Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Arturo Gil · Oscar Martinez Mozos · Monica Ballesta · Oscar Reinoso

A Comparative Evaluation of Interest Point Detectors and Local Descriptors for Visual SLAM

Received: date / Accepted: date

Abstract In this paper we compare the behavior of different interest points detectors and descriptors under the conditions needed to be used as landmarks in vision-based simultaneous localization and mapping (SLAM). We evaluate the repeatability of the detectors, as well as the invariance and distinctiveness of the descriptors, under different perceptual conditions using sequences of images representing planar objects as well as 3D scenes. We believe that this information will be useful when selecting an appropriate landmark detector and descriptor for visual SLAM.

Keywords Interest point detectors · Local descriptors · Visual landmarks · Visual SLAM

1 Introduction

Acquiring maps of the environment is a fundamental task for autonomous mobile robots, since the maps are required in different higher level tasks, such as navigation and localization. In consequence, the problem of simultaneous localization and mapping (SLAM) has received significant attention during the last decades. The SLAM problem considers the situation in which an autonomous mobile robot moves through an unknown space and incrementally builds a map of this environment while simultaneously uses this map to compute its absolute location. It is an inherently hard problem because noise in the estimate of the robot pose leads to noise in the estimate of the map and *vice-versa*. Typical SLAM approaches use laser range sensors to build maps in two and three dimensions (e.g., [8,9,3,5,28]). However, in recent years the interest on using cameras as sensors in SLAM has increased. These approaches are normally denoted as visual SLAM. The main reason for this interest stems from the fact that cameras offer a higher amount of information and are less expensive than lasers. Moreover, they can provide 3D information when stereo systems are used.

The underlying SLAM algorithms used in laser and vision are basically the same. However, the main problem in visual SLAM is the selection of adequate landmarks. That means that it is unclear which are the best visual landmarks. In the case of laser-based SLAM, different landmarks have been proposed with demonstrated good results, such as lines or other geometrical features extracted from the range scans [1,21].

Common approaches in visual SLAM are feature-based. In this case, a set of significant points in the environment are used as landmarks. Mainly, two steps must be distinguished in the selection of visual landmarks. The first step involves the detection of interest points in the images that can be used as landmarks. The points should be detected at several distances and viewing angles, since they will be observed by the robot from different poses in the environment. This situation is represented in figure 1, where the same points in the space are observed by the robot from different poses in the environment. At a second step the interest points are described by a feature vector which is computed using local image information. This descriptor is used in the data association problem, that is, when the robot has to decide whether the current observation corresponds to one of the landmarks in the map or to a new one. Typically, when the robot traverses previously explored places, it re-observes landmarks seen before. In this case, if the current observations are correctly associated with the visual landmarks, the robot will be able to find its location with respect to these landmarks, thus reducing

This work has been supported by the EU under project CoSy FP6-004250-IPCoSy, by the Spanish Government under projects DPI2004-07433-C02-01 and CICYT DPI2007-61197 and by the Generalitat Valenciana under grant BFPI/2007/096.

A. Gil, M. Ballesta, O. Reinoso
Dept. of Industrial Systems Engineering, Miguel Hernández University (Spain)
E-mail: {arturo.gil|m.ballesta|o.reinoso}@umh.es

O. Martinez Mozos
Dept. of Computer Science, University of Freiburg (Germany)
E-mail: omartine@informatik.uni-freiburg.de

the error in its pose. If the observations cannot be correctly associated to the landmarks in the map, the map will be inconsistent. To sum up, the data association is a fundamental part of the SLAM process, since wrong data associations will produce incorrect maps.

Nowadays, a great variety of detection and description methods have been proposed in the context of visual SLAM. In our opinion, there exists no consensus on this matter and this means that the question of which interest point detector and descriptor is more suitable for visual SLAM is still open. This situation motivated the work presented here. The problem of finding the best detector and descriptor can be tackled in two different manners:

1. At the SLAM stage: This involves using a particular interest point detector, a description method and building a map using a SLAM technique. Finally, the quality of the results should be analyzed to evaluate different detectors and descriptors. In some cases, the path of the robot estimated using a visual SLAM approach is compared to the estimation using laser range data [7,29], since laser-based SLAM usually produces more accurate results. However, the visual SLAM results greatly depend on the SLAM algorithm used and several implementation details. In consequence, the results obtained with this procedure may not be general.
2. At the landmark extraction stage: In this case the focus is on measuring the quality of the detectors and descriptors in terms of stability and robustness under image transformation, which are the properties needed in a visual landmark. This evaluation is independent of the SLAM algorithm and is thus more general. This is the approach we consider.

In this paper we compare the behavior of different interest points detectors and descriptors under the conditions needed to be used as landmarks in vision-based simultaneous localization and mapping (SLAM). We evaluate the repeatability of different interest point detectors, as well as the invariance and distinctiveness of several description methods, under changes in scale, viewpoint and illumination. In order to do this we use sequences of images representing planar objects as well as 3D scenes. We have divided the problem of finding the most suitable detector and descriptor for visual SLAM into two parts. First, we concentrate on the selection of the most suitable interest point detector. Second, we analyze several description methods.

In the case of the interest point detectors, we analyze the repeatability of the points in a set of images obtained from the same scene when viewed at different distances, angles and light conditions. This situation typically occurs in visual SLAM applications when the robot explores the environment and observes the same points from different poses. In order to do this we analyze whether a point extracted in a reference image is detected in the remaining images in the sequence.

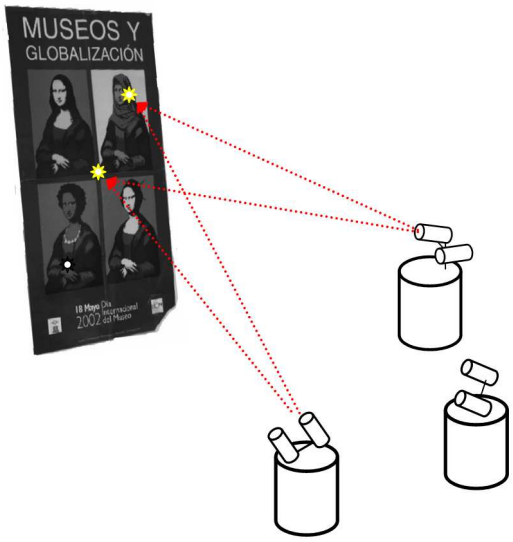


Fig. 1 Some points in the scene are observed by the robot from different poses

In order to evaluate the descriptors we have performed a series of experiments in a matching context. First, by means of *precision* and *recall*. Second, we apply clustering measurements [27] to estimate how well the descriptors representing the same landmark are grouped in the different descriptor subspaces. These measurements can be used to decide which descriptor has better separability properties.

An extensive set of experiments has been carried out with sequences of real indoor environment images. The sequences include significant changes in scale, viewing angle, as well as illumination changes. We believe that these results would help the selection of visual landmarks for SLAM applications.

We consider the case in which SLAM is performed with a mobile robot that carries a camera. The camera maintains always the same position with respect to the robot and it rotates only over the vertical axis. This assumption has been applied to many SLAM approaches [7, 24,29]. We additionally assume that the same landmark, when rotated, converts into a different one. This assumption is based on the fact that sometimes the same landmark indicates different information when rotated, for example, a panel with different arrows indicating different directions.

The remainder of the paper is organized as follows: In Section 2 we introduce related work. Next, Section 3 presents the set of interest point detectors evaluated in this work. Following, Section 4 describes the different local image descriptors analysed here. In Section 5 the methods to evaluate the detection methods are presented. Section 6 deals with the techniques used to evaluate the performance of the descriptors. Finally, in Section 7 we present experimental results.

2 Related Work

To the present days, different combinations of detectors and descriptors have been used for mapping and localization using monocular or stereo vision. For example, in the context of monocular SLAM, Davison and Murray [4] used the Harris corner detector to find significant points in images and described them using a window patch centered at the detected points. Lowe [16] presented the SIFT transform, which combines an interest point detector and a description method, which was initially applied to object recognition applications [15]. Later, Se *et al.* [24] used SIFT features as landmarks in the 3D space. Little *et al.* [14] and Gil *et al.* [7] additionally tracked the detected SIFT points to keep the most robust ones. Jensfelt *et al.* [11] use a rotationally variant version of SIFT in combination with a Harris-Laplace detector for monocular SLAM. Recently, Herbert *et al.* [2] presented the SURF features, which also proposes an interest point detector in combination with a descriptor. Lately, Murillo *et al.* used the SURF features in localization tasks [22] using omnidirectional images.

In the context of matching and recognition, many authors have presented their works evaluating several interest point detectors and descriptors. For example, Schmid *et al.* [23] evaluate a collection of detectors by measuring the quality of these features for tasks like image matching, object recognition and 3D reconstruction. However, the mentioned work does not consider the interest point detectors that are most frequently used nowadays in visual SLAM.

Several comparative studies of local region detectors and descriptors have been presented so far. For instance, Mikolajczyk *et al.* [20] present a comparison of several local affine region detectors. Similarly, Fraundorfer and Bischof also present in [6] an evaluation of detectors, but introducing a new tracking method in order to use non-planar scenes. On the other hand, Mikolajczyk and Schmid [19] use different detectors to extract affine invariant regions, but they focus on the comparison of different description methods. A set of local descriptors are evaluated using a criterion based on the number of correct and false matches between pairs of images. All these previous approaches perform the evaluation of detection and description methods using pairs of images and analyze different imaging conditions.

In contrast to the previous approaches, we present a novel approach in which the stability and invariability of the interest points are evaluated along a sequence of images obtained from the same scene, emulating the different viewpoints from which the robot observes a point while performing visual SLAM tasks. The total variation in the camera position from the first image to the last image in the sequences is significant, as typically occurs in visual SLAM. Instead of having pairs of correspondent points in images, we consider clusters of points. A cluster is composed of a point which has been observed from different viewpoints in some images of the sequence and its associated descriptor in each frame. Furthermore, we evaluate separately the detectors and descriptors under the particular conditions of visual SLAM.

3 Interest Point Detectors

In the following, we present five different interest point detectors that are suitable to extract visual landmarks in visual SLAM applications.

Harris Corner Detector: The Harris Corner Detector [10] is one of the most widely used interest point detectors. For each point in the image the eigenvalues of the second moment matrix are computed. The associated eigenvectors represent two perpendicular directions of greatest change. A corner is characterized as a point with two large eigenvalues, corresponding to a strong change in both directions. In [4] Harris points are used to extract visual landmarks in monocular SLAM.

Harris-Laplace: The interest points extracted by Harris-Laplace are detected by a scale adapted Harris function and selected in scale-space by the Laplacian operator. This detector has previously been used in image indexing applications [18], as well as in bearing only visual SLAM [11].

SUSAN: The Smallest Univalue Segment Assimilating Nucleus (SUSAN) is an approach to low level image processing [26]. It works by placing a circular mask over the pixel in the image to be evaluated. The decision whether a point is a corner or not depends on the number of pixels similar to the central that lie inside the mask. SUSAN has been traditionally used in object recognition applications.

SIFT: The Scale-Invariant Feature Transform (SIFT) is an algorithm that detects distinctive points in images by means of a difference of Gaussian function (DoG) applied in scale space [16]. The points are selected as local extrema of the DoG function. Next, a descriptor is computed for each detected point, based on local image information at the characteristic scale. The algorithm was initially presented by Lowe [15] and used in object recognition tasks. Lately, it has been used in visual SLAM applications [25, 29, 7]. In this work we separate the detection process from the description, thus when used as a detector, points are extracted using a DoG function.

SURF: The Speeded Up Robust Features (SURF) were introduced by Bay *et al.* [2]. According to its authors [2] SURF features are said to outperform existing methods with respect to repeatability, robustness and distinctiveness of the descriptors. The detection method is based on the Hessian matrix and relies on integral images to

reduce the computation time. As with SIFT features, we concentrate only on the detected points when used as an interest point detector.

MSER: Maximally Stable Extremal Regions were introduced by Matas *et al.* [17]. The regions are extracted with a method similar to the watershed segmentation algorithm. The method has been tested in wide-baseline stereo images with significant scale and perspective differences. This fact encourages to evaluate this detection method, since, to the best of our knowledge, MSER has not yet been applied to the visual SLAM problem.

Kadir: Kadir's detector measures the entropy of pixel intensity histograms computed for elliptical regions in order to find salient points [12].

4 Local Descriptors

In this work we have evaluated descriptors according to the requirements for visual SLAM. Some of the methods have been used previously in the context of visual SLAM. Next, we list the set of descriptors that have been studied.

SIFT: The SIFT transform assigns a global orientation to each point based on local image gradient directions. Next, a descriptor is computed based on orientation histograms at a 4×4 subregion around the interest point, resulting in a 128-dimensional vector [16]. To obtain illumination invariance, the descriptor is normalized by the square root of the sum of squared components.

GLOH: Gradient location-orientation histogram is an extension of the SIFT descriptor [19], designed to increase its robustness and distinctiveness. In order to compute the GLOH descriptor, the SIFT descriptor is computed at a log-polar location grid with three bins in radial direction and 8 bins in angular direction. Initially, the descriptor is of size 272 but is reduced to a final length of 128 by means of PCA analysis. According to [19] the GLOH descriptor outperforms the SIFT descriptor in several tests.

SURF: The SURF descriptor represents a distribution of Haar-wavelet responses within the interest point neighbourhood and makes an efficient use of integral images. Three different versions of the descriptor have been studied: the standard SURF descriptor, which has a dimension of 64, the extended version (E-SURF) with 128 elements and the upright version (U-SURF). The U-SURF version is not invariant to rotation and has a length of 64 elements [2].

Gray level patch: This method describes each landmark using the gray level values at a 10×10 subregion around the interest point. This description has been used in [4] in the context of monocular SLAM.

Orientation Histograms: The computation of orientation histograms is based on the gradient image. For each pixel a module and an orientation are computed. The orientation is divided in a number of bins and a histogram is formed with the values of the module. In [13] orientation histograms are applied in mobile robot navigation.

Zernike Moments: The moment formulation of Zernike polynomials [30] appears to be one of the most popular in terms of noise resilience, information redundancy and reconstruction capability. Complex Zernike moments are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc.

5 Evaluation Methods for Interest Point Detectors

In order to evaluate the different interest point detectors introduced in Section 3, we use series of images obtained from the same scene under different scales, viewpoints and illumination. In Fig. 2 we show example images extracted from each sequence in different indoor environments. We are interested in evaluating which detector allows us to extract the same points in the space when the scene is observed at different distances and angles.

For each image in a sequence, we first extract the interest points using the methods explained in Section 3. Next, we would like to evaluate if a point detected in one of the images appears in the other images in the sequence. To do this we have implemented two different algorithms, suitable for 2D and 3D scenes. We consider that a 2D sequence is constituted by a planar object (e.g. a poster), whereas a 3D scene contains both planar and non-planar objects. In both cases, the sequences were obtained from a typical laboratory. The evaluation was performed independently for 2D and 3D scenes, since the same detectors behave differently in each situation.

5.1 Matching the Position of Interest Points

The algorithms described in this section are capable of predicting the position of the detected points along the sequences by only establishing geometric constraints, thus allowing us to perform the matching of the points independently of the description method. The points detected in the scene with a particular position and orientation of the camera are searched in the other images of the sequence. Ideally, the same points should be detected in every image of the sequence. However, due to

Fig. 2 Some example images from the sequences used in the experiments. Fig. 2(a) and (b) show two images from a scale changing sequence (2D). Fig. 2(c) and (d) present two examples from a viewpoint changing sequence (2D). In Fig. 2(e) and (f) examples from a 3D scale changing sequence are presented. Fig. 2(g) and (h) show two more examples with changes in scale. In Fig. 2(i) and (j) two images with changes in illumination are shown.

image variations, some points tend to disappear in some frames. The performance of the matching method is not the scope of this paper; on the contrary, we employ it as a tool in order to obtain the correspondences between the detected points in the images. These correspondences are used in the evaluation methods explained in Section 5.2.

In the 2D case, we used a method based on the homography matrix as in [23]. In this case, given a point X in 3D space, we assume that this point projects at position $x_1 = K_1 X$ in image I_1 and at position $x_i = K_i X$ in image I_i, where K_1 and K_i are projection matrices. If we suppose that the point X is detected in both images, then

$$x_i = H_{1i} x_1, \text{ with } H_{1i} = K_i K_1^{-1} \ . \quad (1)$$

The homography matrix H_{1i} can be computed by selecting manually four correspondences of coplanar points between images 1 and i. In consequence, in order to find correspondences between the points found in image 1 and j, we proceed in this manner: first, we compute the homography matrix H_{1j} by selecting manually four correspondences between images 1 and j. Next, we can predict the position of point x_1 in the first image, which is computed as: $x_j = H_{1j} x_1$. If the predicted position lies at a distance below ε pixels from an interest point detected in the image j, then we consider that both interest points are correspondent and the interest point is successfully found. If no interest point lies in the neighborhood of the predicted point, then we consider that the point is not detected. In this case, we still look for the same point x_1 in the remaining images of the sequence. The process is repeated for every point detected in the first image in the sequence. This method has been applied to sequences of images containing planar objects, such as posters.

In the case of 3D scenarios, we have implemented a tracking method based on the fundamental matrix. The fundamental matrix is a 3×3 dimensional matrix with rank 2 which relates the corresponding points between two stereo images. Given a point x_1 in image I_1, the fundamental matrix F computes the epipolar line on the second image I_2 where the corresponding point x'_1 must lie. The epipolar line is computed as $l' = F x_1$ (see Figure 3). In consequence, two corresponding points will satisfy the following equation

$$x_i'^T F x_i = 0 \ . \quad (2)$$

For each point x_i the correspondent point in the other image x'_i is selected as the one with smallest distance to the epipolar line. This strategy can cause a large number of false correspondences, since several points can lie next to a line. To restrict the correspondences, the point x'_i must lie inside a 10×10 pixel window centered at the point x_i. This is valid for the sequences used in the experiments, since the camera moves slightly between consecutive images. In addition the correspondences were checked manually in order to avoid false matches.

The computation of the matrix F is done in two steps. First, seven correspondences between each pair of consecutive images are selected manually, which allows us to compute a fundamental matrix F. Second, using this fundamental matrix F we find a set of preliminary correspondences that are used as input for the computation of a second fundamental matrix F'. In this second step, the fundamental matrix is computed using a RANSAC approach [31], which results in a more accurate matrix F', that permits to find the final correspondences with more precision. Fig. 4 shows examples of tracked points along different sequences.

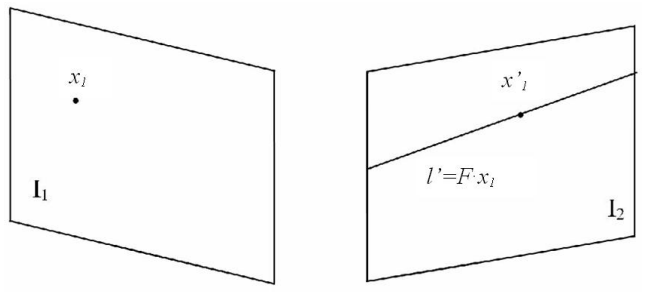


Fig. 3 The point x_1' is the corresponding point of x_1 in the image I_2. This point lies in the epipolar line l' computed with the fundamental matrix F.

In both 2D and 3D scenes we track the points extracted with any of the previously exposed algorithms. Note that the tracking of the points is based only on geometric restrictions and does not use any visual description of the points. Thus, the tracking is independent of the description and we can study the detection and description problems separately.

An example of a tracking using one of these methods is shown in Figure 4. First, Harris points were extracted at each image. Second, the correspondence of the points is found along the sequence. When a point is not detected in one of the images, we still look for it in the remaining images of the sequence. In Figure 4 the points that could be tracked along the whole sequence are shown.

5.2 Evaluation Criteria

To evaluate the different interest point detectors we study their repeatability under changes in scale and viewpoint. Once the correspondence of the points in a sequence has been performed, we can define the repeatability rate rr_i in the frame i of a sequence as:

$$rr_i = \frac{np_i}{np_r} \ , \tag{3}$$

where np_i is the number of interest points found in image i in the sequence and np_r is the number of interest points detected in the reference image of the sequence. This definition is similar to the employed in [23], but extended to the case where the correspondence is made across a set of images and not only a pair. A perfect detector would detect the same points in the first and the last frame, i.e. $rr_i = 1$ for every frame. However, as we will see in the experiments, we normally observe a decreasing tendency in rr_i, indicating that some of the points observed in the first frame are lost in subsequent frames.

When the robot explores the environment, it is desirable to obtain a set of visual landmarks that are robust and stable. In order to do this a common technique consists of tracking each point across consecutive frames and include only the points that can be detected in p consecutive frames [14,7]. As a result, the number of landmarks in the map is reduced and also the complexity of the SLAM problem. Taking into account this requirement we analyze for how many frames we should track a landmark before integrating it in the map. We use the following conditional probability:

$$P(f_j|f_i) = \frac{np_{1:j}}{np_{1:i}}, \quad j \geq i, \tag{4}$$

where, now, $np_{1:k}$ is the number of points that could be tracked from the first frame until frame k in the sequence. This value represents the probability of an interest point to be tracked until frame f_j given that it was tracked until frame f_i. This value ranges between 0 and 1. It is 0 when all points tracked until frame f_i are lost in frame f_j, and 1 if both frames f_j and f_i contain the same tracked points. This value is particularly interesting when f_j is the last frame in the sequence. In this case $P(f_j|f_i)$ gives us the probability of a detected point to be in the last frame, given that it was tracked until frame f_i. Expression (4) gives a prediction of the survival of an interest point in future frames if the movement of the robot remains similar. This expression can be used to estimate the number of frames p a landmark needs to be tracked before it is incorporated in the map.

6 Evaluation Methods for Local Descriptors

The evaluation of the different local visual descriptors is done in three steps. First, we choose one of the interest point detectors described in Section 3 and we apply it to every frame in the sequence we want to analyze. Second, we apply the correspondence method of Section 5.1 to match the interest points along the sequence. Third, we extract each of the descriptors presented in Section 4 at the position of each interest point found at least in two images in the sequence.

As a result of the previous steps, an interest point x, which was found in N images in a sequence $\{f_1, \ldots, f_N\}$, will be represented by M different sets $D_1^x, \ldots D_M^x$, where the set $D_m^x = \{d_m^x(1), \ldots, d_m^x(N)\}$ represents the point along the trajectory using the descriptor method m, and each element $d_m^x(i)$ indicates the descriptor representing the point x in frame f_i. In our case $m \in$ {Patch, SIFT, SURF, E-SURF, U-SURF, Zernike, Histogram, GLOH}.

An example of this process is shown in Fig. 5. Here, three interest points $\{x_1, x_2, x_3\}$ are tracked in three frames. At each frame, each interest point is described by two descriptors $d_1, d_2 \in \Re^2$. From this tracking we obtain three vector sets for the descriptor d_1 which represent the points along the frames:

$$D_1^{x_1} = \{d_1^{x_1}(i-1),\ d_1^{x_1}(i),\ d_1^{x_1}(i+1)\},$$
$$D_1^{x_2} = \{d_1^{x_2}(i-1),\ d_1^{x_2}(i),\ d_1^{x_1}(i+1)\},$$
$$D_1^{x_3} = \{d_1^{x_3}(i-1),\ d_1^{x_3}(i),\ d_1^{x_3}(i+1)\},$$

Fig. 4 The top images are examples of a planar object under different changes in viewpoint. The bottom images depict a 3D scene under different scales. These are the first, the central and the last image of a sequence of 21 images (viewpoint transformation) and of 12 (scale transformation). The tracked points are indicated by white marks.

In the same way we can obtain three sets for the second descriptor d_2. In the general case we will have V sets of vectors for each descriptor m, where V is the number of points that where found at least in two images in the sequence. We consider that each tracked point is a visual landmark, and, in consequence, there exists V visual landmarks. Using the description method m the landmark is represented by N descriptors, each one corresponding to a different view.

Let us concentrate on only one descriptor, e.g. d_1. Each of the V sets $D_1^{x_1}, \ldots, D_1^{x_V}$, corresponding to the selected descriptor, forms a cluster in the descriptor subspace. Each cluster represents the point x in the images where it could be found. An example is given in Fig. 6. Here, the three points $\{x_1,\ x_2,\ x_3\}$ of Fig. 5 are found in 10 images, thus $D_1^{x_v} = \{d_1^{x_v}(1), \ldots,\ d_1^{x_v}(10)\}$ for each point x_v. For this example we assume again that $d_1 \in \Re^2$ and has two components $d_1 = \{a, b\}$. Depending on the performance of the description method 1, each set $D_1^{x_v}$ forms a different cluster in $\Re^2$ and represents the same interest when viewed at different distances and angles. Fig. 6 shows three possible clusterings, that depend on the method used to describe the points.

6.1 Feature Matching

We would like to further evaluate the performance of the descriptors in a feature matching context, by means of *recall* and *precision* curves. In this way, not only do we evaluate the quantity of correct correspondences obtained, but also the relative cost of false positives. However, in contrast to the work exposed in [19] the evaluation is made considering that each landmark is represented by a cluster, consisting of descriptors obtained at different views of the same point in space. We consider that this situation is common in visual SLAM, since in most approaches the robots observe the same visual landmarks from different distances and angles [24, 7, 29].

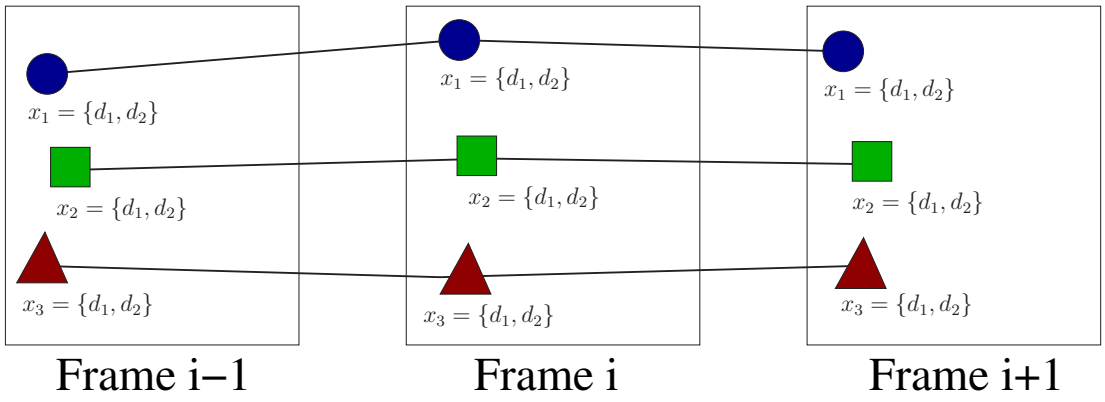


Fig. 5 Three interest points are tracked along three consecutive frames. At each frame, each interest point x_p is represented by two descriptors: d_1 and d_2.

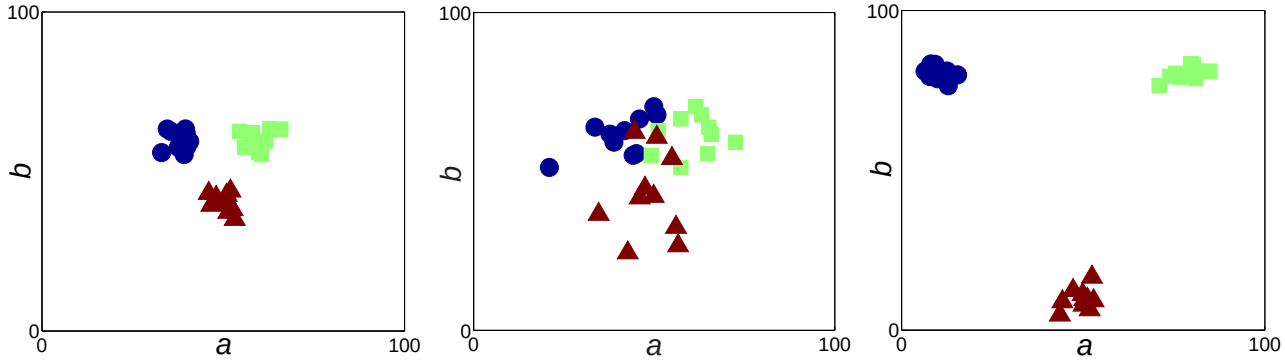


Fig. 6 Three examples of clusterings representing the three tracked interest points of Fig. 5. The clusters are created using one descriptor with two features $d_1 = \{a, b\}$. Fig. 6(a) shows three clusters with small within-class variance and small between-class distances. Fig. 6(b) shows clusters with large within-class variance and small between-class distances. Finally, in Fig. 6(c) clusters with small within-class variance and large between-class distances are shown. This clustering example is based on [27].

By means of the tracking explained in Section 5.1, an interest point x, which was found in, at least, two images in a sequence, will be represented by M different sets $D_1^x, \dots D_M^x$. Thus, the set $D_m^x = \{d_m^x(1), \dots, d_m^x(N)\}$ is considered as a cluster and represents the point along the trajectory using the descriptor method m when viewed from different viewpoints, being N the number of images where the point was found. We consider that there exists a total of V clusters, which correspond to the total number of points that were tracked. Given one descriptor that represents a particular view of a visual landmark, we would like to find its correspondent cluster using a distance measure. To do this we make use of the Euclidean distance, defined as:

$$E = \sqrt{(d_m^{x_i} - d_m^{x_j})^T (d_m^{x_i} - d_m^{x_j})} \quad (5)$$

where $d_m^{x_i}$ is a descriptor belonging to the class ω_i and $d_m^{x_j}$ is a descriptor associated to class ω_j.

For each descriptor $d_m^{x_i}$ that we want to classify, we compute the Euclidean distance to all the clusters in the data set and look for the cluster that minimizes this distance. In addition, we know the correct correspondences for all the descriptors: the correspondence is true when the descriptor $d_m^{x_i}$ is assigned to the cluster ω_i and false when the descriptor is assigned to a different cluster. As a result, we have a list of descriptors, each one with an associated minimum Euclidean distance to a cluster and a $true/false$ correspondence. Next, we sort in ascending order the list of matches according to the minimum Euclidean distance. We use this list in order to compute the *precision* and *recall* parameters, which are defined as:

$$recall = \frac{\#correct\ matches\ retrieved}{\#total\ correct\ matches}$$

$$precision = \frac{\#correct\ matches\ retrieved}{\#matches\ retrieved}$$

In the expressions above, *recall* expresses the ability of finding all the correct matches, whereas *precision* represents the capability to obtain correct matches when the number of matches retrieved varies. Selecting different thresholds for the Euclidean distance in the

ranked list produces different sets of retrieved matches, and therefore different values of *recall* and *precision*. The factor '*#matches retrieved*' represents the number of matches in the list whose distance is below a given threshold value, and varies from 1 to the total number of matches that compose the list (AV). The variable '*#correct matches retrieved*' is the number of correct correspondences obtained for a given threshold in the ranked list. The factor '*#total correct matches*' is a constant value, which expresses the total number of correct correspondences in the list.

In a *precision vs. recall* curve, a high *precision* value with a low *recall* value means that we have obtained correct matches, but many others have been missed. On the other hand, a high *recall* value with a low *precision* value means that we have obtained mostly correct matches but there are also lots of incorrect matches. For this reason, the ideal situation would be to find a descriptor that obtains high values of both parameters simultaneously, thus having values located at the upper-right corner in the *precision vs. recall* curve.

6.2 Cluster Separability

The *recall vs. precision* curves give a result that needs to be carefully interpreted, since it represents the performance of the descriptors in different situations. For this reason we are also interested in obtaining a value that would allow us to rank the descriptors according to its suitability for visual SLAM. In order to do this a separability criterion is introduced in this section.

Ideally, a description method should be invariant to image changes and should also provide a good distinctiveness. In consequence, the cluster representing the same interest point should be compact (small within-class variance) and have a large distance to other clusters representing different interest points (large between-class distance). In this case, we would have a good separability and the landmarks could be distinguished easier. However, if the clusters are very spread and have small separability between them, then the classification of landmarks becomes difficult. Fig. 6 shows three examples of clustering in $\Re^2$. We can see here that the rightmost clusters provide a better separation according to the previous criteria, as they have a small within-class variance and a large between-class distance.

To study the separability of the clusters representing the interest points, i.e. within-class variance and between-class distance, we use the J_3 separability criterion [27]. This measure is based on two scatter matrices: S_w and S_b. S_w is called *within-class scatter matrix*, and measures the within-class variance of a cluster. The *between-class scatter matrix* S_b measures the between-class distance between different clusters. In our case, S_w measures the invariance of the descriptor to viewpoint and scale changes, whereas S_b measures the distinctiveness of the points described. For a given clustering, the within-class scatter matrix S_w is computed as:

$$S_w = \sum_{i=1}^{V} P_i S_i, \tag{6}$$

where S_i is the covariance matrix for the class ω_i:

$$S_i = E[(\mathbf{x} - \mu_\mathbf{i})(\mathbf{x} - \mu_\mathbf{i})^T], \tag{7}$$

P_i the *a priori* probability of class ω_i and μ_i the mean descriptor for the class ω_i. In this case we consider that all classes are equiprobable. Obviously, $trace(S_w)$ is a measurement of the average variance, over all classes, of the samples representing each class. In our case, each class ω_i represents, for a given descriptor m, the set of vectors of a visual landmark along N frames, i.e. $D_m^{x_i}$. The between-class scatter matrix S_b is calculated as:

$$S_b = \sum_{i=1}^{V} P_i(\mu_\mathbf{i} - \mu_\mathbf{0})(\mu_\mathbf{i} - \mu_\mathbf{0})^T, \tag{8}$$

where $\mu_\mathbf{0}$ is the global mean computed as:

$$\mu_\mathbf{0} = \sum_{i=1}^{V} P_i \mu_\mathbf{i}. \tag{9}$$

$trace(S_b)$ is a measurement of the average distance (over all classes) of the mean of each individual class to the global mean value. The J_3 criterion is defined as:

$$J_3 = trace(S_w^{-1} S_m), \tag{10}$$

where S_m is the *mixture scatter matrix* and is computed as $S_m = S_w + S_b$. A good descriptor should have a low value of S_w, since the variability of the vectors describing the same class should be small. Furthermore, it is desirable that vectors describing different points are as distinctive as possible, resulting in a high value of S_b. In consequence, a suitable descriptor would have a high value of J_3. This descriptor would have good results in terms of the data association problem, despite of changes in imaging conditions, such as viewpoint and scale changes.

To compare descriptors with different length we use a normalized version of the criterion: $J_3' = \frac{J_3}{L}$, where L is the descriptor length.

7 Experiments

In order to evaluate the different interest point detectors and local descriptors, we captured 12 sequences of viewpoint changing images, each containing 21 images. Additionally, we captured 14 sequences of images with scale changes, each containing 12 images. Additionally, 2 more sequences present changes in illumination, obtained using different combinations of natural and artificial light. The images were obtained by opening and shutting the window lids at our laboratory. It is worth noting the presence of shadows and non-linear effects in these sequences.

All sequences were captured using a camera (Videre Design MDCS3) mounted on a robotic arm in order to achieve constant variations of viewing angle and distance change. In the case of viewpoint changing sequences, we moved the camera following a semicircular trajectory with center at a point in the scene. We moved the camera 2.5 degrees between consecutive images. Since we captured 21 images, the total amount of angle variation between the first and the last image in the sequence is therefore 50 degrees. In the case of scale changing sequences the camera followed a linear trajectory, moving 0.1 meters between consecutive frames. The total displacement of the camera between the first and the last image in the sequence is 1.1 m. The sequences represent scenes with planar objects (such as posters) and 3D scenes (images at our laboratory). Examples of both types of images are shown in Fig. 4. Finally, the images were captured at different resolutions (320×240, 640×480 and 1280×960), so that the set of images was as much representative as possible. The complete image data set contains approximately 500 images and can be downloaded from `http://www.isa.umh.es/arvc/vision/imgsDataBase/`.

7.1 Evaluation of Interest Point Detectors

In the first experiment we extracted interest points at each image of the sequences using the methods described in Section 3. Next, the points were followed using the techniques explained in Section 5.1. We computed the repeatability rate using Equation (3) for each sequence. The results are shown in Fig. 7 and Fig. 8 respectively. The figures show the mean value and 2σ intervals obtained for several sequences with the same image variations. In most of the cases, the Harris detector shows the best results. For example, in Fig. 7(a) it achieves a repeatability rates above 0.7 in all the images in the sequence. In addition, in the $3D$ case, the Harris corner detector obtains a bigger difference with respect to the other detectors. For example, in Fig. 8(b) the Harris corner detector is able to find approximately a 60% of the points in the last image of the sequence. It is worth noting that the MSER detector obtained results comparable with the Harris corner detector, outperforming Harris in the case of 2D images with changes in scale. In general, we can observe that Harris-Laplace, SIFT and SURF behave in a similar way, with poorer results compared to the Harris corner detector. In the case of 3D scenes the SIFT detector obtained worse results. On the other hand, Kadir's detector obtained good results in some of the experiments. However in the case of 2D viewpoint changing scenes, the results were unsatisfactory. In general the worst results are obtained by the SUSAN detector.

The results obtained with illumination changes are presented in Fig. 9(a) and Fig. 9(b) corresponding to two different scenes the repeatability rate is plotted against the mean brightness in the whole image, however, the illumination changed in an non-linear way, with different variations in the shadows. In this case, the Harris detector obtained the best results with difference. It is worth noting that, in this case, the MSER detector obtained substantially worse results. This degradation can be explained by the fact that MSER tries to extract regions, which have been specially altered by shadows and other artifacts.

Next, using the same sequences and tracked points, we computed $P(f_n|f_i)$ using Equation (4), that represents the probability that a point is found in the last frame n given that it was tracked until the frame f_i. In this case, a tracking of the points was performed from the first to the last image in the sequence. Once a point is lost it is never considered again. The values were computed for sequences with the same transformation, computing at each frame a mean value and 2σ error bounds. Fig. 10 and Fig. 11 show the results. For instance, in Fig. 11(b) it can be observed that a point detected with Harris which is tracked until frame 3 has a probability of 0.6 of being tracked until the last frame. These curves allow us to make a slightly different interpretation of the data. The value of $P(f_n|f_1)$ indicates the fraction of detected points in the first frame that appeared in all the frames of the sequence. For example, in Fig. 10(b) approximately a 55% of the points detected by Harris in the first image could be tracked along the whole sequences. Normally, the probability curves show an increasing tendency, meaning that, if a point has been tracked successfully for a number of frames, it is normally more stable, and consequently the probability that it also appears in the last frame increases. Using this criterion, the Harris detector and the MSER detector obtained comparable results, followed by the remaining methods.

The results showed that the Harris corner detector has demonstrated a high stability under changes in scale and viewpoint in most of the experiments. The good results obtained by the Harris corner detector can be explained by the high amount of corner-like structures that appear in the images that can be robustly extracted. It is worth noting that the results highly depend on the set of images used. For example, the Harris-Laplace detector shows a good behaviour when blob-like structures are present in the environment. However, the main purpose here was to find the best detector that would allow us to perform visual SLAM in this particular environment.

7.2 Evaluation of Local Descriptors

First, we present the evaluation of local descriptors using a feature matching criterion as explained in Section 6.1. Figures 12 and 13 show the results obtained in viewpoint and scale changing images respectively, both for 2D and 3D scenes. The figures represent the *recall* and *precision* curves for each descriptor. The results

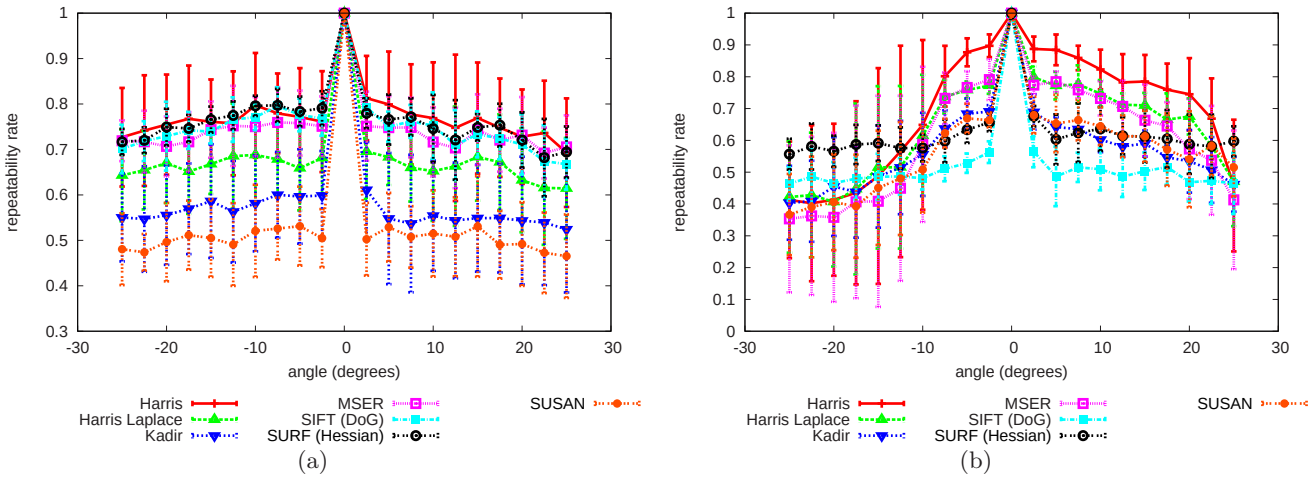


Fig. 7 The left image shows the average repeatability rate of the interest points in all sequences of 2D scenes with changes in viewpoint. The right image depicts the same values but for sequences of 3D scenes. Both figures show 2σ error bounds computed for the sequences with the same changes in the image.

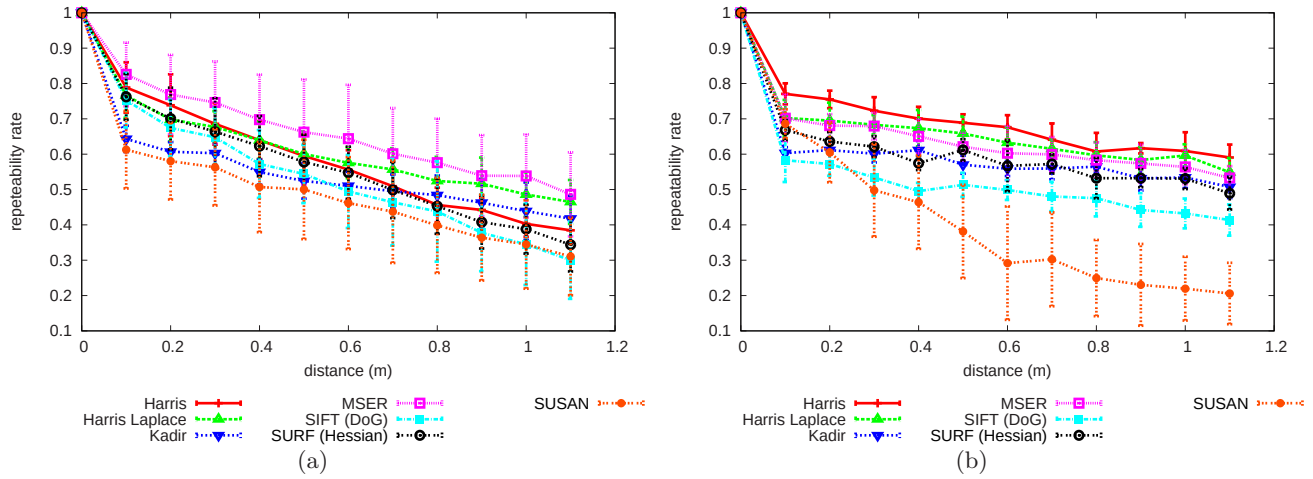


Fig. 8 The left image shows the average repeatability rate of the interest points in all sequences of 2D scenes with changes in scale. The right image depicts the same values but for sequences of 3D scenes. Both figures show 2σ error bounds computed for the sequences with the same changes in the image.

are presented in Fig. 12 (viewpoint changes), Fig. 13 (scale changes) and Fig. 14 (illumination changes). In the case of viewpoint changing images (Fig. 12), the three versions of SURF (SURF, U-SURF and E-SURF) and GLOH obtained similar results, achieving high values of *recall* and *precision* both in 2D and 3D. The U-SURF descriptor shows good results with this criterion, although it is not invariant to rotation. These results, can be explained by the fact that the camera does not rotate around its optical axis in the sequences.

In the case of scale changing images (Fig. 13) the results are similar, in this case the GLOH descriptor obtains the best results in the 2D case, outperforming SURF-based descriptors. In the 3D case GLOH obtains results similar to the SURF versions. Both SURF descriptors and GLOH outperform the SIFT descriptors in all the cases.

In Fig. 14 we show the results obtained with sequences with variations in illumination. In this case, the SURF descriptors and GLOH present similar results. In this case, the SIFT descriptor shows comparable results with the before mentioned.

In the next experiment we computed the descriptors at a local neighborhood of the points detected by the Harris corner detector in each frame of the sequences. The different description methods explained in Section 4 were applied. Tables 1 and 2 show the results of applying the J_3' criterion to different sequences of 2D and 3D scenes. Maximum values are indicated in bold face. The U-SURF descriptor achieves the highest value of separability in 96% of the sequences, outperforming signifi-

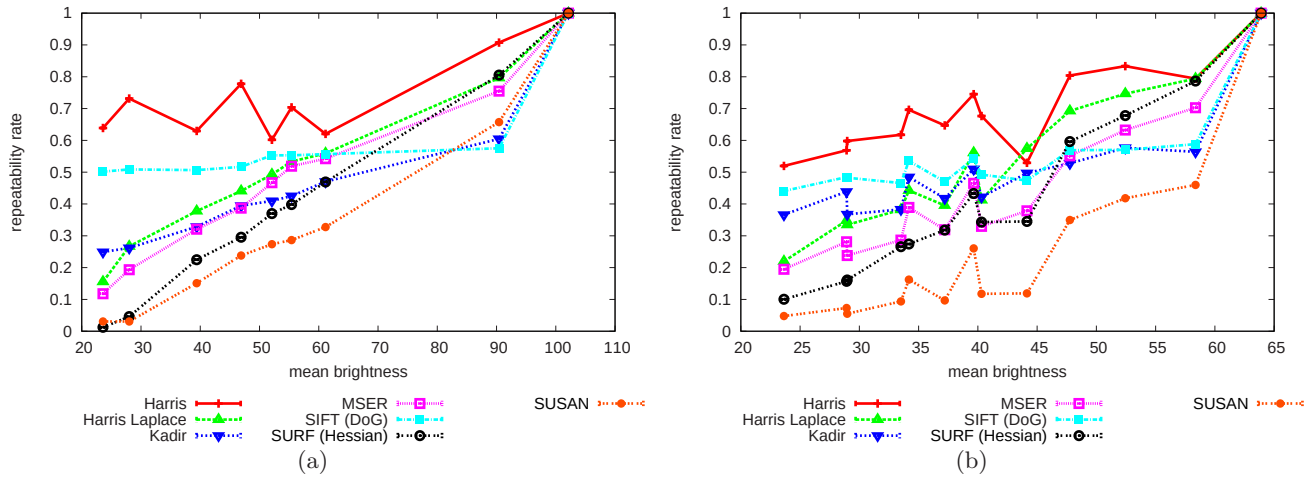


Fig. 9 The figures show the repeatability rate of a sequence with illumination changes.

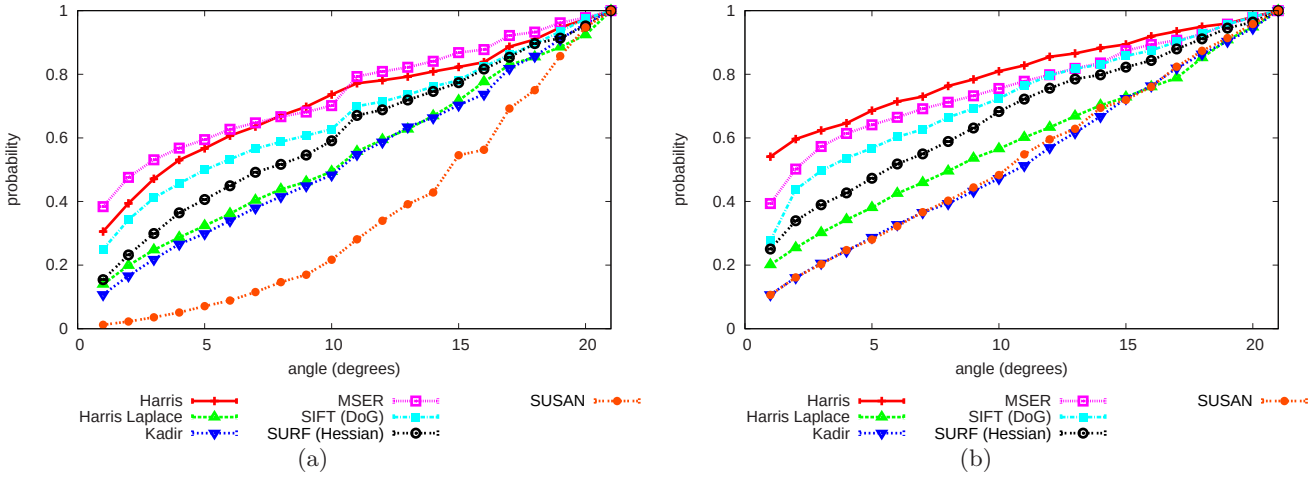


Fig. 10 The left image shows the average value of Equation (4) of the interest points in all sequences of 2D scenes with changes in viewpoint. The right image depicts the same values but for sequences with 3D scenes.Both figures show 2σ error bounds computed for the sequences with the same changes in the image.

cantly the other descriptors. It is worth noting that the sequences that have been used in the experiments do not present changes in rotation. The reason for this restriction is that, in most visual SLAM applications [7,29,11] the camera moves parallel to the ground, and rotates only around its vertical axis. This fact can explain the high performance that the U-SURF has obtained in the experiments.

When comparing only rotationally invariant descriptors (SURF, E-SURF, SIFT and GLOH), it is remarkable that SURF and E-SURF present similar results. In this case, the computational cost of computing the extended version of SURF is not worth the trouble, since the results are not improved substantially. Comparing SURF with GLOH, in the 3D case, SURF always outperforms GLOH in the viewpoint changing images and in the scale changing images, whereas GLOH obtains better results in the 2D sequences.

With reference to the rest of description methods (patch, histogram and zernike moments), is is observable that they do not present remarkable results. It is worth noting that, in the case of the patch description, the results presented may underestimate the maximum capacity of the descriptor to produce good matches. Typically, the matches between different patches are obtained using the normalized cross correlation, which improves the matching capability with illumination changes (e.g. [4]). However, the Euclidean distance to the nearest neighbour is often used to compare descriptors in a matching context (e.g. [19]).

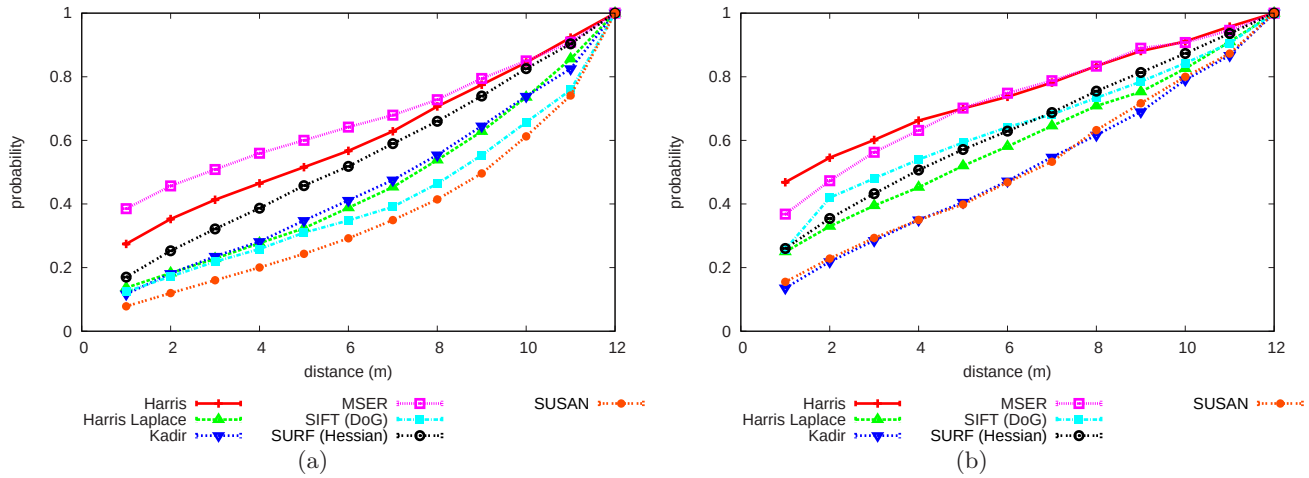


Fig. 11 The left image shows the average value of Equation (4) of the interest points in all sequences of 2D scenes with changes in scale. The right image depicts the same value but for sequences of 3D scenes.Both figures show 2σ error bounds computed for the sequences with the same changes in the image.

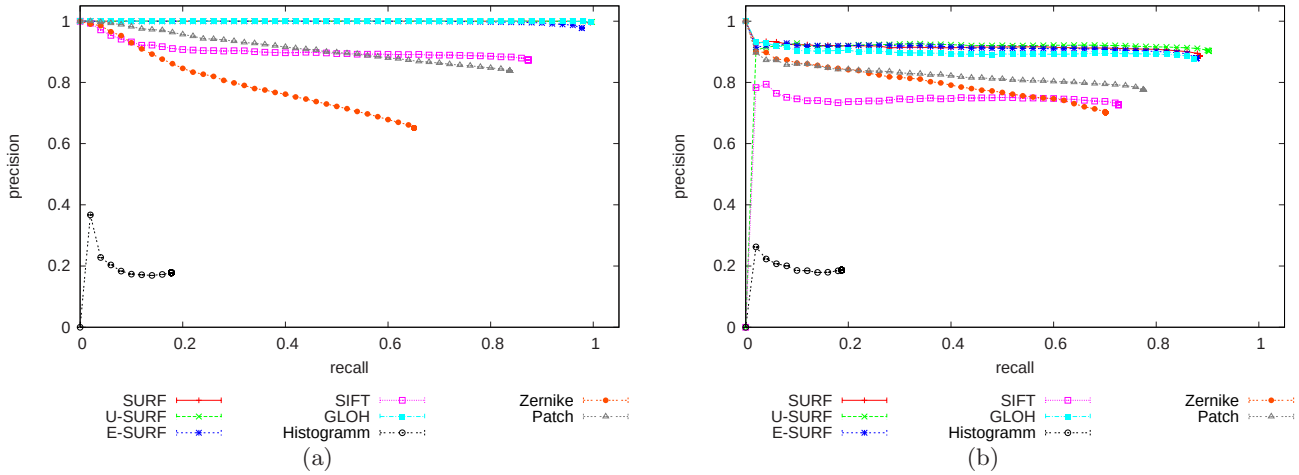


Fig. 12 The left image shows the recall *vs.* precision curve for the 2D sequences with changes in viewpoint. The image on the right shows the results for the 3D sequences.

8 Conclusions

In this paper, we have focused on the evaluation of detection and description methods for visual landmarks under the requirements of vision-based SLAM.

First, we analyzed each detector according to the properties desired for visual landmarks. To do this, we analyzed the stability of the points extracted using different detection methods. The evaluation was performed in image sequences where the total movement of the camera is significant, as usually occurs in visual SLAM applications. On the one hand, we used the repeatability rate in order to analyze the percentage of points found in one image that can be found in the remaining images of the sequence, and therefore are more stable. On the other hand, the evaluation was also performed using the conditional probability. This measure is really profitable for performing SLAM tasks, since estimates for how many frames a landmark should be tracked before being incorporated in the map.

In the case of local descriptors we analyze its clustering properties and matching performance. Two different evaluation methods have been used in order to study the local descriptors under changes in viewpoint and scale. First, the descriptor were evaluated in a matching context. To do this, each descriptor was assigned to the cluster that minimizes the Euclidean distance over all clusters. Next, *recall* and *precision* curves were computed to compare the descriptors. It is noticeable that both evaluation methods agree to conclude that GLOH and the SURF versions are the most suitable descriptors in the experiments that have been performed. The U-SURF

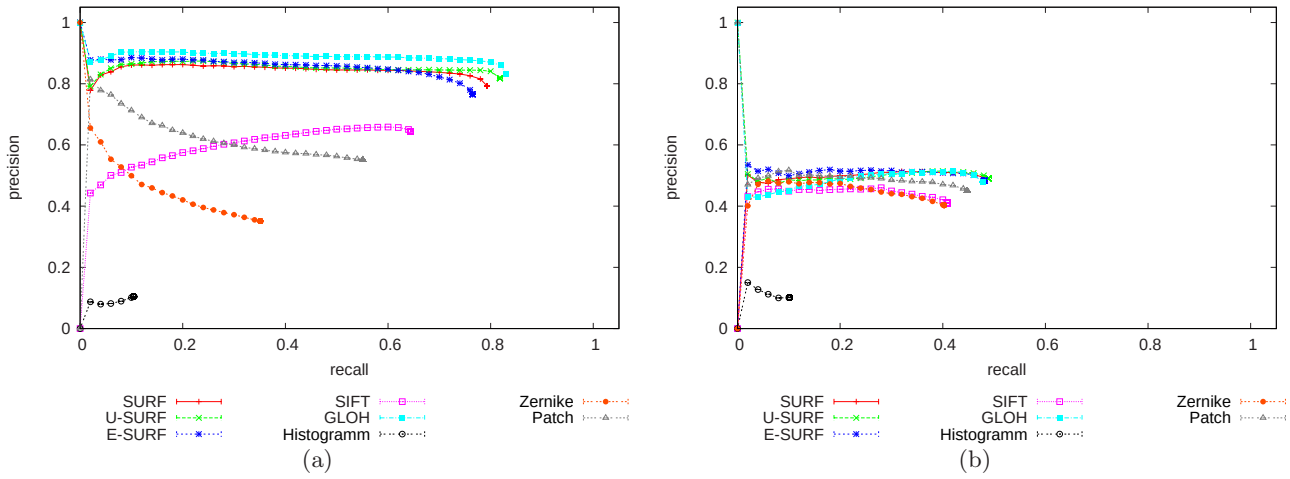


Fig. 13 The left image shows the recall *vs.* precision curve for the 2D sequences with changes in scale. The image on the right shows the results for the 3D sequences.

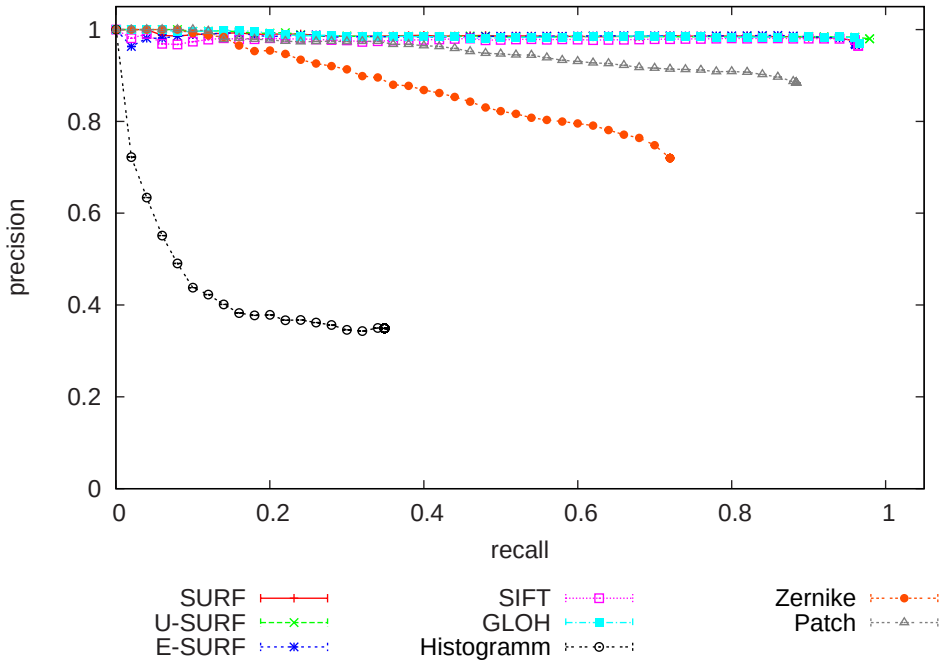


Fig. 14 The left image shows the recall *vs.* precision curve for the images with illumination changes.

descriptor is not invariant to rotation. For this reason, it is limited to applications where the camera only rotates around the vertical axis, which is the case studied in this paper. If we analyze separately the rotationally invariant descriptors (SURF, E-SURF and SIFT), the results obtained with different criteria show that SURF is the best descriptor. In consequence, the SURF descriptor would be the most suitable in situations where the rotation of the camera is not constrained. Next, the evaluation was performed using a clustering measurement, by means of the J'_3 criterion [27], that allows to study the behaviour of the local image descriptors associated to the same point when observed from different viewpoints.

It is also relevant, that the SURF descriptors and GLOH outperformed SIFT in all the situations analyzed in this paper.

Finally, there are other factors that should be considered in the election of a detector and descriptor for visual SLAM. An important parameter in this selection is the computational cost of the detection and description methods. However, this is not the scope of this paper. As an example, and according to its authors [2], the SURF descriptor has a lower computational cost com-

Table 1 J'_3 values computed in the viewpoint changing sequences.

Seq.	SIFT	GLOH	SURF	E-SURF	U-SURF	Patch	Hist.	Zernike
2D								
1	22.90	24.93	36.87	34.18	**126.63**	15.53	2.48	6.39
2	15.89	18.37	39.45	34.00	**119.58**	9.03	1.83	2.93
3	10.18	15.00	30.49	25.81	**118.64**	6.06	1.85	2.90
4	27.24	58.99	68.32	57.81	**184.06**	15.78	2.13	6.54
5	23.75	16.49	27.60	28.32	**55.94**	13.59	2.02	5.87
6	13.38	10.05	29.45	23.47	**67.36**	6.83	1.77	3.68
3D								
7	5.71	4.84	10.70	10.70	**35.93**	2.59	1.46	2.13
8	17.62	13.02	16.45	18.96	**73.23**	5.99	1.51	4.33
9	7.11	5.08	7.83	7.65	**25.17**	3.33	1.72	2.35
10	16.44	9.70	14.47	16.60	**50.58**	7.37	1.54	5.54
11	6.22	3.43	9.60	9.41	**30.33**	2.76	1.78	2.25
12	10.26	8.70	9.63	11.13	**41.09**	4.00	1.43	3.43

Table 2 J'_3 values computed in the scale changing sequences.

Seq.	SIFT	GLOH	SURF	E-SURF	U-SURF	Patch	Hist.	Zernike
2D								
1	7.10	2.43	3.29	2.87	**8.82**	2.32	1.78	2.15
2	7.97	2.94	6.27	5.89	**13.67**	2.59	1.51	2.45
3	9.42	2.85	4.47	4.50	**13.03**	3.45	1.92	2.81
4	14.09	3.07	7.00	9.05	**26.89**	4.22	1.94	2.70
5	103.36	4.98	17.58	38.58	**131.54**	27.73	0.87	14.20
6	4.24	2.69	3.51	3.22	**8.56**	2.81	1.12	2.32
7	7.34	2.46	4.03	4.90	**12.71**	4.87	1.77	2.73
8	**26.49**	2.8311	5.99	10.62	22.65	12.34	2.89	9.05
3D								
9	7.06	1.74	10.12	10.24	**28.01**	4.47	1.70	3.10
10	14.48	2.34	10.39	14.97	**47.48**	5.98	1.67	4.54
11	8.76	1.97	9.18	10.02	**24.72**	3.47	2.48	3.95
12	22.22	2.61	15.53	23.09	**67.38**	8.50	2.15	5.61
13	6.28	1.84	8.84	10.00	**25.56**	3.56	1.94	3.06
14	17.45	2.12	11.10	16.86	**42.37**	7.37	2.10	5.88

pared to SIFT, and this fact would facilitate the online extraction of visual landmarks. In our opinion, the time required detect points and compute descriptors depends highly on implementation details. As a consequence, a general comparison is difficult to obtain.

References

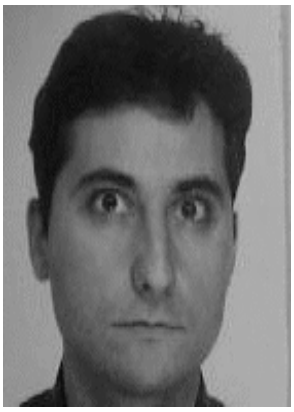
1. Arras, K.O.: Feature-based robot navigation in known and unknown environments. Ph.D. thesis, Swiss Federal Institute of Technology Lausanne (EPFL), Thèse No. 2765 (2003)
2. Bay, H., Tuytelaars, T., Gool, L.V.: SURF: Speeded up robust features. In: European Conference on Computer Vision (2006)
3. Biber, P., Andreasson, H., Duckett, T., Schilling, A.: 3D modelling of indoor environments by a mobile robot with a laser scanner and panoramic camera. In: IEEE/RSJ Int. Conf. on Intelligent Robots & Systems (2004)
4. Davison, A.J., Murray, D.W.: Simultaneous localisation and map-building using active vision. IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)
5. Eustice, R., Singh, H., Leonard, J.: Exactly sparse delayed-state filters. In: IEEE Int. Conf. on Robotics & Automation (2005)
6. Fraundorfer, F., Bischof, H.: A novel performance evaluation method of local detectors on non-planar scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (2005)
7. Gil, A., Reinoso, O., Burgard, W., Stachniss, C., Martínez Mozos, O.: Improving data association in rao-blackwellized visual SLAM. In: IEEE/RSJ Int. Conf. on Intelligent Robots & Systems (2006)
8. Grisetti, G., Stachniss, C., Burgard, W.: Improved techniques for grid mapping with rao-blackwellized particle filters. IEEE Transactions on Robotics **23**(1) (2007)
9. Hähnel, D., Burgard, W., Fox, D., Thrun, S.: An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements. In: IEEE/RSJ Int. Conf. on Intelligent Robots & Systems. Las Vegas, NV, USA (2003)
10. Harris, C.G., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference (1998)
11. Jensfelt, P., Kragic, D., Folkesson, J., Björkman, M.: A framework for vision based bearing only 3D SLAM. In: IEEE Int. Conf. on Robotics & Automation (2006)
12. Kadir, T., Brady, M., Zisserman, A.: An affine invariant method for selecting salient regions in images. In: Proc. of the 8th European Conf. on Computer Vision, pp. 345–457 (2004)
13. Kosecka, J., Zhou, L., Barber, P., Duric, Z.: Qualitative image based localization in indoor environments. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (2003)
14. Little, J., Se, S., Lowe, D.: Global localization using distinctive visual features. In: IEEE/RSJ Int. Conf. on Intelligent Robots & Systems (2002)
15. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision (1999)
16. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **2**(60), 91–110 (2004)
17. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. of the 13th British Machine Vision Conf., pp. 384–393 (2002)
18. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Int. Conf. on Computer Vision (2001)
19. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(10) (2005)
20. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of computer Vision **65(1/2)**, 43–72 (2005)
21. Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B.: Fastslam: a factored solution to the simultaneous localization and mapping problem. In: Eighteenth national conference on Artificial Intelligence, pp. 593–598. American Association for Artificial Intelligence, Menlo Park, CA, USA (2002)
22. Murillo, A.C., Guerrero, J.J., Sagüés, C.: Surf features for efficient robot localization with omnidirectional images. In: IEEE Int. Conf. on Robotics & Automation (2007)
23. Schmid, C., Mohr, R., Bauckhage, C.: Evaluaton of interest point detectors. International Journal of computer Vision **37**(2) (2000)
24. Se, S., Lowe, D.G., Little, J.: Vision-based mobile robot localization and mapping using scale-invariant features. In: IEEE Int. Conf. on Robotics & Automation (2001)

25. Sim, R., Elinas, P., Griffin, M., Little, J.: Vision-based slam using the rao-blackwellised particle filter. In: IJCAI Workshop on Reasoning with Uncertainty in Robotics (RUR) (2005)
26. Smith, S.: A new class of corner finder. In: British Machine Vision Conference (1992)
27. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, third edn. Academic Press (2006)
28. Triebel, R., Burgard, W.: Improving simultaneous mapping and localization in 3D using global constraints. In: National Conference on Artificial Intelligence (AAAI) (2005)
29. Valls Miro, J., Zhou, W., Dissanayake, G.: Towards vision based navigation in large indoor environments. In: IEEE/RSJ Int. Conf. on Intelligent Robots & Systems (2006)
30. Zernike, F.: Diffraction theory of the cut procedure and its improved form, the phase contrast method. Physica **1**, 689–704 (1934)
31. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. In: Artificial Intelligence, vol. 78 pp. 87-119 (1995)

Arturo Gil Aparicio received the M. Eng. degree in Industrial Engineering from the Miguel Hernández University (UMH), Elche, Spain, in 2002, receiving also the best student academic award in Industrial Engineering by the UMH. He has actively participated in several projects in the Systems Engineering and Automation area of the UMH since 2000. Since 2003, he works as a lecturer and researcher at the UMH, teaching subjects related to Control, Computer Vision and Informatics. His research interests are focussed on mobile robotics, visual SLAM and cooperative robotics. He is currently working on techniques for building visual maps by means of mobile robots. He is member of CEA-IFAC.

Óscar Martínez Mozos is a PhD student at the Department of Computer Science at the University of Freiburg, working in the group for Autonomous Intelligent Systems headed by Prof. Dr. Wolfram Burgard. His research focuses on mobile robotics and artificial intelligence. Previously, he was doing research on Visual Neuroscience at the Bioengineering Institute at the University Miguel Hernández in Spain. In 2004 he received a M.Sc. degree in Applied Computer Science at the University of Freiburg. In 2001 He received a M.Sc. degree (Spanish Advance Studies Degree) in Bioengineering at University Miguel Hernández (Spain). In 1997 he completed a M.Eng. in Computer Science at the University of Alicante (Spain). He has also worked as software developer in several companies.

Mónica Ballesta Galdeano received the M. Eng. degree in Industrial Engineering from the Miguel Hernández University (UMH), Elche, Spain in 2005, receiving also the academic award in Industrial Engineering by the UMH. In 2006, she worked as a researcher at the Systems Engineering and Automation area of the UMH. Since 2007, she has a research position as scholarship holder in the area of Systems Engineering and Automation of the UMH, receiving a pre-doctoral grant (FPI) by the Valencian Government (Generalitat Valenciana). Her research interests are focussed on mobile robots, visual feature extraction and visual SLAM.



Óscar Reinoso García received the M. Eng. degree from the Polytechnical University of Madrid (UPM), Madrid, Spain in 1991. Later, he obtained de Ph.D. degree in 1996. He worked at Protos Desarrollo S.A. company in the development and research of artificial vision systems from 1994 to 1997. Since 1997, he works as professor at the Miguel Hernández University (UMH), teaching subjects related to Control, Robotics and Computer Vision. His main research interests are mobile robotics, teleoperated robots, climbing robots, visual control and visual inspection systems. He is member of CEA-IFAC and IEEE.