make use of image derivatives, gradient histograms, or information from the frequency domain to increase the robustness.

Recently, Mikolajczyk and Schmid published a performance evaluation of various local descriptors (Mikolajczyk and Schmid, 2005). In this review more than ten different descriptors are compared for affine transformations, rotation, scale changes, jpeg compression, illumination changes, and blur. The conclusions of their analysis showed an advantage in performance of the Scale Invariant Feature Transform (SIFT) introduced by Lowe (Lowe, 2004) and one of its variants: Gradient Location Orientation Histogram

(GLOH) (Mikolajczyk and Schmid, 2005). Based on these results, we use these two local descriptors in our experiments. Both SIFT and GLOH descriptors divide the affine covariant region in several subregions and construct a histogram with the orientations of the gradient for each subregion. The output of both methods is a 128-dimension descriptor vector computed from the histograms.

3 EXPERIMENTAL SETUP

In this section, we describe our experimental setup. The data set of images is composed of six sequences of panoramas from different rooms of our research center. Each sequence consists of 11 to 25 panoramas taken every 20 cm. moving along a straight line predefined path. The panoramas have been constructed by stitching together multiple views taken from a fixed optical center with a Directed Perception PTU-46-70 pan-tilt unit and a Sony DFW-VL500 camera.

Apart from the changes in point of view, the images exhibit different problems such as illumination changes, repetitive textures, wide areas with no texture and reflecting surfaces. These nuisances are common in uncontrolled environments.

From each panorama, a constellation of affine covariant regions is extracted and the SIFT and the GLOH descriptors are computed for each region. In Figure 1 a fragment of a panorama with several detected Hessian-Affine regions can be seen.

To find matches between the feature constellations of two panoramas, the matching method proposed by Lowe in (Lowe, 2004) is used. According to this strategy, one descriptor is compared using euclidean distance with all the descriptors of another constellation, and the nearest-neighbor wins. Bad matches need to be rejected, but a global threshold on the distance is impossible to be found for all situations. Instead, Lowe proposes to compare the nearest-neighbor and the second nearest-neighbor distances and reject the point if the ratio is greater than a certain value, which typically is 0.8. Lowe determined, using a database of 40,000 descriptors, that rejecting all matches with a distance ratio higher than this value, 90% of the false matches were eliminated while only 5% of correct matches were discarded.

Finally, the matches found comparing the descriptors of two constellations are used to estimate the essential matrix between the two views with the RANSAC algorithm. As in the case of conventional cameras, the essential matrix in cylindrical panoramic cameras verifies,

$$p_0^\top E p_1 = 0, \tag{1}$$

where p_0 and p_1 are projections of a scene point P in the two cylindrical images related by the essential matrix E. However, the epipolar constraint defines a sinusoid instead of a line. This sinusoid can be parameterized with the following equation,

$$z_1(\phi) = -\frac{n_x cos(\phi) + n_y sin(\phi)}{n_z}, \tag{2}$$

where $z_1(\phi)$ is the height corresponding to the angle ϕ in the panorama, $n_1 = [n_x, n_y, n_z]$ is the epipolar plane normal, obtained with the following expression,

$$n_1 = p_0^\top E. \tag{3}$$

The test performed consists in estimating the essential matrix between the first panorama of the sequence and all the remaining panoramas using different combinations of detectors and descriptors. As random false matches will rarely define a widely supported epipolar geometry, finding a high number of inliers reflects a good performance. To validate the results, ground truth essential matrices between the reference image and all the other images of each sequence have been computed using manually selected corresponding points. These essential matrices are then used to compute the error of the inliers of each combination of detectors and descriptors to the ground truth epipolar sinusoid.

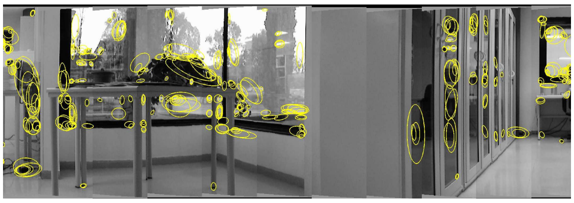

Figure 1: Some Hessian-Affine regions in a fragment of a panorama.

4 RESULTS

To evaluate the performance of each combination of methods, we measured the maximum distance at which each combination of methods passed the three different tests that are explained in the following paragraph. The results of the tests are presented in the Table 1. It is important to notice that these distances are the mean across all the panorama sequences.

Since a minimum of 7 inliers are required in order to find the essential matrix, the first test shows at which distance each method achieves less than 7 inliers. For the second test, the inliers that do not follow equation 1 for the ground truth essential matrices are rejected as false matches. Again, the distance at which the number of correct inliers drops below 7 is checked. Finally, the third test evaluates at which distance the percentage of correct inliers drops below 50%, which is the theoretic breakdown point for the RANSAC algorithm. This third test is the hardest and the more realistic one.

Table 1: Results of the comparison. For convenience we have labeled M:MSER, HA:Harris-Affine, HE:Hessian-Affine, S:SIFT, G:GLOH.

	Test 1	Test 2	Test 3
M+G	180cm	140cm	83cm
HA+G	320cm	200cm	106cm
HE+G	380cm	220cm	108cm
M+S	180cm	120cm	84cm
HA+S	400cm	220cm	101cm
HE+S	480cm	200cm	107cm
M+HE+G	480cm	200cm	106cm
HA+HE+G	480cm	220cm	100cm
M+HA+G	480cm	220cm	99cm
M+HE+S	480cm	180cm	99cm
HA+HE+S	480cm	260cm	111cm
M+HE+S	480cm	220cm	109cm
M+HA+HE+G	480cm	240cm	87cm
M+HA+HE+S	480cm	260cm	82cm

The results of the first test show that, except for the Hessian-Affine and SIFT, the combination of various detectors performs better than one detector alone. In the second test we can see that the performance of all the methods is greately reduced, and the combination of two methods (except for the Harris-Affine, Hessian-Affine and SIFT) drops to a similar level to that of one method alone. Finally, regarding the third test, the performance drops again, putting all combinations at a similar level (around 100 cm). For the third test an exponential function has been fitted to the data sets to aproximate the behaviour of the noisy data and find the estimated point where the ratio falls below 0.5.

In Figure 2 the ratio of inliers for the best combinations of each category is shown (namely, Harris-Affine and GLOH, Hessian-Affine and Harris-Affine and SIFT, and all the detectors and GLOH) as well as the fitted exponential of each of the three combinations. As can be observed in the point's cloud, the method using two regions in general achieved better results than the other two methods, and several times achieved a performance above 0.5 after the estimated point.

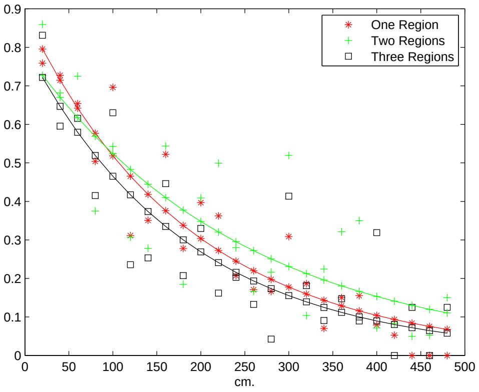

Figure 2: Ratio of inliers for the best combinations of one region (Hessian-Affine and GLOH), two regions (Harris-Affine and Hessian-Affine and SIFT) and three regions (MSER, Harris-Affine, Hessian-Affine and GLOH). Additionally, the exponential fitting of the different data sets is shown.

To obtain an estimation of the precision, the mean distance error of the inliers to the estimated epipolar sinusoid and the corresponding ground truth epipolar sinusoid has been computed for the three selected combinations (Hessian-Affine and GLOH, Hessian-Affine and Harris-Affine and SIFT, and all the detectors and GLOH). The results of this comparison are presented in figure 3. It can be seen that for the first 250 cm. all the methods have a similar error, both for the estimated epipolar sinusoid and for the ground truth one. Discontinuities are due to failures of the combinations to compute a valid essential matrix at a given distance.

Finally, a performance test has been done to compare the processing speed of the different region detectors and descriptors. This results, shown in Table

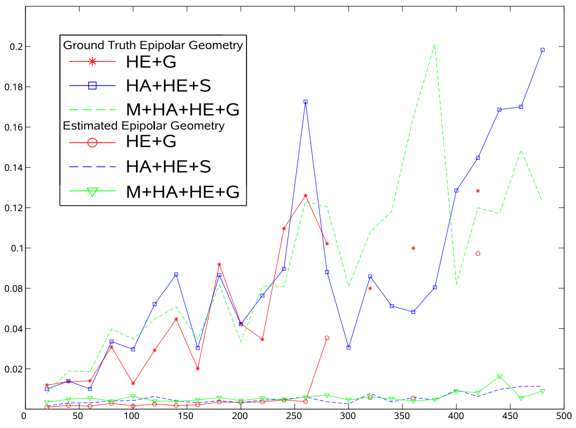

Figure 3: Mean distance error of a match to the estimated epipolar sinusoid and to the ground truth epipolar sinusoid.

2, are the mean of 50 runs of a 4946x483 panoramic image. The implementation used to perform the tests is the one given in `http://www.robots.ox.ac.uk/~vgg/research/affine/` by the authors of the different region detectors (Mikolajczyk et al., 2005). It is important to note that this implementations are not optimal, and a constant disk reading and writting time is entailed. The tests where done in a AMD Athlon 3000MHz computer.

Table 2: Time Comparision of the different region detectors and descriptors.

	Time (sec)	Regions Processed
MSER	0.95	828
Harris-Affine	8.47	3379
Hessian-Affine	3.34	1769
SIFT	14.87	3379
GLOH	16.64	3379

5 CONCLUSIONS

In this paper, we have evaluated different combinations of affine covariant region detectors and descriptors to correclty estimate the essential matrix between pairs of panoramic images. It has been shown that the direct combination of region detectors finds a higher number of corresponding points but this, in general, does not translate in a direct improvement of the final result, because also a higher number of new outliers are introduced.

No significant differences in performance have been found between the detectors individually, except that MSER is notably faster than the other methods thanks to its very simple algorithm; nevertheless the detected regions are very robust. However MSER finds a low number of regions and, as the robot moves away from the original point, the matches become to few to have a reliable estimation of the essential matrix. Regarding combinations of two region detectors, they find estimations of the essential matrix at longer distances. However, as shown in the results, the distance at which the estimations are reliable is similar to that of one detector alone. Finally, the combinations of three descriptors have shown to perform worse than the combinations of two. The reason is probably the number of false matches, that confuse the RANSAC method.

Regarding the descriptors, no significant differences have been found between the SIFT and the GLOH. The GLOH gives a slightly better performance than the SIFT, but also requires a bit more processing time.

Additionally we have found the practical limits in distance between the panoramic images in order to have a reliable estimation using these methods. This value can be used as the distance that a robot using this method to navigate in an office environment is allowed to travel before storing a new node in the map.

Future work includes experimentig with a larger data set and with different kinds of environments to verify and extend the presented results. Another interesting line of continuation would be investigating better matching and model fitting methods to reduce the proportion of false matches, and also researching ways to combine the different types of regions taking into account the scene content or estimated reliability of each region detector.

ACKNOWLEDGEMENTS

The authors want to thank Adriana Tapus for her valuable comments and help. This work has been partially supported by the FI grant from the Generalitat de Catalunya, the European Social Fund and the MID-CBR project grant TIN2006-15140-C03-01 and FEDER funds.

REFERENCES

Booij, O., Zivkovic, Z., and Kröse, B. (2006). From sensors to rooms. In *Proc. IROS Workshop From Sensors to Human Spatial Concepts*, pages 53–58. IEEE.

Castellanos, J. A. and Tardos, J. D. (2000). *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach*. Kluwer Academic Publishers, Norwell, MA, USA.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *Int. J. Comput. Vision*, 30(2):79–116.

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

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2002). 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.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V.

(2005). A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72.

Ramisa, A., Aldavert, D., and Toledo, R. (2006). A panorama based localization system. In *1st CVC Research and Development Workshop, 2006,* pages 36–41. Computer Vision Center.

Se, S., Lowe, D., and Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2051–2058, Seoul, Korea.

Tapus, A. and Siegwart, R. (2006). A cognitive modeling of space using fingerprints of places for mobile robot navigation. In *Proceedings the IEEE International Conference on Robotics and Automation (ICRA), Orlando, U.S.A., May 2006*.

Thrun, S. (2002). Robotic mapping: A survey. In Lakemeyer, G. and Nebel, B., editors, *Exploring Artificial Intelligence in the New Millenium*. Morgan Kaufmann.

Tomatis, N., Nourbakhsh, I., and Siegwart, R. (2002). Hybrid simultaneous localization and map building: Closing the loop with multi-hypotheses tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Washington DC, USA, May 2002*.