Unsupervised Identification of Useful Visual Landmarks Using Multiple Segmentations and Top-Down Feedback

Pablo Espinace, Daniel Langdon, and Alvaro Soto
Department of Computer Science, Pontificia Universidad Catolica de Chile
Casilla 306 - Santiago 22 - CHILE

Abstract

In this paper, we tackle the problem of unsupervised selection and posterior recognition of visual landmarks in images sequences acquired by an indoor mobile robot. This is a highly valuable perceptual capability for a wide variety of robotic applications, in particular autonomous navigation. Our method combines a bottom-up data driven approach with top-down feedback provided by high level semantic representations. The bottom-up approach is based on three main mechanisms: visual attention, area segmentation, and landmark characterization. As there is no segmentation method that works properly in every situation, we integrate multiple segmentation algorithms in order to increase the robustness of the approach. In terms of the top-down feedback, this is provided by two information sources: i) An estimation of the robot position that reduces the searching scope for potential matches with previously selected landmarks, ii) A set of weights that, according to the results of previous recognitions, controls the influence of each segmentation algorithm in the recognition of each landmark. We test our approach with encouraging results in three datasets corresponding to real-world scenarios.

Key words: Visual landmarks, robot localization, autonomous navigation, top-down visual feedback, topological maps.

1. Introduction

Vision is an attractive option to provide an intelligent agent, such as a mobile robot, with suitable perceptual capabilities to successfully deal with the complexity of an unstructured natural environment. In particular, the unsupervised selection and posterior recognition of relevant visual landmarks is a highly valuable perceptual capability for a wide variety of robotic applications, such as object recognition and autonomous navigation.

In the context of mobile robot navigation, the problems of automatic construction of maps of the environment and accurate estimation of the position of the robot within a map, tasks known as mapping and localization, have been a long time aspiration for the Robotics community. In fact, most practical applications of mobile robots in natural environments require some sort of solution to these problems. Particularly, mapping and localization are highly relevant issues for the case of indoor environments, where globally accurate positioning systems, such as GPS, are not available.

At present, the state of the art solutions to indoor mapping and localization problems are mainly based on using 2D laser range finders and metric map representations, such as evidence-grids [4]. Although, this type of approaches has shown a high degree of success when operating in real time in natural environments [19], they still suffer from some limitations. For example, the usual structural symmetries of indoor building produce data association problems that are hard to solve with the 2D view of a laser range finder. Furthermore, problems such as modifications of the environment due to changes in the position of furniture, uncertainties due to the state of doors, or partial occlusions due to people walking around, also diminish the robustness of solutions based on 2D laser range finders.

Recently, advances in the area of computer vision [22] [9] have increased the interest in including vision as one of the main sensor modalities to support the perceptual needs of autonomous navigation. In this respect, the robustness and flexibility exhibited by the navigation systems of most seeing beings is a clear proof of the advantages of counting with a suitable visual perception system.

At present, several studies show that biological beings, such as bees and humans, successfully navigate using representations of the world based on visual landmarks [2] [13]. These landmarks are characterized not only by bottom-up

visually perceivable characteristics, such as shape, physical continuity, or color, but also by top-down information like context, purpose, and memories of learned landmarks and their spatial relations.

After the work of Marr [10] the usual approach in machine vision has been the bottom-up model. A data driven bottom-up model, however, is not able to learn and use the acquired knowledge in order to reduce the heavy load of a powerful vision system. A natural extension of the bottom-up model is the use of top-down feedback from high-level semantic representations. This feedback can guide a vision system to allocate its resources in the most prominent places of the image space, improving the robustness and efficiency of the resulting system.

In this paper we present an unsupervised method for the automatic selection and subsequent recognition of suitable visual landmarks using images acquired by a mobile robot. To achieve this goal, we combine bottom-up visual features based on color, intensity, and depth cues, with top-down feedback given by spatial relations and memories of the most successful predicting features of previously recognized landmarks. Our overall goal is to select interesting, meaningful, and useful landmarks that can be used by a mobile robot to achieve indoor autonomous navigation.

We base our bottom-up approach for the selection of candidate landmarks on the integration of three main mechanisms: visual attention, area segmentation, and landmark characterization. Visual attention provides the ability to focus the processing resources on the most salient parts of the input image. This eliminates the detection of irrelevant landmarks and significantly reduces the computational cost. Area segmentation provides the ability to delimit the spanning area of each salient region. This spanning area defines the scope of each relevant landmark. Finally, landmark characterization provides a fingerprint for each landmark given by a set of specific features. These features allow the system to recognize and distinguish each landmark from others in subsequent images.

Although, considerable effort has been put on finding solutions to each of the three individual visual mechanisms mentioned above, the integration of these steps has not been deeply pursued. In general, a poor integration of attention, segmentation, and characterization reduces the efficiency and robustness of the resulting approach. In particular, we believe that an incorrect segmentation of a landmark produces severe problems in the next steps of the image analysis. An inaccurate segmentation might cause that the characterization of the landmark is imprecise, more complex, or even erroneous. As an example, an incorrect segmentation of a landmark that includes parts of the background in the segmented area, might lead to situations where the system only recognizes the landmark when the specific background is present, or even worse, to situations where the system confuses the landmark with parts of the background image.

This paper addresses the integration problem mentioned above, focusing on the segmentation step. Our hypothesis is that no segmentation algorithm can work correctly in every situation, so multiple segmentation algorithms need to be used to obtain a correct landmark from the attended spot. The best segmentation or a combination of them can be then used to achieve a more robust recognition.

The direct application of the three-steps approach mentioned above poses two problems. On one hand, as the size of the working environment increases, the number of salient landmarks selected by the pure bottom-up approach reaches unmanageable levels. Particularly, the search for a match in a database with previously selected landmarks starts to collapse. On other hand, for some landmarks only some of the available segmentation algorithms provide useful information, while other segmentations just add noise to the detection. A naive use of these noisy segmentations can confuse the recognition.

To alleviate the previous problems, we complement our bottom-up landmark selection approach introducing two modalities of top-down feedback that increase the efficiency and robustness of our approach. First, to increase efficiency, we use an estimation of the robot position to reduce the searching scope for a potential match with a previously selected landmark. The estimation of the robot position is based on traditional Sequential Monte-Carlo Localization methods [19] using a metric map representation augmented with topological information. Second, to increase robustness, we keep a record of the previous successful recognitions of each landmark. We use this information to bias the influence that each bottom-up segmentation plays in the recognition. We achieve this goal by adaptively updating a set of weights that control the relevance of each segmentation in the recognition of each landmark.

Accordingly, the main contributions of this work are: i) To develop a bottom-up approach to select and recognize relevant visual landmarks integrating three main mechanisms: visual attention, adaptive fusion of multiple segmentations, and landmark characterization. ii) To incorporate top-down feedback mechanisms that increase the efficiency and robustness of the pure bottom-up approach, iii) To implement and test the resulting approach using images acquired by a mobile robot navigating in a natural unconstrained indoor environment.

This document is organized as follows. Section 2 provides background information and describes related previous work. Section 3 presents the proposed method and relevant implementation details. Section 4 presents our empirical results. Finally, Section 5 presents the main conclusions of this work and future avenues of research.

2. Previous Work

In the computer vision literature, there is an extensive list of works that, individually, target the problems of attention, segmentation, and characterization. We briefly review some relevant works in each area. We also review some relevant works concerning selection and recognition of visual

landmarks in the context of robot navigation, particularly the mapping and localization problems.

Attention is the process of selecting visual information from an image based on a measure of saliency. The saliency is influenced by the concrete relevance of a part of the image, such as strong color contrast. In computational terms, saliency is usually employed to focus the processing resources in key parts of an image, in order to improve efficiency, performance, or both. Previous work in the area includes several models of visual attention. Tsotsos et al. [21] selectively tuned neuron models at the salient location with top-down mechanisms and winner-take-all networks. Itti et al. [7] introduced a model for selecting locations from a saliency map according to decreasing saliency. Sun and Fisher [17] proposed a hierarchical object-based attention framework that integrates visual saliency from bottom-up groupings with top-down object-based selectivity mechanisms. The same authors [18] also extended Duncan's Integrated Competition Hypothesis [3] with a framework for location-based and object-based attention using grouping.

Segmentation is the process of partitioning an image in non-overlapping regions according to relevant visual properties and metric distances. As such, there are many ways in which an image can be segmented. In the computer vision literature, a great number of segmentation techniques have been proposed [5], however, it is not possible to find a single general purpose segmentation algorithm that is effective in all situations. As a consequence, there has been an increasing interest in combining the results of multiple segmentations algorithms. As an example, in the context of object recognition, independent works by Roth and Ommer [14], Rabinovich et al. [12], and, recently, Malisiewicz and Efros [20], have empirically shown the advantages of combining multiple segmentations algorithms over using a single one. This is similar to one of our goals, but we adaptively combine the segmentations using top-down feedback.

Characterization is the process of finding a set of descriptors, or fingerprint, that provides a simple and ideally unambiguous way to identify each particular landmark. One of the most relevant trends in this area looks for the presence of stable features in the input image. These features must be stable, even under slight variations on the input image, such as changes in lighting conditions, field of view, or partial occlusions. This approach became very popular after Harris and Stephens [6] presented their corner detector. Recently, Lowe [9] presented a refinement of this idea, called the Scale Invariant Feature Transform (SIFT), which gained great popularity due to its success in several applications.

There is also related work in the area of simultaneous localization and mapping (SLAM) using visual sensors. Here, most current works [8], [25] consider the complete input image as the relevant area where a feature detector is applied. As an example, Karlsson et al. [8] presented a solution for the robot localization problem based on SIFT features extracted from the complete input images. Although the high redundancy in the feature vector extracted from the complete image is a good fingerprint for recognition, a problem arises with the scaling properties of the approach, as the number of relevant views of the environment increases. Furthermore, the performance of these approaches presents a high degradation with changes in the lighting conditions. In our case, the use of an attention mechanism and top down feedback helps to alleviate these problems.

Among approaches related to robot navigation that use visual attention mechanisms, Walther et al. [23] used a bottom-up approach similar to the one proposed in this paper, but for the problem of detecting moving objects using a remotely operated underwater vehicle. In their work, Walther et al. also demonstrate the influence of visual saliency in recognition. In our case, we focus on demonstrating the influence of segmentation in recognition, besides that we include top-down feedback mechanisms to improve the performance of our system. Recently, Siagian and Itti [15] present preliminary results about a system that combines gist and attention mechanisms to achieve outdoor robot localization. In our case, instead of gist, we use a traditional Sequential Monte-Carlo localization method, as a top-down mechanism to achieve a probabilistic estimation of the current robot position. Furthermore, we also include a second top-down mechanism that adaptively biases the influence of multiple segmentations in the detection of each landmark.

3. Our Approach

In this section, we first present our bottom-up approach to select and recognize landmarks. This approach integrates visual attention, area segmentation, and landmark characterization mechanisms. Afterward, we explain in detail the top-down mechanisms added to improve our final system. Figure 1 shows an overall view of the complete approach. We explain the details next.

3.1. *Bottom-Up Approach*

3.1.1. *Attention*

We use the bottom-up saliency map of Itti et al. [7] to extract salient locations from an input image. In our implementation, we slightly modify the original algorithm by introducing an adaptive scheme that dynamically selects an appropriate number of relevant salient locations. A brief explanation of the algorithm follows.

Each image is processed to extract multi-scale maps of orientation, intensity, and color. After computation of center-surround differences, the algorithm creates feature maps for each scale. These maps are then combined to form a color map, an intensity map, and an orientation map. These maps are grayscale representations, where bright areas represent highly salient locations. A single saliency map is obtained by merging these individual maps. Once the saliency map has been extracted, a winner-take-all (WTA) neural network is used to extract the image coor-

dinates of the center of the most salient location. Then, an inhibition-of-return method is applied to prevent the selection of the same location multiple times. The WTA neural network iterates several times over the saliency map, obtaining the potential location of the relevant landmarks.

In the original implementation proposed by Itti et al., the algorithm has no limitations on the number of locations to be obtained. The saliency map is normalized after each iteration, therefore, a new salient location can be found every time. Previous approaches find a fixed number of locations in each image [24]. Images from different scenarios, however, produce a highly variable number of relevant locations, so an adaptive scheme is needed.

In our implementation, we adapt the number of locations selected, by limiting the evolution of the WTA neural network that finds each salient location. The key observation is that the evolution time required by the network to find a salient location is proportional to the intensity and distribution of saliency in the saliency map. In other words, the network needs less time to evolve with an image with distinctive and interesting objects. In this way, we calibrate the evolution time on the WTA network, according to the distribution of bright pixels in the saliency map, using training images. As a training criteria we use the average evolution time to match the number of relevant landmarks detected by a human operator in the training images.

3.1.2. *Segmentation*

Although a large amount of research has been made on segmentation, there is not yet a complete solution to this problem. Every segmentation algorithm copes with certain situations but fail to produce adequate results in others. Since, in general, we can not predict a priori the conditions of each input image, we can use multiple segmentation algorithms to increase the adaptability and robustness of our landmark detection algorithm.

As a testbed, we select three existing segmentation algorithms based on color, saliency maps, and depth information, to find the area defined by the underlying landmarks. Each of the algorithms relies on highly independent visual information, therefore, we expect that their behaviors differ depending on the input conditions. We refer to these three algorithms as the color-based, saliency-based, and stereo-based segmentation algorithms, respectively.

The color-based segmentation algorithm is based on a technique proposed in [11] to segment color food images. The original algorithm includes three main steps. First, a grayscale image is obtained from the input image using an optimal linear combination of the RGB components found in the pixels around a specific image location. Then, a statistical approach is used to estimate a global threshold used by a region growing segmentation step. Finally, a morphological operation is used to fill possible holes that might appear in the final segmented area.

In terms of our application, one of the advantages of the previous algorithm is that it does not segment the full input image but it only extracts a single object from it. The challenge is then finding a suitable image region to calibrate the parameters of the color model. This is easily solved in our implementation by automatically calibrating the color model, using as foreground the area around a salient location.

The saliency-based segmentation algorithm is based on a technique proposed in [24]. This algorithm was also designed to work with the attention model proposed by Itti, so its application to our case is straightforward. The basic idea is that while Itti's model identifies the most salient location in the image, it does not provide the extent of the image region that is salient around this location. To find this region, the authors in [24] proposed to apply a region growing segmentation algorithm that uses as the starting seed the center of the salient area. This segmentation algorithm is applied over the feature map with the greatest influence over the selected section of the saliency map, see [24] for details. Although this algorithm includes a color saliency map in its process, its color model has important differences from the model used by the color-based segmentation algorithm described above, and thus, the algorithms are highly independent.

The stereo-based segmentation algorithm is based on the disparity image obtained from a pair of images provided by a calibrated pair of video cameras. The segmentation algorithm assumes that for each particular landmark, two adjacent pixels present depth continuity. Under this assumption, the landmark pixels form a connected area in the disparity image, that can be computed through a region-growing procedure. For more details about the stereo-based segmentation algorithm, see [16].

3.1.3. *Characterization*

For the characterization of the segmented patches, we use the SIFT feature extraction algorithm [9]. This algorithm provides highly discriminative features that, to some extent, are robust to the presence of affine distortion, noise, changes of viewpoint, and changes in illumination.

Using SIFT, each landmark is characterized by a group of redundant individual SIFT descriptors. Given this redundancy, a landmark can be recognized even when only a subset of the original features presents a correct match. This produces a certain degree of robustness under occlusion problems.

3.1.4. *Integration*

To accomplish our goal of unsupervised selection and subsequent recognition of landmarks, we need to integrate the three steps described above. Figure 1 shows the different steps of our integration scheme. The procedure is as follows.

An input image is received and then its saliency map is computed by the attention algorithm. The first salient location is extracted, and the three segmentation algorithms are used to extract landmark candidates. Inhibition of return is calculated from the shape of the segmented land-

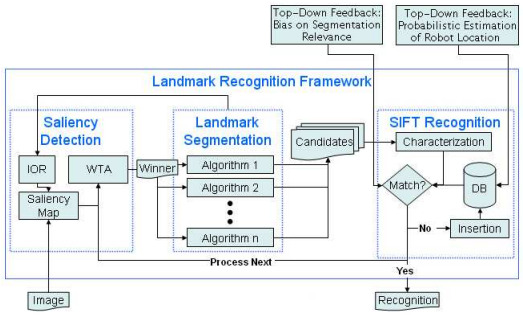

Fig. 1. Schematic operation of our approach. This integrates attention, segmentation, and characterization mechanisms in addition to top-down feedback for unsupervised recognition of relevant visual landmarks.

marks and applied to the saliency map. This avoids selecting the same location in posterior iterations. The previous steps are repeated until the time to evolve the WTA network indicates that there are no more relevant salient regions to consider. At the end of this process, we obtain a list of candidate landmarks corresponding to the resulting segmented areas around the selected saliency regions.

Once we have obtained the candidate landmarks, we extract SIFT features for each available segmentation of each landmark in the list. These features are then compared to the features of the landmarks in our database, that, at the beginning of the process is empty. To estimate if a candidate landmark matches the SIFT description of any of the landmarks included in the database, we use a similarity score based on a nearest-neighbors technique, as described in [9]. Given that, in our case, for each landmark we keep 3 possible descriptions corresponding to each of the available segmentations, we combine the respective similarity scores using a set of importance weights. Section 3.2.2 provides the details of the method used to set the values of these importance weights, that define the final score P_{match}, in eq. 2, used to verify the recognition of a candidate landmark.

If a candidate landmark matches one of the landmarks in the database, we report the match and modify our record of the number of times the landmark has been successfully recognized by the different segmentations. As we describe in section 3.2.2, we use later this information to update the importance weights associated to the influence that plays each segmentation in the recognition of each landmark.

If a candidate landmark does not match any of the landmarks in the database, we add this landmark to the database provided that certain constraints are satisfied. Our premise is to keep in the database only landmarks that are highly distinctive and easy to detect. According to this, we only include in the database landmarks that present SIFT features with an strength about a fixed threshold. Furthermore, we require that the number of relevant SIFT features for each landmark in the database is greater than 12 SIFT features. Our experiments indicate that landmarks with fewer features usually do not produce enough matches to trigger a robust detection.

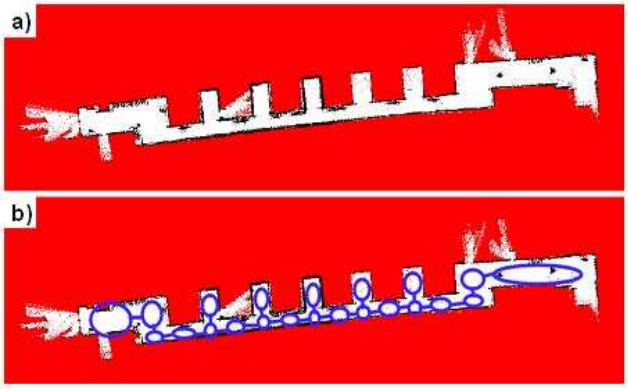

Fig. 2. Map representations of our Computer Science Department. a) Metric map. b) Topological map superimposed over the metric map.

3.2. *Top-Down Feedback*

As we pointed out before, a pure bottom-up approach does not scale properly with the size of the environment and does not learn which segmentations are most useful to detect each landmark. Next, we explain the top-down mechanisms that we use to alleviate these problems. First, we explain our method to restrict the searching scope for each landmark match, based on an estimation of the robot position. Then, we explain our method to integrate the information of the segmentation algorithms based on a set of weights that bias the integration according to the results of previous detections.

3.2.1. *Use of an estimate of robot position*

One of the more computationally expensive parts of our approach is database operation, particularly, insertion and searching for a landmark match. This is especially critical when the database contains a great number of stored landmarks, that must be compared to each new candidate.

To improve efficiency, we divide the database into several smaller databases. Each of these databases stores a group of landmarks that belongs to a significant part of the environment, such as a corridor or a room. Using an estimation of the position of the robot at the time each image is taken, we estimate the location of the potential visible landmarks by using the distance information available from the stereo-based segmentation. This procedure results in a much faster execution, as new landmark candidates are only compared to a reduced set of local landmarks.

To divide the environment in a set of local relevant places, we augment our metric map representation with topological information. Each node of the topological map covers one part of the metric map that might correspond to a hall, a room, a corridor, or an intersection in the environment. At this point, we manually label the topological map. Figure 2-a shows the metric map representation of our Computer Science Department (DCC). This map was automatically built by our robot using the SLAM algorithm presented in [1]. Figure 2-b shows the corresponding topological map superimposed over the metric map.

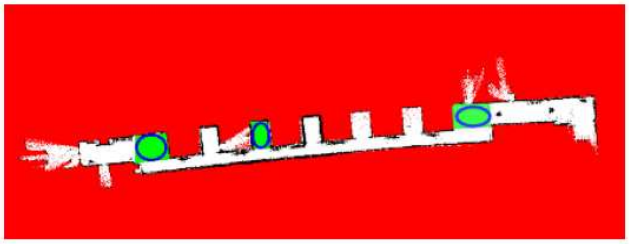

Fig. 3. Map representation displaying a case where the Sequential Monte-Carlo estimation of the position of the robot indicates that the robot might be located in any of 3 possible topological nodes.

In our current implementation, we associate one landmark database to each node in the topological map. Furthermore, each node in the topological map stores information about the cells of the grid map that are contained in it. In this way, when searching for a landmark match, we restrict the search to the databases corresponding to topological nodes associated to grid cells with high likelihood under the current estimation of the robot position. Figure 3 shows a situation with 3 topological nodes associated to cells with high probability.

3.2.2. *Biased integration of segmentations*

In order to integrate the three available segmentations, an importance weight is associated to each segmentation for each landmark. This importance weight controls the influence of each segmentation algorithm in the recognition of each landmark. The idea is to adaptively assign the importance weights according to the historic performance of the respective segmentation in the recognition of each landmark. We explain next the procedure to estimate and update the importance weights.

Initially, when a new landmark is accepted in a database, the corresponding importance weights are estimated based on the number of SIFT points detected in each segmentation of the landmark. As an example, let w_c^i be the importance weight for the color-based segmentation for landmark i, and let S_c^i, S_s^i, S_t^i be the corresponding number of SIFT points calculated over the regions obtained by the color, saliency, and stereo-based segmentations, respectively. The initial value for w_c^i is calculated by:

$$w_c^i = S_c^i/(S_c^i + S_s^i + S_t^i). \quad (1)$$

The initial values for the importance weights w_s^i and w_t^i of the saliency and the stereo-based segmentations are initialized in a similar way. This initialization scheme assigns a greater weight to segmentations that have more SIFT points. Our rationale is that a segmentation with more SIFT descriptors should provide a better characterization of the landmark. To simplify the notation from now on, we drop the superindex i.

As we mentioned before, to achieve the recognition of a candidate landmark, we use the similarity score between SIFT descriptions of objects proposed in [9]. Using this score and the importance weights for each segmentation, a recognition probability (P_{match}) is calculated. Let M_c, M_s and M_t be the similarity scores for the SIFT descriptions of the color, saliency, and stereo-based segmentations, respectively. The recognition probability for a candidate landmark is calculated as:

$$P_{match} = w_c \times M_c + w_s \times M_s + w_t \times M_t. \quad (2)$$

Finally, every time a landmark is recognized, we use the results of the recognition to update its importance weights. We perform the update by considering the new matching scores associated to each of the segmentations. As an example, for the color-based segmentation, the importance weight w_c is updated as follows:

$$w_c = \alpha \times w_c + (1 - \alpha) \times \hat{M}_c, \quad (3)$$

where $\hat{M}_c$ is a normalized version of the similarity score M_c calculated with respect to the similarity scores of the 3 segmentations, and $\alpha \in [0, 1]$ is a constant that controls the influence of the result of the last recognition in the updating of w_c. In this work we set the value of α to 0.2. The weights w_s and w_t are updated in a similar way, using the same constant value α, but considering the respective normalized scores $\hat{M}_s$ and $\hat{M}_t$.

4. Results

In this section we describe the results of testing our approach using several datasets corresponding to real-world scenarios. We start the section providing details of these datasets. Next, we present the overall performance of the approach, displaying representative landmarks selected by the system and the recall capabilities to recognize landmarks previously inserted in the databases. Afterwards, we describe a set of specific experiments designed to highlight some particular features of our approach. We include: i) An experiment designed to show the impact of combining several segmentations in the performance of the system, ii) Two experiments designed to highlight the advantages of each of the top-down feedback mechanisms used in this work. Finally, we end the section with a discussion of our main results.

4.1. *Datasets*

We obtained two different datasets to test our algorithm. The first dataset corresponds to 185 images captured by a digital camera inside three different rooms of a house environment using a resolution of 640x480 pixels (see Figure 4b). To test the robustness of the algorithm, we consider images with different viewpoints, rotations, illuminations, and even blur produced by the motion of the camera.

The second dataset consists of 10 video sequences, with an average number of 1605 image-pairs per sequence, captured by a stereo vision system in an indoor office environment using a resolution of 320x240 pixels (see Figure 4a).

In this case, the images were automatically obtained by a mobile robot navigating inside an office building. Each sequence corresponds to the trip of the robot crossing the long corridor displayed in the map of Figure 2. The images of this dataset do not feature relevant illumination changes but they include moving objects.

As a third testing set, we also include one of the datasets available from the workshop: From Sensors to Human Spatial Concepts (FS2HSC) [26]. We use the dataset corresponding to Home 2, run 1. This dataset includes 1401 omnidirectional images that were taken by a camera with a hyperbolic mirror. The images were transformed to panoramic images with a resolution of 320x240 pixels using the toolbox available in the website. The dataset also provides information about the position and time where each image was acquired.

4.2. *Overall Performance*

We independently applied our approach to each of the datasets mentioned above. After processing the different test sets, we obtained a set of databases containing a total of 60 landmarks for the home dataset, 200 landmarks for the first sequence of the office dataset, and 951 landmarks for the FS2HSC dataset. It is interesting to note that although the office and FS2HSC datasets have a similar number of images, the number of landmark selected in the case of the FS2HSC dataset is significantly larger. This is explained by the large number of objects and clutter present in the images of the FS2HSC dataset. In contrast, in the office dataset large image areas, mainly in the main corridor, consist of planar walls without many salient features that can be used as useful visual landmarks.

Figure 4 shows images that highlight the typical landmarks detected in the sequences. Rectangles with identification labels are superimposed around new landmarks (ADD) and recognized landmarks (REC). In particular, Figure 4c) shows two landmarks that are aggregated to the database, while Figure 4d) shows the recognition of these landmarks in a posterior frame captured from a different position.

To test the landmark recognition capabilities of our system, we measure its recall performance. Given that our final goal is to achieve autonomous robot navigation, recall is measured as the number of times that a selected landmark is correctly recognized in posterior times that the robot visits a given place of the environment. To run this experiment, we use the 10 sequences of the office dataset. Given that in each sequence the robot visits each place only once, the ideal recall is 10. Table 1 shows the average recall performance of our system that has a value of 5.8. We analyze this result in more detail in Section 4.4.

It is important to note that in the current version of our method, we still do not incorporate a mechanism to filter out unreliable landmarks associated to dynamic targets, such as moving people. Therefore, in the estimation

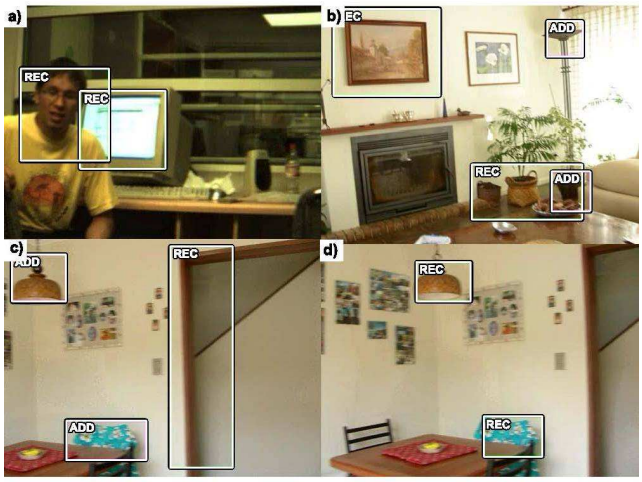

Fig. 4. a-b) Images from the test sets used in this work. Examples of typical landmarks detected by the system are highlighted. c) Two selected landmarks are added (ADD) to a database. d) The two selected landmarks are recognized (REC) in a posterior frame taken from a different point of view.

Table 1
Average Recall Performance of Our Complete Approach

Dataset	Ideal recall	Av. recall of our system
Office	10	5.8

of the average recall, we only consider landmarks that are correctly recognized in at least two independent image sequences.

4.3. *Specific Experiments*

4.3.1. *Impact of combining multiple segmentations*

In our first test, we demonstrate that the different segmentation algorithms present a dissimilar performance. To show this, we use the home dataset to count the number of successful recognitions independently achieved by the color and the saliency-based segmentations. Figure 5 shows the result of this test for 50 of the landmarks stored in the databases. The differences in heights on each pair of adjacent bars show that there is a large variation in the performance of each segmentation algorithm. We observe a similar behavior when we include the stereo-based segmentation and when we apply a similar test to the other datasets.

The next test is oriented to measure the impact of the segmentation step on the performance of the system. To achieve this, we compare a case where we use a segmentation step with a situation where not segmentation is used. In this last case, the area spanned by each landmark is set by using a fixed sized bounding box around each area detected by the visual attention step. Several bounding box sizes were tested. We show here the results obtained with the most suitable size. Table 2 shows the results obtained for the three datasets used in this work. We can see that by combining multiple segmentations algorithms that adaptively defines a suitable image patch spanned by each landmark, we obtain a considerable impact on the performance

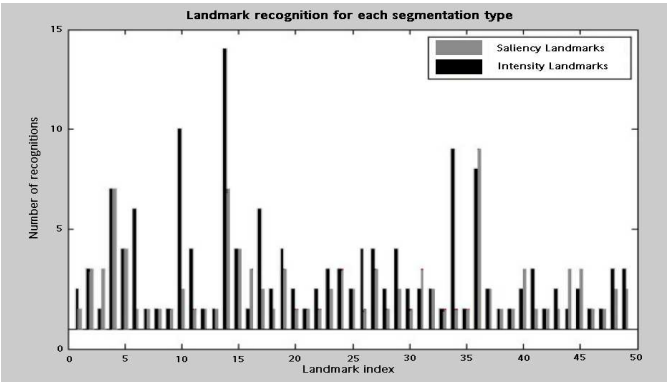

Fig. 5. Landmark recognitions for the home dataset for the color and the saliency-based segmentations. Recognition performance depends on the segmentation algorithm used.

Table 2
Number of Landmarks Correctly Recognized

Dataset	No segmentation	With segmentation Step	% Increase
Home	26	90	246.15
Office	180	856	375.56
FS2HSC	152	587	286.18

Table 3
Number of Landmark Correctly Recognized

Dataset	Saliency-based segm.	Color-based segm.	Both
Home	248	270	358
Office	272	275	473

of the system.

Finally, we use the home dataset and one sequence of the office dataset to compare the recognition results achieved by combining the color and saliency-based segmentation algorithms with respect to using each segmentation algorithm alone. As expected, Table 3 shows an increase in the recognition performance when we combine both segmentation algorithms.

4.3.2. *Impact of using an estimation of the robot position*

The next experiment shows the impact on the efficiency of the system of including an estimation of the position of the robot. In this case, we use the office dataset running all the segmentations algorithms. We compare a case where all the selected landmarks are stored in a single database against a case where the environment is divided into several databases according to the topological map representation presented in Figure 2-b. Figure 6 shows that as the number of landmarks inserted in the database increases, the impact of reducing the search scope to recognize a candidate landmark becomes highly significant.

4.3.3. *Impact of adaptively combining the segmentation algorithms*

This experiment is designed to demonstrate the advantages of including the second top-down mechanism. This mechanism adaptively controls the relevance of each segmentation algorithm in the recognition of each landmark. To achieve this, we compare the average recall performance achieved by including an adaptive combination of the segmentation algorithms with respect to including a fixed scheme to combine these algorithms. In this last case, we calculate the value of P_{match} by just averaging the similarity scores of the 3 segmentation algorithms available. Table 4 shows the results. We observe that by adaptively combining the segmentation algorithms, we are able to achieve a significant increment in the recall performance of the system.

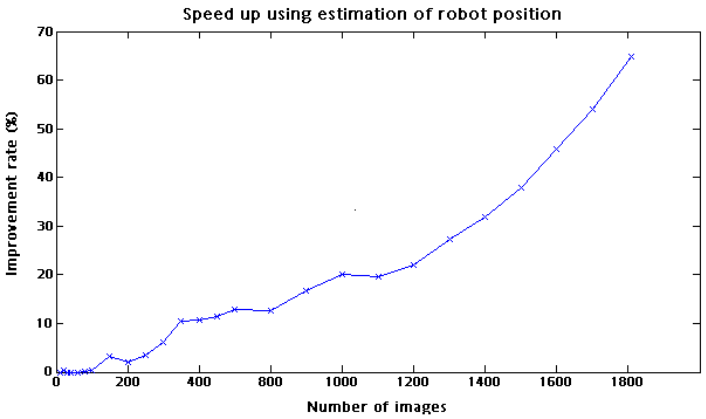

Fig. 6. Relative speed up in the processing time when reducing the spatial scope for searching for landmark matches for the different number of images. As expected the percentage of the speed up grows with the number of images processed by the system.

Table 4
Average Recall Performance of Fixed vs Adaptive Combination of Similarity Scores for the Segmentation Algorithms

Dataset	Fixed combination	Adaptive combination	Ideal recall
Office	3.7	5.8	10

4.4. *Discussion*

In our experiments, we notice that although our approach uses no semantic information about the objects in the scenes, the landmarks selected by the system often have a good correlation to the objects perceived by a human being. Furthermore, as our goal is not object recognition but robot localization, the system can afford some degree of redundancy by storing several different views of the same object as independent landmarks. As an example, Figure 7 shows a case where different views of a sofa, selected as a landmark from the FS2HS dataset, are stored in the corresponding database as three different landmarks.

With respect to the overall performance of the system, the average recall of 5.8, obtained for the sequences of the office dataset, results suitable to provide useful information to support robot navigation tasks. As an example, Figure 8 shows the positions of some of the landmarks recognized in one of the sequences of the office dataset. The distribution of the landmarks correctly identified by the system, span-

Fig. 7. Different views of a sofa, selected as a landmark from the FS2HS dataset, are stored in the database as three different landmarks.

ning most of the environment, shows the potential of using this approach to support tasks related to robot localization and mapping.

A further analysis of our experiments indicates three main reasons that can explain the gap between the average recall of 5.8 and the ideal recall of 10 for the office dataset. This analysis also suggest further work that can improve our current system. First, we notice that over different sequences the robot observes the environment from slightly different points of view. This produces that visual landmark associated to the same object are archived as independent landmarks instead of being recognized. This suggest, the use of clustering strategies to jointly store different views of a given landmark. Second, in some cases, we observe that new views of the relevant objects do not trigger the visual saliency attention mechanism. This suggest the use of top-down feedback that can be used to improve the performance of the attention mechanism. Third, we notice that due to occlusions and rotations of the robot, particularly at the extreme parts of the main corridor, some of the selected landmarks are not visible at all in some of the sequences. This suggest the use of active perception strategies that can guide the sensors of the robot to the most informative places.

With respect to the combination of the segmentation algorithms, we observe that the performance of these algorithms presents large variations depending on the characteristics of the selected landmark and the image conditions. Examples of the behavior of some of the segmentation algorithms can be seen in Figure 9 and Figure 10. Landmark borders are smoothed to avoid the artificial creation of high contrast, which can have a negative effect on the posterior

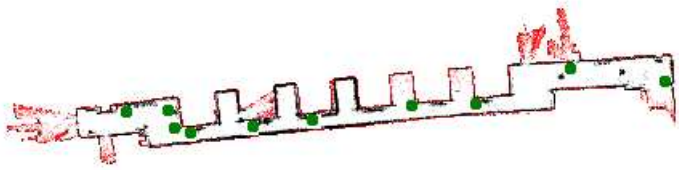

Fig. 8. The dots display the positions of some of the landmarks correctly recognized in one of the sequences of the office dataset. The great area coverage of the landmarks shows the potential of using this information for mobile robot localization.

Fig. 9. Segmentation of a lamp in the home sequence. Left: Using the saliency-based segmentation algorithm. Right: Using the color-based segmentation algorithm. The color-based segmentation totally fails because in this case it uses the background pixels to set its color model.

Fig. 10. Segmentation of a paint in the home sequence. Left: Using the saliency-based segmentation algorithm. Right: Using the color-based segmentation algorithm. Here the color-based segmentation presents better results, segmenting the paint without considering distracters or missing parts.

characterization of the landmark. Figure 9 shows a situation where the saliency-based segmentation provides a better result than the color-based segmentation, while Figure 10 shows the opposite case. The fact that we cannot predict which of these situations we are facing, validates the use of an approach that adaptively integrates multiple segmentations.

In our results, we observe that landmarks with lack of textural information are often not recognized, even when they are very salient in the original image and easy to distinguish from other objects in the scene using color or other properties. Figure 11 shows an example where a very salient and colorful t-shirt initially included in the database (bottom-left) is later missed in two occasions. The first failure (bottom-center) is due to intensive blurring originated from the camera motion. This eliminates the textural in-

Fig. 11. Pitfalls on recognition. Top: Original input image. Bottom-Left: segmented area of a t-shirt initially included in the landmark database. Bottom-Center: The landmark is not recognized in a posterior frame due to significant blur produced by camera motion. Bottom-Right: Again the landmark is not recognized due to significant differences when it is observed from a different point of view.

formation used by SIFT features to recognize the object. The second failure (bottom-right) is due to observing the t-shirt from a different point of view. This produces radically different SIFT features that confuse the system. This observation indicates a need for complementary feature descriptors, that can handle cases where SIFT features do not provide suitable results.

Finally, it is important to note that, although our main goal is to create a visual landmark detection system to achieve robot localization, in our approach we use information from a second localization system based on metric and topological information. We believe that this situation does not present conflicts, because both localization methods depend on different information types. On the contrary, these methods are highly complementary, and their synergistic combination can produce a more robust localization approach.

5. Conclusions and Future Work

In this work, we presented an unsupervised method for selection and posterior recognition of visual landmarks using images acquired by a mobile platform. Our results indicate that we achieved three main goals. First, we developed a robust bottom-up working solution based on three main steps: (i) Visual attention by means of a saliency algorithm, (ii) Landmark delimitation by means of multiple segmentations, and (iii) Landmark characterization by means of SIFT descriptors. Second, we demonstrated that by adaptively combining multiple segmentation algorithms, we can effectively increase the performance of the system. Third, we proved the relevance of including top-down feedback mechanisms to increase the efficiency and robustness of the approach.

In terms of the overall performance of the system, the average recall obtained shows the potential of this approach to support task related to autonomous mobile robot navigation. This is granted by the usual redundancy in the number of landmarks visible by the robot from each possible position. Furthermore, in our tests we also observed that different views of a relevant object in the environment can be archived as independent landmarks. Although this adds some inefficiencies to the system, it still provides useful information to the task of autonomous robot navigation.

In terms of the integration of multiple segmentations algorithms, we were able to significantly increase the average number of recognitions of the selected landmarks with respect to the case with no segmentation, and the case of using a single segmentation algorithm. In particular, with respect to the case that does not consider a segmentation algorithm, the increase rate was over 240% for the home and FS2HSC dataset, where we combined two segmentation algorithms, and over 370% in the case of the office dataset, where we combined three different segmentation algorithms.

In terms of the top-down feedback mechanisms used by the system, we demonstrated that the adaptive integration of the segmentation algorithms using a set of importance weights, plays a relevant role in the performance of the system. Furthermore, by using a probabilistic estimate of the robot position, we were able to speed up the processing time of the system by more than 60%.

As a future work, an important issue will be to delete certain landmarks in the databases, if they are not recognized for a long time. This will help to discard certain spurious landmarks that can be associated to wrong segmentations or dynamic objects. In this sense, motion cues can also be included to filter-out non stable image regions as candidate landmarks, such as a human walking close to the robot. Another important addition will be to augment the current adaptation scheme used to combine the influence of each segmentation algorithm. As an example, some segmentation algorithms can be activated and deactivated dynamically, in order to save computational resources, and also to diminish the information actually stored in the databases. Finally, we can also increase the influence of the top-down feedback to other steps of the algorithm besides segmentation. As an example, it is possible to include a top-down feedback mechanism to bias the influence of each saliency map in the detection of each landmark.

6. ACKNOWLEDGMENTS

This work was partially funded by FONDECYT grant 1050653. We would like to thank Anita Araneda, Domingo Mery and Laurent Itti for valuable comments.

References

[1] A. Araneda, S. Fienberg, and A. Soto. A statistical approach to simultaneous mapping and localization for mobile robots. *Annals of Applied Statistics*, 1(1):66–84, 2007.

[2] T. Collett. Landmark learning and guidance in insects. *Proc. of the Royal Society of London, series B*, pages 295–303, 1992.

[3] J. Duncan. Integrated mechanisms of selective attention. *Current Opinion in Biology*, 7:255–261, 1997.

[4] A. Elfes. *Occupancy grids: A Probabilistic Framework for Robot Perception and Navigation*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1989.
[5] R. Gonzalez and R. Woods. *Digital Image Processing, 2nd ed.* Addison-Wesley, 2002.
[6] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conferences*, pages 147–152, 1988.
[7] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
[8] N. Karlsson, L. Goncalves, M. Munich, and P. Pirjanian. The vSLAM algorithm for navigation in natural environments. *Korean Robotics Society Review*, 2(1):51–67, 2005.
[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
[10] D. Marr. *Vision*. Freeman Publishers, San Francisco, 1982.
[11] D. Mery and F. Pedreschi. Segmentation of colour food images using a robust algorithm. *Journal of Food engineering*, 66:353–360, 2004.
[12] A. Rabinovich, A. Vedaldi, and S. Belongie. Does image segmentation improve object categorization? Technical Report CS2007-0908, University of California, San Diego, 2007.
[13] J. Rieser and S. Kramer. The calibration of space perception as a basic for navigation. In *AAAI Spring Symposium Series*, pages 65–69, 1989.
[14] V. Roth and B. Ommer. Exploiting low-level image segmentation for object recognition,. In *Pattern Recognition, Symposium of the DAGM, LNCS 4174*, 2006.
[15] C. Siagian and L. Itti. Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, IROS*, 2007.
[16] A. Soto, M. Saptharishi, J. Dolan, A. Trebi-Ollennu, and P. Khosla. CyberATVs: dynamic and distributed reconnaissance and surveillance using all terrain UGVs. In *Proc. of the Intl. Conf. on Field and Service Robotics*, pages 329–334, 1999.
[17] Y. Sun and R. Fisher. Hierarchical selectivity for object-based visual attention. In *Proc. 2nd Biologically Motivated Computer Vision Workshop, BMCV*, 2002.
[18] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146:77–123, 2003.
[19] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. Cambridge University Press, New York, 2006.
[20] T. Tomasz Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference (BMVC)*, 2007.
[21] J. K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.
[22] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):37–154, 2004.
[23] D. Walther, D.R. Edgington, and C. Koch. Detection and tracking of objects in underwater video. In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition CVPR*, pages 544–549, 2004.
[24] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100:41–63, 2005.
[25] B. Williams, G. Klein, and I. Reid. Real-time SLAM relocalisation. In *Proc. Intl. Conference on Computer Vision, ICCV*, 2007.
[26] Z. Zivkovic and J. Kosecka. Workshop: From sensors to human spatial concepts. In *In Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, IROS*, 2007.