A Panoramas-Based Localization System

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

* *Institut d'Investigació en Intel.ligencia Artificial, Campus de la UAB, 08193 Bellaterra, SPAIN*
E-mail:aramisa@iiia.csic.es

× *Centre de Visió per Computador, Campus de la UAB, 08193 Bellaterra, SPAIN*
E-mail:aldavert@cvc.uab.es

+ *Centre de Visió per Computador, Campus de la UAB, 08193 Bellaterra, SPAIN*
E-mail:ricardo.toledo@cvc.uab.es

Abstract In this paper a global localization method is proposed to be used in a topological navigation scheme. Such method characterizes, in a distinctive way, places visited by a robot. The method extracts a constellation of various types of local affine covariant features from a panoramic image and then computes a local descriptor for each detected region. This constellation is then added to a map represented as a graph where nodes are places represented by local feature panoramas and edges are adjacency relations between known places. Then the robot find its localization in this map by acquiring a new local feature constellation and comparing it to those stored in the graph.

Keywords: Robot Vision, Machine Vision, Pattern Recognition, Topological Robot Navigation.

1 Introduction

Affine covariant local region detectors and descriptors have received a considerable attention recently. They also have been shown to work well for different tasks such as image stitching [1], object and object class recognition [2, 3, 4], or autonomous robot navigation [5]. In this work we will use them to characterize a place, usually a room in indoor environments. Such characterization will be used for localize a robot in topological navigation.

Related work

This approach is similar in spirit to the one presented by Tapus et al. [6]. In her work, Tapus proposes a *fingerprint* to characterize every place the robot visits. This *fingerprint* is composed of various features extracted from an omni-directional image and a 2D laser range-scanner reading. From the image the features extracted are color blobs and vertical lines and from the laser reading 3D corners. A circular string is constructed using all the features extracted. Each character in the string represents a particular type of feature. Then a string matching method that is inspired by the minimum energy algorithm used in stereo vision is used to decide in which of the mapped places is located the robot.

The differences between our proposed approach and the one of Tapus is that in our case the only perceptual information used is a panoramic image. In addition, the position of the robot in relation to the reference panorama can be computed up to a scale ambiguity.

This paper is organized as follows. Section 2 reviews the affine covariant regions and local descriptors used, as well as the matching procedure between regions. Section 3 presents the proposed strategy to find correspondences between panoramas and to estimate the local position of the robot. Section 4 shows our global localization results in real environments and, finally, section 5 presents the conclusions and future work.

2 Affine covariant local features

Extracting local features from an image provides a way to reduce the dimensionality of the data, making it manageable for high-level vision tasks. In addition, this simplification provides robustness against noise,

aliasing or acquisition conditions.

Local features can be defined as points or regions with high information content, which correspond to local extrema of a function over the image. An additional requirement for local features is that they should be resistant to image transformations like, for example, changes in the point of view or illumination.

This robustness to changes makes local features well suited to tasks such as matching and recognition. In addition, its local nature makes them resistant to partial occlusion and background clutter. In order to compare local features they are characterized using descriptors.

To correctly match two instances of the same local feature, their descriptors must be as similar as possible. Therefore, the region used to compute the descriptor should be composed by the same pixels independently of the differences in point of view, illumination or scale between the two images.

In a recent article, Mikolajczyk et al. [7] reviewed the state of the art of affine covariant region detectors and compared several of the latest techniques. Based on Mikolajczyk work, and in the results of some experiments carried out by ourselves, we have chosen three affine covariant local feature detectors: *Harris affine*, *Hessian affine* and *MSER*. These three region detectors have a high repeatability rate with a good precision and a reasonable computational cost. The Harris affine detector is an improvement of the Harris corner detector [8] and locates regions around the corners in the image in a scale and affine invariant way using the scale-space approach proposed by Lindeberg [9].

The Hessian affine is similar to the Harris affine, but in this case the detected regions are blobs instead of corners. Local maximums of the determinant of the Hessian matrix are used as base points.

The Maximally Stable Extremal region detector proposed by Matas et al. [10] detects connected components at all the possible thresholding levels of an image. The concept of *extremal region* defines a set of pixels with a value either higher or lower than all the neighboring pixels, which can be seen as a local maximum or minimum (an extremum) of the surface defined by pixel intensities. Finally, *maximally stable* refers to extremal regions where the intensity values of the pixels of the region is several levels higher (or lower) compared to the neighbors.

Figure 1: Example of normalized regions.

The regions detected with this methods are denoted in the image by an ellipse that encloses all the region pixels. Later, this ellipse is remapped to a circle to normalize the region. Figure 1 shows an example of this normalization. Even though both normalized regions are very similar, in the top right corner we can observe one of the drawbacks of the affine covariant region detectors used: the planarity assumption. Regions detected in an affine invariant manner can be recovered under projective transformations as long as the surface imaged in the region can be approximated by a plane.

Local features are of little use if they can not be compared and matched with regions from other images.This step implicitly involves the use of a *local descriptor*. The objective of these descriptors is to provide a compact and distinctive representation of the local feature to simplify the matching stage, and at the same time induce robustness to the remaining variations in the measurement regions. These variations can be illumination changes, noise and changes in the measurement region introduced by deep discontinuities or non-flat surfaces.

Recently, Mikolajczyk and Schmid published a performance evaluation of various local descriptors [11]. 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 this analysis observe an advantage in performance of the *Scale Invariant Feature Transform* introduced by Lowe [12, 2] and its variants GLOH [11] and, at a certain distance, PCA-SIFT [13].

Taking these results and the growing number of applications that make use of the SIFT into account, we have chosen this descriptor and its variant GLOH.

To compute the SIFT descriptor, the local region is divided into 16 sub-regions and an histogram of the orientations of the gradient is computed for every sub-region. A gradient can move within a sub-region and still produce the same descriptor, in this way the shift in position is allowed to compensate for small 3D rotations.

The orientations are quantized by the magnitude of the gradient to lower the contribution of instable orientations from sample points in flat zones of the image. The histograms have eight bins, each of 45 degrees. A trilinear interpolation is used to distribute gradient samples across adjacent bins of an histogram, to avoid boundary effects in the gradient orientation. In this way, a small change in the orientation will not change the descriptor abruptly. Once all the histograms are constructed, the values of all bins are arranged as a vector of 128 dimensions.

Gradient location-orientation histogram (GLOH) proposed by Mikolajczyk and Schmid [11] is an extension of the SIFT descriptor. The algorithm to compute the descriptor is the same except for the distribution of the sub-regions. Here a log-polar location grid is used, with three sub-regions in radial direction, each divided into 8 sub-regions in angular direction except the central sub-region. This results in 17 sub-regions. Instead of 8 orientation bins for each histogram, 16 bins are used. The resulting feature vector has 272 values, which are reduced to 128 with PCA.

To compare two descriptors, the Euclidian distance between the descriptors is computed. In principle, a global threshold on the distance could be used to filter the improbable matches, but such a value is difficult to adjust to perform well in every possible situation. The proposal of Lowe [2] is to compute matches comparing the distance of the local feature with the nearest-neighbor and with the second nearest-neighbor. If these two distances are too similar, the nearest-neighbor is not disctintive enough to be considered a correct match, so it is discarded. The threshold to the relation between the distances of the first two nearest-neighbors proposed by Lowe is 0.8.

3 Global localization method

The main goal of this work is to construct a topological localization system using only a conventional panning camera or an omni-directional camera.

In our case, the *fingerprint* is a constellation of feature regions extracted from a panoramic image of each room in the map. All these feature region panoramas are arranged as the nodes of a graph, where edges represent accessibility information between places. This graph is the topological map for the robot. The local feature region detectors used are the MSER [10], the Harris affine and the Hessian affine [14]. To characterize the detected regions, we have used the descriptors with better results as stated in [11]: the Scale Invariant Feature Transform (SIFT) [12, 2] and one of its variants, the Gradient Location and Orientation Histogram (GLOH) [11].

The motivations to use local features to construct the *fingerprint* of each room are:

- A *fingerprint* based on local features has the advantage of being more resistant to occlusions and partial changes in the image. This robustness is obtained because many individual local regions are used for every *fingerprint* and, thus, if some of them disappear the constellation can still be recognized. If new or different local features appear in a new constellation of a place, the matching can, still be successfully done against the stored *fingerprint*. Lowe showed in [2], in the context of object recognition, that correct subsets of features where selected even if they represented only 1% of the total local features.

- Local features have demonstrated its usefulness in many tasks, some of them of great interest to mobile robotics. For example, the same features can be used in motion estimation, 3D object recognition [2] and panorama construction [1], reducing the computational load for the robot.

- These features are robust to several distortions common to the images acquired by the robot.

- Finally, some local navigation information can be obtained through the geometric interpretation of the changes occurred in the constellation of feature points.

The steps of the proposed method are the following

- A new panoramic image is acquired.
- The local affine covariant regions are detected using the three region detectors.
- The detected regions are described using either the SIFT or the GLOH descriptors.
- The constellation of described feature regions is compared with every constellation in the map of the robot.
- To discard false matches, the essential matrix between the panoramas is computed using RANSAC, and as a result the matches will be classified as inliers or outliers.
- Finally, the panorama with the highest number of inliers is selected as the corresponding panorama where the robot is located.

As mentioned above, an interesting property of our approach is that it implicitly recovers metric information up to a certain degree. The room is modelled as a cylinder and using the essential matrix, which in any case has to be computed, we can recover the position of the robot in relation to the reference panorama up to a scale factor, which can be later estimated using stereovision techniques if at least one more panorama of the room is available and the metric distance between two of the panoramas is available [15].

4 Results

The test set consists of 27 panoramas from the IIIA research center captured rotating a Sony DFW-VL500 camera and stitching the images in a cylindrical panorama. The location of the test panoramas can be observed in figure 2, and one of the panoramas acquired in figure 3 with the detected affine covariant regions. The panoramas correspond to different rooms and corridors of the center, and there are at least two panoramas of every mapped place.

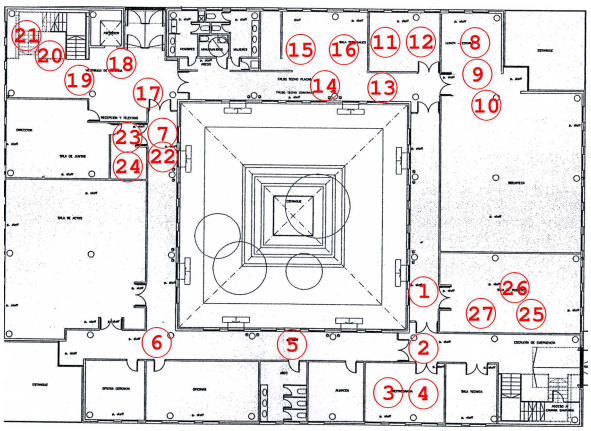

Figure 2: Map of the center with the location of the acquired panoramas.

Figure 3: Example of panorama.

For every panorama a constellation of feature regions was constructed using the three region detectors proposed, and the regions where described using the SIFT and the GLOH descriptors.

Even though a mobile robot can capture multiple panoramas when navigating within a room, our test was performed assuming only one panorama was available as a first measure of the distinctiveness of the proposed *fingerprint*. Therefore, two tests for every panorama were performed: a test with feature regions described with the SIFT descriptor and another test using the GLOH descriptor. The following results where obtained: using the SIFT 13 of the 27 panoramas where correctly matched, thus giving an accuracy of 48.1% of matches at the first attempt; in the case of the GLOH, 15 panoramas where correctly matched, which represent an accuracy of 55.6%.

5 Conclusions

The conclusions of this work are the following: Even though the presented schema obtained some success, it needs to be improved in order to reliably localize a mobile robot. Possible ways to improve this method are:

- Using other local feature detectors and descriptors.

- Specifying the contribution of every type of local feature employing, for example voting schemas.
- Finding a different matching strategy between local region descriptors. If many similar objects and, thus, similar regions are present in the image, the relation of the distance between the two nearest-neighbors gives poor results. A matching strategy based on lists of tentative matching could improve the performance.
- Color information is neglected in this approach but it could be very helpful in the localization process.
- Finally, to reliably estimate the essential matrix, it is important that features are homogeneously distributed over the panorama. In consequence, a method to discriminate which panorama will better represent a place is necessary.

References

[1] M. Brown and D. G. Lowe, "Recognising panoramas," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 1218.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Interantional Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] G. Dorkó and C. Schmid, "Object class recognition using discriminative local features," INRIA - Rhone-Alpes, Rapport de recherche RR-5497, February 2005. [Online]. Available: http://lear.inrialpes.fr/pubs/2005/DS05a

[4] S. Helmer and D. G. Lowe, "Object class recognition with many local features," in *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 12*. Washington, DC, USA: IEEE Computer Society, 2004, p. 187.

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

[6] A. Tapus, N. Tomatis, and R. Siegwart, "Topological global localization and mapping with fingerprint and uncertainty," in *International Symposium on Experimental Robotics(ISER04)*, Singapore, June 2004.

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

[8] C. Harris and M. Stephens, "A combined corner and edge detector," in *4th Alvey Vision Conference*, 1988, pp. 147–151.

[9] T. Lindeberg and J. Gårding, "Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure." *Image Vision Comput.*, vol. 15, no. 6, pp. 415–434, 1997.

[10] 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.

[11] 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.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*. Washington, DC, USA: IEEE Computer Society, 1999, p. 1150.

[13] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, 2004, pp. II–506–II–513 Vol.2.

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

[15] S. B. Kang and R. Szeliski, "3-D scene data recovery using omnidirectional multibaseline stereo," in *CVPR '96: Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*. Washington, DC, USA: IEEE Computer Society, 1996, pp. 364–370.