

# Robust Vision-based Robot Localization using Combinations of Local Feature Region Detectors

Arnaud Ramisa · Adriana Tapus · David Aldavert · Ricardo Toledo ·  
Ramon Lopez de Mantaras

Received: date / Accepted: date

**Abstract** This paper presents a vision-based approach for mobile robot localization. The model of the environment is topological. The new approach characterizes a place using a signature. This signature consists of a constellation of descriptors computed over different types of local affine covariant regions extracted from an omnidirectional image acquired rotating a standard camera with a pan-tilt unit. This type of representation permits a reliable and distinctive environment modeling. Our objectives were to validate the proposed method in indoor environments and, also, to find out if the combination of complementary local feature region detectors improves the localization versus using a single region detector. Our experimental results show that if false matches are effectively rejected, the combination of different covariant affine region detectors increases notably the performance of the approach by combining the different strengths of the individual detectors. In order to reduce the localization time, two strategies are evaluated: re-ranking the map nodes using a global

---

This work was 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 and the grant 2005-SGR-00093 and the MIPRCV Consolider Ingenio 2010.

A. Ramisa · R. Lopez de Mantaras  
Artificial Intelligence Research Institute (IIIA, CSIC), UAB  
Campus, E-08193 Bellaterra, Barcelona, Spain  
E-mail: aramisa@iiia.csic.es, mantaras@iiia.csic.es

A. Tapus  
University of Southern California, Los Angeles, CA, 90089, USA  
E-mail: adriana.tapus@ieee.org

D. Aldavert · R. Toledo  
Computer Vision Center (CVC), UAB Campus, E-08193 Bellaterra, Barcelona, Spain  
E-mail: david.aldaert@cvc.uab.cat, ricardo.toledo@cvc.uab.cat

similarity measure and using standard perspective view field of 45°.

**Keywords** Topological Localization · Vision Based Localization · Panoramic Vision · Affine Covariant Region Detectors

## 1 Introduction

Finding an efficient solution to the robot localization problem will have a tremendous impact on the manner in which robots are integrated into our daily lives. Most tasks for which robots are well suited demand a high degree of robustness in their localizing capabilities before they are actually applied in real-life scenarios (e.g., assistive tasks).

Since localization is a fundamental problem in mobile robotics, many methods have been developed and discussed in the literature. The existing approaches can be broadly classified into three major types: metric, topological and hybrid. Metric approaches (Dissanayake et al, 2001; Castellanos and Tardos, 1999; Thrun, 1998, 2000) are useful when it is necessary for the robot to know its location accurately in terms of metric coordinates (i.e. Cartesian coordinates). However, the state of the robot can also be represented in a more qualitative manner, by using a topological map (i.e. adjacency graph representation) (Choset and Nagatani, 2001; Tapus and Siegwart, 2006; Beeson et al, 2005). Because the odometry does not provide enough and complete data in order to localize a mobile autonomous robot, laser range finders and/or vision sensors are usually used to provide richer scene information. The rapid increase in computational power in the last few years had a significant impact in the development of better approaches to solve the simultaneous localization and

mapping (SLAM) problem, by using qualitative information provided by vision. Furthermore, vision units are cheaper, smaller and more practical than large expensive laser scanners.

In this work, we propose a topological vision-based localization approach of a mobile robot evolving in dynamic indoor environments. Robot visual localization and place recognition are not easy tasks, and this is mainly due to the perceptive ambiguity of acquired data and the sensibility to noise and illumination variations of real world environments. We propose to approach this problem by using a combination of affine covariant detectors so as to extract a robust spatial signature of the environment.

We decided to use combinations of the following three feature region detectors: MSER (Maximally Stable Extremal Regions) (Matas et al, 2002), Harris-Affine (Lindeberg, 1998), and Hessian-Affine (Mikolajczyk and Schmid, 2004), which have shown to perform better when compared to other region detectors.

When a new signature is acquired, it is compared to the stored panoramas from the a priori map. The panorama with the highest number of matches is selected. To improve the results and discard false matches, the essential matrix is computed and the outliers filtered. Finally, the panorama with the highest number of inliers is selected as the best match. In our approach images are acquired using a rotating conventional perspective camera. When a set of images covering 360 degrees is acquired, they are projected to cylindrical coordinates and the feature regions are extracted and described. The descriptors constellation is next constructed automatically. Hence, by using feature regions to construct the signature of a location, our approach is much more robust to occlusions and partial changes in the image than the approaches using global descriptors. This robustness is obtained because many individual regions are used for every signature of a location and, thus, if some of them disappear the constellation can still be recognized.

Nevertheless, combining different region detectors increases the computational time and memory requirements. For this reason we show that a re-ranking mechanism based on a global appearance-based similarity measure can be used to prioritize the most similar map nodes.

This framework gives us an interesting solution to the perceptual aliasing problem (one of the main difficulties when dealing with qualitative navigation and localization). Our approach is validated in real world experiments and is compared with other vision-based localization methods.

The remainder of this paper is organized as follows: In Section 2 we will first present a review of most recent related work on visual-based localization and navigation. Section 3 summarizes the different affine covariant region detectors and descriptors that we used in our work. Section 4 describes the localization procedure in details and the experimental design. Experimental results are reported in Section 5. And finally, Section 6 concludes our paper.

## 2 Related Work

Over the last decade, many appearance-based localization methods have been proposed (Owen and Nehmzow, 1998; Franz et al, 1998; Se et al, 2002). SIFT (Scale Invariant Feature Transform) features (Lowe, 2004) have been widely used for robot localization. The SIFT approach detects and extracts feature region descriptors that are invariant to illumination changes, image noise, rotation and scaling. In (Se et al, 2002), the authors used SIFT scale and orientation constraints so as to match stereo images; least-square procedure was used to obtain better localization results. The model designed by Andreasson et al (2005) combines SIFT algorithm for image matching and Monte-Carlo localization; their approach takes the properties of panoramic images into consideration. Another 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). Therefore, Silpa-Anan and R. Hartley (2004) construct an image map based on Harris Affine feature Regions with SIFT descriptors that is later used for robot localization.

The work proposed by Tapus and Siegwart (2006) defined fingerprints of places as generic descriptors of environment locations. Fingerprints of places are circular lists of features and they are represented as a sequence of characters where each character is an instance of a specific feature type. The authors used a multi-perceptional system and global low-level features (i.e., vertical edges, color blobs, and corners) are employed for localization.

Moreover, recently, several robot global localization methods similar to the one proposed in this paper have been presented. Booij et al (2007) build first an appearance graph from a set of training images recorded during exploration. The Differences of Gaussians (DoG) feature detector and the SIFT descriptor are used to find matches between omnidirectional images in the same manner as described in (Lowe, 2004), and the essential matrix relating every two images is computed with the 4-point algorithm with planar motion assumption in RANSAC. The similarity measure between each

pair of nodes of the map is the ratio between the inliers according to the essential matrix and the lowest number of features found in the two images. Appearance based navigation is performed by first localizing the robot in the map with a newly acquired image and then using Dijkstra's algorithm to find a path to the destination. Several navigation runs are successfully completed in an indoor environment even with occlusions caused by people walking close to the robot. Valgren and Lilienthal (2008) evaluate an approach focusing on visual outdoor localization across seasons using spherical images taken with a high resolution omnidirectional camera. Then, Upright Speeded Up Robust Features (U-SURF) (Bay et al, 2008), that are not invariant to rotation, are used to find matches between the images and the 4-point algorithm is used to compute the essential matrix. Indoor localization differs from outdoor localization in that, typically, distances to objects and walls is much shorter, and therefore the appearance of objects changes faster if one moves away from a reference point. Furthermore, indoor locations tend to have few texture and repetitive structure that complicates the data association problem, but they are positively less affected by environmental changes (e.g., time of the day; seasons).

(Cummins and Newman, 2008) proposed an approach that uses a probabilistic bail-out condition based on concentration inequalities. They have applied the bail-out test to accelerate an appearance-only SLAM system. Their work has been extensively tested in outdoor environments. Furthermore, the work presented by Angeli et al (2008) describes a new approach for global localization and loop detection based on the bag of words method.

### 3 Affine Covariant Region Detectors

An essential part of our approach is the extraction of discriminative information from a panoramic image so it can be recognized later under different viewing conditions. This information is extracted from the panoramic image using affine covariant region detectors. These detectors find regions in the image that can be identified even under severe changes in the point of view, illumination, and/or noise.

Recently (Mikolajczyk et al, 2005) reviewed the state of the art of affine covariant region detectors individually. In this review they concluded that using several region detectors at the same time could increase the number of matches and thus improve the results. Hence, based on their results, we have used all the combinations of the following three complementary affine covariant region detectors: (1) Harris-Affine, (2) Hessian-Affine, and (3) MSER (Maximally Stable Extremal Re-

gions), so as to increase the number of detected features and thus of potential matches. Examples of detected regions for the three region detectors can be seen in Fig. 1. These three region detectors have a good repeatability rate, a reasonable computational cost and they are briefly detailed below.

1. The Harris-Affine detector is an improvement of the widely used Harris corner detector. It first detects Harris corners in the scale-space with automatic scale selection using the approach proposed by Lindeberg (1998), and then estimates an elliptical affine covariant region around the detected Harris corners. The Harris corner detector finds corners in the image using the description of the gradient distribution in a local neighborhood provided by the second moment matrix:

$$M = \begin{bmatrix} I_x^2(x, \sigma) & I_x I_y(x, \sigma) \\ I_x I_y(x, \sigma) & I_y^2(x, \sigma) \end{bmatrix}, \quad (1)$$

where  $I(x, \sigma)$  is the derivative at position  $x$  of the image smoothed with a Gaussian kernel of scale  $\sigma$ . From this matrix, the cornerness of a point can be computed using the following equation:

$$R = \text{Det}(M) - k \text{Tr}(M)^2, \quad (2)$$

where  $k$  is a parameter usually set to 0.4. Local maxima of this function are found across the scales, and the approach proposed by Lindeberg is used to select the characteristic scales.

Next, 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 isotropy of the region is measured using the eigenvalue ratio of the second moment matrix:

$$Q = \frac{\lambda_{\min}(\mu)}{\lambda_{\max}(\mu)} \quad (3)$$

where  $Q$  varies from 1 for a perfect isotropic structure to 0, and  $\lambda_{\min}(\mu)$  and  $\lambda_{\max}(\mu)$  are the two eigenvalues of the second moment matrix of the selected region at the appropriate scale. For a detailed description of this algorithm, the interested reader is referred to (Mikolajczyk and Schmid, 2004).

2. The Hessian-Affine detector is similar to the Harris-Affine, but the detected regions are blobs instead of corners. The base points are detected in scale-space as the local maxima of the determinant of the Hessian matrix:

$$H = \begin{bmatrix} I_{xx}(x, \sigma) & I_{xy}(x, \sigma) \\ I_{xy}(x, \sigma) & I_{yy}(x, \sigma) \end{bmatrix}, \quad (4)$$



**Fig. 1** Example of regions for the three affine covariant region detectors, from left to right: Harris-Affine, Hessian-Affine and MSER.

- where  $I_{xx}$  is the second derivative at position  $x$  of the image smoothed with a Gaussian kernel of scale  $\sigma$ . The remainder of the procedure is the same as the Harris-Affine: base points are selected at their characteristic scales with the method proposed by Lindeberg and the affine shape of the region is found.
3. The Maximally Stable Extremal Regions (MSER) 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 the intensity of all the neighboring pixels of the region. Regions selected with this procedure may have an irregular shape, so the detected regions are approximated by an ellipse.

Because affine covariant regions must be compared, a common representation is necessary. Therefore all the regions detected with any method are normalized by mapping the detected elliptical area to a circle of a certain size. Once the affine covariant regions are detected and normalized, to reduce even more the effects caused by changes in the viewing conditions, these regions are characterized using a local descriptor. In our work, we have used Scale Invaraint Feature Transform (SIFT) (Lowe, 2004) and Gradient Location-Orientation Histogram (GLOH) (Mikolajczyk and Schmid, 2005). These two descriptors were found to be the best in a comparison of various state of the art local descriptors (Mikolajczyk and Schmid, 2005). The SIFT descriptor computes a 128 dimensional descriptor vector with the gradient orientations of a local region. In short, to construct the descriptor vector, the SIFT procedure divides the local region in 16 rectangular sub-regions and then, for every sub-region, it builds a histogram of 8 bins with the gradient orientations weighted with the gradient magnitude to suppress the flat areas with unstable orientations. The descriptor vector is obtained by concatenating the histograms for every sub-region.

The GLOH descriptor is similar to SIFT, with two main differences: the sub-regions are defined in a log-polar way, and the resulting descriptor vector has 272 dimensions but it is later reduced to 128 with a PCA.

These two descriptors are based on the same principle but with slightly different approaches. As they have no complementary properties, our objective in this com-

parison is to determine which one achieves the best performance. Therefore we have not combined them.

## 4 Experimental Design

The objective of the present work is twofold: On the one hand, we want to validate the proposed method for indoor global localization and, on the other hand, we target to experimentally determine if using different region detectors simultaneously improves significantly the localization results. Although successive images acquired by the robot while moving in the room could be used to incrementally refine the localization, in our experiment, we wanted to evaluate if combining different region detectors improves the robustness to viewpoint change for the presented global localization method and therefore, we have only considered the worst case scenario, where only one image per room is available to localize the robot.

### 4.1 Dataset description

The test-bed data used in this work consists of 17 sequences of panoramas from rooms in various buildings<sup>1</sup>. Each sequence consists of several panoramas acquired every 20 cm following a straight line predefined path. This type of sequences are useful to check the maximum distance at which a correct localization can be performed. In order to make the data set as general as possible, rooms with a wide range of characteristics have been selected (e.g., some sequences correspond to long and narrow corridors, while others have been taken in big hallways, large laboratories with repetitive patterns and others in smaller rooms such as individual offices). Panoramic images of the environment are shown (Figure 8) in the Annex. A short description of each sequence is given below:

- **iiia01** consists of 11 panoramas, and the sequence has been taken in a large robotics laboratory type of space.
- **iiia02** and **iiia03** contain 14 panoramas each, and have been taken at the conference room of the IIIA. In our experiments only the map node of iiia02 is used.
- **iiia04** is 19 panoramas long, and has been acquired in a long and narrow corridor.
- **iiia05** and **iiia06** have 25 and 21 panoramas, respectively. They have been taken in the library of the IIIA, the first one is from the library entrance

<sup>1</sup> The data-set can be downloaded from <http://www.iiia.csic.es/~aramisa>.



**Fig. 2** The camera and pan-tilt unit used to take the images.

and librarian desk, while the second is from a narrow corridor with book shelves. Both share the first panorama of **iiia05** as map node.

- **iiia07** is 19 panoramas long. This represents another section of the robotics laboratory, and corresponds to a small cubicle.
- **iiia08** is 10 panoramas long, and has been acquired in a small machinery room.
- **iiia09** has 21 panoramas that have been taken at the back entrance hall. This sequence has been taken in a tilted floor, which is a challenge for the 4-point algorithm, because of the flat world assumption.
- **iiia10** is 19 panoramas long and has been taken in the coffee room.
- **iiia11** has 21 panoramas and has been acquired in the entrance hall of the IIIA.
- **cvc01** is 21 panoramas long and corresponds to a long corridor of the CVC research center. As one of the corridor walls is made out of glass, the view field is wider than a normal corridor. However, direct sunlight affects the white balance of the image.
- **cvc02** is 21 panoramas long, and has been acquired in a large office with many desks.
- **cvc03** has 14 panoramas taken in a small office with just one working desk.
- **cvc04** has 22 panoramas and has been taken in a wide corridor with posters.
- **etse01** is the main hall of the engineering building and is 20 panoramas long.
- **etse02** has 21 panoramas and has been taken in a very wide corridor of the engineering building.

#### 4.2 Panorama Construction

Instead of using an omnidirectional camera, the panoramas have been constructed by stitching together multiple views taken from a Sony DFW-VL500 camera mounted on a Directed Perception PTU-46-70 pan-tilt

unit. The camera and pan-tilt unit can be seen in Fig. 2. In order to build a panorama using a rotating camera, we had to take into consideration the following: the image sequence employed to build the panorama must have a fixed optical center. Translations of the optical center would introduce motion parallax, making the image sequence inconsistent to build a panorama. However, if the objects in the scene are sufficiently far from the camera, small translations could be tolerated. The steps to stitch all the images in a panorama are the following:

1. The first step consists of projecting all the images of the sequence to a cylindrical surface. The points are mapped using the transformation from Cartesian to cylindrical coordinates:

$$\theta = \tan^{-1}\left(\frac{x}{f}\right), \quad v = \frac{y}{\sqrt{x^2 + f^2}} \quad (5)$$

where  $x$  and  $y$  are the position of the pixel,  $f$  is the focal distance measured in pixels and  $\theta$  and  $v$  are respectively the angular position and the height of the point in the cylinder. The cylinder radius is the focal length of the camera used to acquire the images, in this way the aspect ratio of the image is optimized (Shum and Szeliski, 1997). Taking this into account, the size of the panoramas is 5058x500.

2. Once all the images have been projected to cylindrical coordinates, the rotation between each pair of images must be estimated. In principle, only panning angles need to be recovered but, in practice, to correct vertical misalignment and camera twist, small vertical translations are allowed. Therefore, a displacement vector  $\Delta t = (t_x, t_y)$  is estimated for every pair of input images. The implemented method to compute  $\Delta t$  distinguishes between three situations:

*i* If sufficient feature points are found in the shared part of the images,  $\Delta t$  is computed by means of matches between pairs of feature points. To find the translation with most support among matches, and to exclude false matches and outliers, RANSAC is used.

*ii* In those cases where there is not enough texture in the images to extract sufficient feature points,  $\Delta t$  is computed looking for a peak in the normalized correlation between the Canny edges (Canny, 1986) of the two images. This method has the advantage over other correlation-based approaches of being independent of the illumination conditions and the vignetting effect (intensity decreases towards the edge of the image). In addition, as all the image is used, even with small amounts of texture a reliable translation can be



**Fig. 3** Intensity jumps between successive images caused by automatic camera gain. Applying linear blending solves the problem.

estimated. However, this technique is computationally more expensive than feature matching and is not invariant to rotations or other deformations in the image.

- iii If no texture exists at all and the above procedure fails, the only remaining solution is to compute the expected translation if the angular displacement  $\varphi$  (in radians) between the images is known:  $t_x = f\varphi$  and  $t_y = 0$
- 3. Due to automatic camera gain, vignetting or radial distortion, an intensity jump may appear between two images as can be seen in Figure 3. In this work we the most straightforward solution is taken, that consists in blending linearly every two consecutive images. This method produces results good enough for visualization purposes and is suitable for static scenes. However techniques such as multi-band blending and deghosting can be used (Shum and Szeliski, 1997; Brown and Lowe, 2003; Szeliski and Shum, 1997; Uyttendaele et al, 2001) to improve the result by eliminating stitching artifacts and dynamic objects that created *ghosts* in the panorama.

Although the panoramic images were constructed for validation purposes, the constellations of feature region descriptors were not extracted from them. Instead, the features from the original images projected to cylindrical coordinates where used. The reason for this is to avoid false regions introduced by possible new artifacts created during the stitching process. The panoramas built with the stitching method where all correctly constructed, even in the case of changes in lightning, reflections, multiple instances of objects or lack of tex-

ture. The sequences have been acquired in uncontrolled environments.

### 4.3 Panorama Matching

The region detectors and descriptors provided by Mikolajczyk et al (2005)<sup>2</sup> were used to extract the affine-covariant regions from the images and compute the SIFT descriptor vectors.

In this work, the procedure to compare two panoramas is relatively straightforward. First, matches are established as nearest neighbors between the feature descriptors of both panoramas using the Euclidean distance as similarity measure. Potentially false matches are rejected comparing the distance of the first and the second nearest neighbor in the same way as proposed by Lowe (2004). Additionally, reciprocal matching is used to filter even more false matches: if feature  $f_a$  from the first panorama matches feature  $f_b$  of the second panorama, but feature  $f_b$  does not match feature  $f_a$ , the match is discarded.

Next, the epipolar constraint between the panoramas is enforced by computing the essential matrix. The most straightforward way to automatically compute the essential matrix is using the normalized 8-point algorithm (Hartley and Zisserman, 2004). However, assuming that the robot will only move through flat surfaces, it is possible to use a simplified version where only 4 correspondences are necessary.

$$E = \begin{bmatrix} 0 & e_{12} & 0 \\ e_{21} & 0 & e_{23} \\ 0 & e_{32} & 0 \end{bmatrix} \quad (6)$$

Therefore, with a set of at least four correspondences of points of the form

$$p = [x, y, z] = [\sin(2\pi\tilde{x}), \tilde{y}, \cos(2\pi\tilde{x})] \quad (7)$$

where  $\tilde{x}$  and  $\tilde{y}$  are the normalized point coordinates in the planar panorama image, the following equations can be written:

$$\begin{bmatrix} y'_1 x_1 & x'_1 y_1 & z'_1 y_1 & y'_1 z_1 \\ \vdots & \vdots & \vdots & \vdots \\ y'_n x_n & x'_n y_n & z'_n y_n & y'_n z_n \end{bmatrix} \begin{bmatrix} e_{12} \\ e_{21} \\ e_{23} \\ e_{32} \end{bmatrix} = 0 \quad (8)$$

where  $(x_i, y_i, z_i)$  and  $(x'_i, y'_i, z'_i)$  is the  $i^{th}$  pair of corresponding points. As outliers may still be present among the matches, RANSAC is used to automatically compute the essential matrix with most support. Finally, the set of inlier feature matches that agree with the epipolar constraint is used as the evidence of the relation between the two panoramas.

<sup>2</sup> <http://www.robots.ox.ac.uk/~vgg/research/affine/>

Given the high dimensionality of the feature descriptors, matching is expensive in terms of computational cost even for a small set of nodes. An alternative to exhaustive matching is to use a global similarity measure to re-rank the map nodes and estimate the essential matrix only for the  $k$  top map nodes or, taking an *anytime* algorithm approach, until a node with a certain ratio of inliers is met. The global similarity measure should be fast to compute and exploit the differences between the map nodes to improve the re-ranking. We have applied the Vocabulary Tree proposed in Nister and Stewenius (2006) for object categorization to re-rank the map nodes for a new query image as it fulfilled both requirements. In short, this method constructs a visual vocabulary tree of feature descriptors applying hierarchical  $k$ -means on a training dataset. Next, images are described as a normalized histogram of *visual word* counts. To give more emphasis to discriminative *visual words*, they are weighted using a term frequency-inverse document frequency (tf-idf) approach. Finally, training set images can be re-ranked according to its Euclidean distance to the new image signature.

Although the presented method has a very good performance in our experiments, it is time-consuming to acquire a panorama rotating a pan-tilt unit every time a localization has to be performed. Instead, we evaluate the decrease in performance using uniquely a normal planar perspective of 45° field of view to localize the robot.

The simplest way to decide the corresponding node is by the maximum number of matches after computing the essential matrix (Valgren and Lilienthal, 2008; Ramisa et al, 2008). Alternatively, the ratio between the number of matches and the lowest number of keypoints of the two images(Booij et al, 2007). Experimentally, we did not find much difference between both approaches and therefore we have retained the first one.

## 5 Experiments

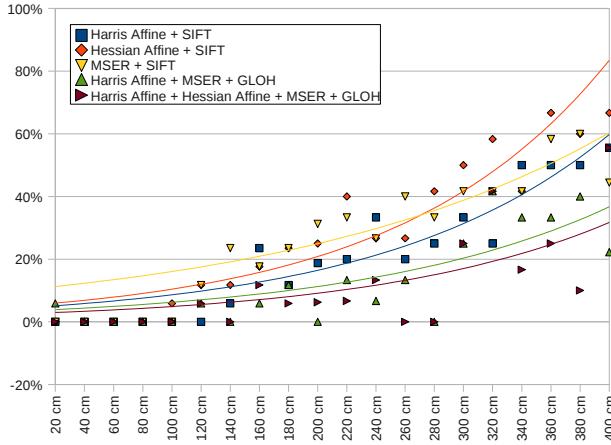
In order to achieve our two objectives, we tested all possible combinations of the three selected region detectors with two different descriptors. Table 1, shows the average percentage of correctly classified test panoramas for each combination. Results are provided using the 8-point algorithm, the 4-point algorithm and also the later with reciprocal matches. From the results illustrated in Table 1 for the 4 point algorithm with reciprocal matches, it can be seen that by reducing the number of false matches with this technique, the performance is substantially improved. Therefore from now on, we only show the results obtained with this last technique.

**Table 1** Average percentage of correctly localized panoramas (acl) across all sequences and standard deviation (std). For convenience we have labeled M: MSER, HA: Harris-Affine, HE: Hessian-Affine, S: SIFT, G: GLOH.

Combination	8 points algorithm		4 points algorithm		4 points and recipr. match	
	acl	std	acl	std	acl	std
HA+S	74%	23%	69%	23%	82%	22%
HA+G	70%	21%	73%	24%	81%	21%
HE+S	58%	24%	73%	26%	75%	25%
HE+G	63%	26%	65%	27%	74%	26%
M+S	62%	28%	78%	18%	76%	23%
M+G	61%	29%	69%	23%	74%	26%
HA+HE+S	64%	15%	78%	19%	86%	14%
HA+HE+G	67%	14%	79%	21%	87%	16%
M+HE+S	56%	23%	75%	23%	87%	15%
M+HE+G	60%	23%	78%	18%	88%	14%
M+HA+S	65%	21%	79%	19%	86%	14%
M+HA+G	70%	25%	79%	19%	88%	11%
M+HA+HE+S	62%	16%	82%	19%	89%	11%
M+HA+HE+G	64%	20%	82%	19%	90%	11%

Standard deviation is also provided in order to assess the stability of combinations along the different sequences. Not much difference is observed among the descriptors GLOH and SIFT, which performed similarly in all cases. Looking at the feature detectors individually, the best results have been obtained by Harris Affine, while Hessian Affine and MSER had a similar performance. Overall, the combinations of detectors outperformed the individual detectors. The best performance in the localization test has been achieved by the combination of the three detectors, which classified correctly 90% of the panoramas. This performance is mainly due to their good complementarity. Furthermore, in Figure 4 the average performance of two selected combinations is compared to the standalone detectors as a function of the distance to the map node. As can be seen, combinations cope better with changes in point of view than individual detectors. Sequences acquired in large rooms typically achieved a good performance no matter the combination used. However, small rooms and specially long and narrow corridors seem to be more difficult environments, even if they are well textured. This can be explained because the distance between the robot and the perceived objects is short and, therefore, the objects' appearance changes rapidly resulting in an unreliable matching in the lateral regions of the panorama.

Some particularly difficult sequences have been cvc01, iiiia04, iiiia06 and iiiia09. Table 2 shows the results with these sequences. As we can see, the performance is notably increased by combining different detectors. On average, standalone detectors achieved around 55%, while combinations increased to around 81% in these environments. The sequence iiiia04 is a specially long and



**Fig. 4** Percentage of incorrectly classified test panoramas as a function of the distance to the map node. The exponential regression of the data points is also provided for clarity.

**Table 2** Average percentage of correctly localized panoramas for some interesting sequences. The naming convention is the same as in Table 1.

Combination	cvc01	iiia04	iiia06	iiia09
HA+S	35%	42%	75%	75%
HA+G	35%	47%	70%	70%
HE+S	20%	47%	75%	45%
HE+G	20%	53%	80%	30%
M+S	85%	42%	30%	65%
M+G	95%	53%	35%	35%
HA+HE+S	80%	84%	70%	95%
HA+HE+G	45%	89%	75%	85%
M+HE+S	90%	84%	80%	65%
M+HE+G	90%	84%	90%	60%
M+HA+S	90%	79%	70%	85%
M+HA+G	90%	89%	65%	80%
M+HA+HE+S	80%	95%	75%	80%
M+HA+HE+G	75%	100%	90%	75%

narrow corridor with very few texture; it is interesting to notice that in this case the combination of all feature types achieved 100% correct classification. Another notable finding is the extremely good performance of MSER on cvc01 when compared to the other detectors.

Most of the similar approaches to global localization (e.g. Booij et al (2007)) use feature detectors only invariant to scale but not affine covariant, mainly because of its more expensive computational cost. For comparability, we have evaluated the performance of the Difference of Gaussians detector (Lowe, 2004). This method uses as initial points the local maxima of the Differences of Gaussians (DoG), defines a circular region around these initial points and finally SIFT is used to describe the selected regions. For our tests we used the implementation provided by Lowe<sup>3</sup>. On average, using points detected with the DoG and SIFT, the average correct

location was 72%. However, it had an irregular performance depending on the environment type (27% standard deviation), with perfect results in large rooms, but very poor results in narrow corridors and small rooms. This was an expected outcome as this detector is less resistant to viewpoint changes.

In terms of computational complexity, the most expensive step of the approach is clearly the bidirectional descriptor matching as can be seen in Table 3. These computational times have been obtained with a C++ implementation of the method running in a Linux Pentium 4 at 3.0 GHz computer with 2Gb of RAM.

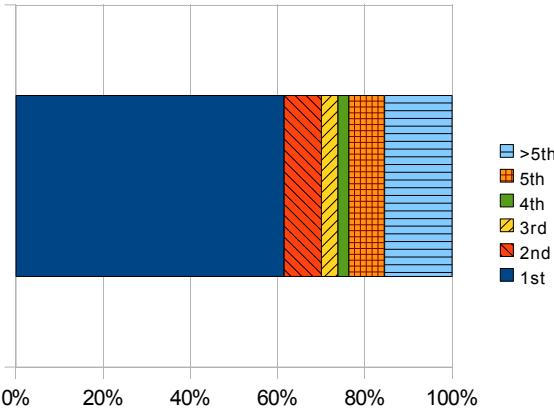
**Table 3** Average feature matching and RANSAC time per map node. It is important to notice the difference in time scale.

Combination	Matching (seconds)	RANSAC (milliseconds)
HA+S	4,31	3,046
HA+G	4,29	2,597
HE+S	2,87	3,016
HE+G	2,88	2,631
M+S	1,24	2,920
M+G	1,24	2,310
HA+HE+S	7,16	6,625
HA+HE+G	7,16	5,401
M+HE+S	4,11	5,827
M+HE+G	4,11	5,361
M+HA+S	5,51	6,682
M+HA+G	5,51	5,382
M+HA+HE+S	8,44	1,3941
M+HA+HE+G	8,47	1,0815

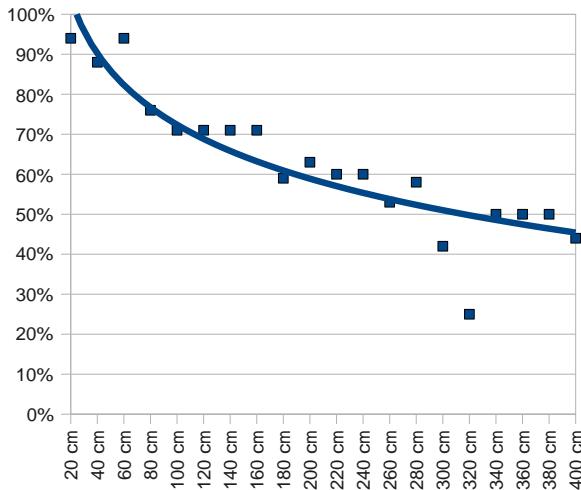
### 5.1 Re-ranking of map nodes

As explained in Section 4.3, the global appearance based image similarity measure from Nister and Stewenius has been used to re-rank the map nodes and prioritize those that appear more similar. We have build the vocabulary tree with Harris Affine features. When used for object classification, this type of approach requires at least tens of training images in order to correctly determine the class of a novel object instance. However, we only used the map nodes to train both the vocabulary tree and the classifier. This gives only one training instance for each class. Despite so limited training data, the approach achieved the notable overall result of re-ranking the correct node in the first position for 62% of the query panoramas, and among the top five nodes 85% of times as can be seen in Figure 5. More detailed results of this re-ranking experiment are in Figure 7, where the performance is shown for each individual sequence.

<sup>3</sup> <http://www.cs.ubc.ca/~lowe/keypoints/>



**Fig. 5** Position of the correct map node after re-ranking using the vocabulary tree.

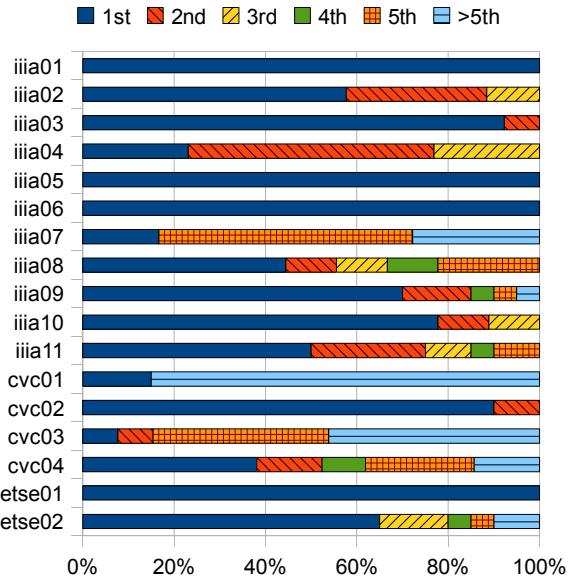


**Fig. 6** Ratio of query images with the correct node re-ranked at the top position against distance to first panorama of the sequence. The logarithmic regression curve is also shown.

As expected, the percentage of times the correct map node is re-ranked at the top position decreases as distance to the query panorama increases (see Figure 6).

## 5.2 Localization with 45° FOV images

Constructing a panoramic image with a rotating camera on a pan-tilt unit is a time-consuming step that requires the robot to stay in a fixed position during the acquisition. In order to assess the decrease in performance that would cause using just a single conventional image to localize the robot we have done the following experiment: For every test panorama, a random area that spans 45° and has at least 100 features is extracted and matched to the map nodes. This procedure is repeated for every test panorama. After a 10 fold



**Fig. 7** Position of the correct map node after re-ranking using the vocabulary tree per sequence.

experiment with all test panoramas, the average number of correct localizations was 73% using Harris Affine combined with MSER and the GLOH descriptor. This result is good considering how limited the field of view is. In addition to the time saved in image acquisition, the matching time is reduced almost one order of magnitude on average.

## 6 Conclusions and Discussion

In this work we have proposed and evaluated a signature to characterize places that can be used for global localization. This signature consists of a constellation of feature descriptors, computed from affine-covariant regions, extracted from a panoramic image, that has been acquired in the place we want to add to the map. Later, these signatures are compared to the constellation extracted from a new panoramic image using geometric constraints, and the most similar signature is selected as the current location. To compare the different signatures, the 4-point algorithm with RANSAC to reject false matches is used. Combinations of feature detectors have been shown to perform best if combined with adequate mechanisms, such as reciprocal matching or distance to the second nearest neighbor, to reject incorrect pairings of features before computing the essential matrix.

Regarding the validation of the global localization schema, the results obtained show that by using the combination of different feature detectors, a room can be reliably recognized in indoor environments from a

distance of up to 4 meters from the point where the reference panorama was obtained. The best results (90% correct localizations) were achieved by combining all three detectors.

Moreover, we have also compared the results of our proposed affine-covariant region detectors approach with the scale-invariant region detectors methodology proposed in Lowe (2004), widely used in robot navigation, and showed that the affine-covariant regions outperformed Lowe's scale-invariant method.

In order to speed-up the otherwise very expensive descriptor matching phase, a global similarity technique usually employed for object recognition, the vocabulary tree of Nister and Stewenius (2006), has been effectively applied to re-rank the map nodes for a given query panorama and save most of the computation time.

Furthermore, we tested how the performance degrades if only a conventional perspective image is used instead of an omnidirectional image. Results of a 10 fold experiment with random 45° sections (with a minimum amount of texture) from all the test panoramas showed a surprisingly good performance.

## References

- Andreasson H, Treptow A, Duckett T (2005) Localization for mobile robots using panoramic vision, local features and particle filters. In: In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'05), Barcelona, Spain
- Angeli A, Filliat D, Doncieux S, Meyer A J (2008) A fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions On Robotics, Special Issue on Visual SLAM*
- Bay H, Ess A, Tuytelaars T, Gool LV (2008) Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)* 110(3):346–359
- Beeson P, Jong K N, Kuipers B (2005) Towards autonomous topological place detection using the extended voronoi graph. In: IEEE International Conference on Robotics and Automaton (ICRA), Barcelona, Spain, pp 4373–4379
- Booij O, Terwijn B, Zivkovic Z, Kroese B (2007) Navigation using an appearance based topological map. In: IEEE International Conference on Robotics and Automation (ICRA), pp 3927–3932
- Brown M, Lowe D (2003) Recognising panoramas. In: ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, p 1218
- Canny J (1986) A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6):679–698
- Castellanos A J, Tardos D J (1999) Mobile Robot Localization and Map Building: Multisensor Fusion Approach. Kluwer Academic Publisher
- Choset H, Nagatani K (2001) Topological simultaneous localization and mapping (SLAM): Toward exact localization without explicit localization. *IEEE Transactions On Robotics and Automation* 17(2):125–137
- Cummins M, Newman P (2008) Accelerated appearance-only SLAM. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pp 1828–1833
- Dissanayake M, Newman P, Clark S M, Durrant-Whyte H, Csorba M (2001) A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation* 17(3):229–241
- Franz M, Schlkopf O, Mallot B, Blthoff A H (1998) Learning view graphs for robot navigation. *Autonomous Robots* 5:111–125
- Hartley R, Zisserman A (2004) Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, ISBN: 0521540518
- Lindeberg T (1998) Feature detection with automatic scale selection. *International Journal of Computer Vision* 30(2):79–116
- Lowe D (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110
- Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: In Proceedings of the British Machine Vision Conference (BMVC'02), Cardiff, UK
- Mikolajczyk K, Schmid C (2004) Scale & affine invariant interest point detectors. *Int J Comput Vision* 60(1):63–86
- Mikolajczyk K, 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 LV Tand Gool (2005) A comparison of affine region detectors. *International Journal of Computer Vision* 65(2):43–72
- Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. *Conf Computer Vision and Pattern Recognition* 2:2161–2168
- Owen C, Nehmzow U (1998) Landmark-based navigation for a mobile robot. In: From Animals to Animats: Fifth International Conference on Simulation of Adaptive Behavior (SAB), MIT Press, Cambridge, MA, pp 240–245
- Ramisa A, Tapus A, Lopez de Mantaras R, Toledo R (2008) Mobile robot localization using panoramic vi-

- sion and combinations of feature region detectors. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pp 538–543
- Se S, Lowe D, Little J (2002) Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research (IJRR)* 21(8):735–758
- Shum H, Szeliski R (1997) Panoramic image mosaics. Tech. Rep. MSR-TR-97-23, Microsoft Research
- Silpa-Anan C, R Hartley A (2004) Localization using an image-map. In: Proceedings of the 2004 Australasian Conference on Robotics and Automation, Canberra, Australia
- Szeliski R, Shum HY (1997) Creating full view panoramic image mosaics and environment maps. In: SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, vol 31, pp 251–258
- Tapus A, Siegwart R (2006) A cognitive modeling of space using fingerprints of places for mobile robot navigation. In: Proceedings IEEE International Conference on Robotics and Automation (ICRA), Orlando, USA, pp 1188–1193
- Thrun S (1998) Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence* 99(1):21–71
- Thrun S (2000) Probabilistic algorithms in robotics. *Artificial Intelligence Magazine* 21:93–109
- Uyttendaele M, Eden A, Szeliski R (2001) Eliminating Ghosting and Exposure Artifacts in Image Mosaics. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, vol 2
- Valgren C, Lilienthal A (2008) Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pp 1856–1861

## Annex



**Fig. 8** Panorama nodes in the same order as described in the text.