

Improving the Selection and Detection of Visual Landmarks Through Object Tracking

P. Espinace and A. Soto

Department of Computer Science, Pontificia Universidad Catlica de Chile
Casilla 306, Santiago 22, Chile

[pespinac, asoto]@ing.puc.cl

Abstract

The unsupervised selection and posterior recognition of visual landmarks is a highly valuable perceptual capability for a mobile robot. Recently, in [6], we propose a system that aims to achieve this capability by combining 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. The top-down feedback is based on 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 different segmentation algorithms in the recognition of each landmark. In this paper we explore the benefits of extending our previous work by including a visual tracking step for each of the selected landmarks. Our intuition is that the inclusion of a tracking step can help to improve the model of each landmark by associating and selecting information from its most significant views. Furthermore, it can also help to avoid problems related to the selection of spurious landmarks. Our results confirm these intuitions by showing that the inclusion of the tracking step produces a significant increase in the recall rate for landmark recognition.

1. Introduction

Autonomous point to point navigation is a key requirement for most practical applications of mobile robots in natural environments. In this context, 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. 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 [5]. Although, this type of approaches has shown a high degree of success when operating in real time in natural environments [18], 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 [21] [11] 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.

In the case of mobile robots, the unsupervised selection and posterior recognition of relevant visual landmarks is a highly valuable perceptual capability to successfully deal with the complexity of an unstructured natural environment. In this respect, in a previous approach [6], we developed 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. In this way, the resulting system is able to select interesting, meaningful, and useful landmarks that can be used by a mobile robot to achieve indoor autonomous navigation.

The bottom-up approach for the selection of candidate

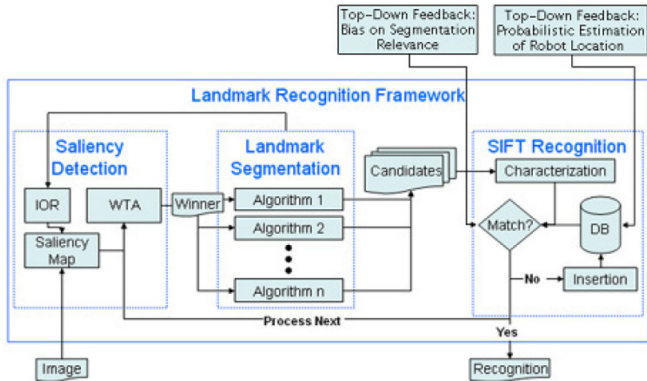


Figure 1. Schematic operation of our previous approach. This method integrates attention, segmentation, and characterization mechanisms in addition to top-down feedback for unsupervised recognition of relevant visual landmarks.

landmarks is based 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 recognition 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.

The bottom-up landmark selection approach is complemented with two modalities of top-down feedback that increase the efficiency and robustness of our approach. First, to increase efficiency, an estimation of the robot position is used 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 [18] using a metric map representation augmented with topological information. Second, to increase robustness, a record of the previous successful recognitions of each landmark is kept. This information is used to bias the influence that each bottom-up segmentation plays in the recognition, by adaptively updating a set of weights that control the relevance of each segmentation in the recognition of each landmark. A diagram that shows how the described system works can be seen in figure 1

In this paper, we extend our previous approach by including a visual tracking step that keeps track of each selected landmark in consecutive image frames. Using this step, we are able to associate different views of a single landmark. Our intuition is that the inclusion of a tracking step can help to improve the model stored by the system for each landmark by selecting information from its most sig-

nificant views. Furthermore, it can also help to avoid problems related to the selection of spurious landmarks. Our results confirm these intuitions by showing that the inclusion of the tracking step produces a significant increase in the recall rate for landmark recognition.

This document is organized as follows. Section 2 provides background information and describes related previous work. Section 3 presents a deeper description of our previous approach. Section 4 provides full details of the addition of the new tracking step. Section 5 presents our empirical results. Finally, Section 6 presents the main conclusions of this work and future avenues of research.

2. Related Work

In the computer vision literature, there is an extensive list of works that, individually, target the problems of attention, segmentation, characterization and visual tracking. We briefly review some relevant works in each area. We also review some relevant works concerning selection and recognition of visual landmarks and visual object tracking 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. [20] selectively tuned neuron models at the salient location with top-down mechanisms and winner-take-all networks. Itti et al. [9] introduced a model for selecting locations from a saliency map according to decreasing saliency. Sun and Fisher [16] 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 [17] also extended Duncan’s Integrated Competition Hypothesis [4] 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 [7], 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. [13], and, recently, Malisiewicz and Efros [19], have empirically shown the advantages of

combining multiple segmentations algorithms over using a single one.

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 [8] presented their corner detector. Recently, Lowe [11] presented a refinement of this idea, called the Scale Invariant Feature Transform (SIFT), which gained great popularity due to its success in several applications.

Visual tracking is the process of estimating the position of an object that follows a certain trajectory in a sequence of images. A great number of tracking techniques have been developed for this purpose, such as Mean Shift based techniques [3][26] and Particle Filter based techniques [2][12]. These techniques use different features extracted from the target object, such as color or shape features.

There is also related work in the area of simultaneous localization and mapping (SLAM) using visual sensors. Here, most current works [10], [25] consider the complete input image as the relevant area where a feature detector is applied. As an example, Karlsson et al. [10] 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.

Among approaches related to robot navigation that use visual attention mechanisms, Walther et al. [22] used a bottom-up approach to detect moving objects using a remotely operated underwater vehicle. In their work, Walther et al. also demonstrate the influence of visual saliency in recognition. Recently, Siagian and Itti [15] present preliminary results about a system that combines gist and attention mechanisms to achieve outdoor robot localization.

Among visual tracking techniques applied to robot navigation, Wang et al. introduced the concept of Simultaneous localization, mapping and moving object tracking (SLAMMOT), that involves both, simultaneous localization and mapping (SLAM) in dynamic environments, and detecting and tracking these dynamic objects [24][23]. These approach focus in tracking of moving objects in dynamic environments, but do not address tracking of static natural visual landmarks for mobile robot localization.

3. Our Previous Approach

In this section, we provide a general overview of our previous approach to select and detect relevant visual landmarks, for further detail see [6]. The system is based on two main parts: i) A bottom-up approach to select and recognize landmarks. This approach integrates visual attention, area segmentation, and landmark characterization mechanisms. ii) A top-down mechanisms added to improve robustness and efficiency to the system.

3.1. Bottom-Up Approach

Figure 1 shows a schematic view of the 3 main steps of the bottom-up part of the approach. It integrates attention, segmentation, and characterization mechanisms.

3.1.1 Attention

The bottom-up saliency map of Itti et al. [9] is used to extract salient locations from an input image. The original algorithm is slightly modified by introducing an adaptive scheme that dynamically selects an appropriate number of relevant salient locations.

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, the conditions of each input image can not be predicted a priori, a multiple segmentation algorithm is used to increase the adaptability and robustness of the landmark recognition algorithm.

Three different existing segmentation algorithms are used 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, their behaviors differ depending on the input conditions.

3.1.3 Characterization

For the characterization of the segmented patches, the SIFT feature extraction algorithm is used [11]. 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 obtain an unsupervised selection and subsequent recognition of landmarks, an integration of the three steps described above is needed. The integration 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 landmarks 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, a list of candidate landmarks corresponding to the resulting segmented areas around the selected saliency regions is obtained.

Once obtained the candidate landmarks, SIFT features for each available segmentation of each landmark in the list are extracted. These features are then compared to the features of the landmarks in the 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, a similarity score based on a nearest-neighbors technique, as described in [11], is used. Given that, for each landmark, 3 possible descriptions corresponding to each of the available segmentations are kept, the respective similarity scores are combined 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, the match is reported and the record of the number of times the landmark has been successfully recognized by the different segmentations is modified. This information is used later 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, this landmark is added to the database provided that certain constraints are satisfied. The premise is to keep in the database only landmarks that are highly distinctive and easy to detect. According to this, only landmarks that present SIFT features with an strength about a fixed threshold are included in the database. Furthermore, the number of relevant SIFT features for each landmark in the database must be greater than 12 SIFT features to include it in the database. Developed experiments indicate that landmarks with fewer features usually do not produce enough matches to trigger a robust recognition.

3.2. Top-Down Feedback

As pointed in [6], 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, the top-down mechanisms that are use to alleviate these problems are described.

3.2.1 Use of an estimate of robot position

One of the more computationally expensive parts of the bottom-up 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, the database is divided 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, the location of the potential visible landmarks is obtained 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, the metric map representation is augmented 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. The topological map is labeled manually.

One landmark database is associated 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, the search is restricted to the databases corresponding to topological nodes associated to grid cells with high likelihood under the current estimation of the robot position.

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.

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. To simplify the notation from now on, we drop the superindex i .

As mentioned before, to achieve the recognition of a candidate landmark, the similarity score between SIFT descriptions of objects proposed in [11] is used. 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, the results of the recognition is used to update its importance weights. The update is performed 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. Extended Approach

In this section we present a variant to the previously described approach that uses visual landmark tracking to complement the landmark selection and recognition system of the original implementation. The tracking algorithm works in the following way. When a new landmark is selected, a particle filter based tracker is applied to the following frames to track each selected landmark. The particles filter

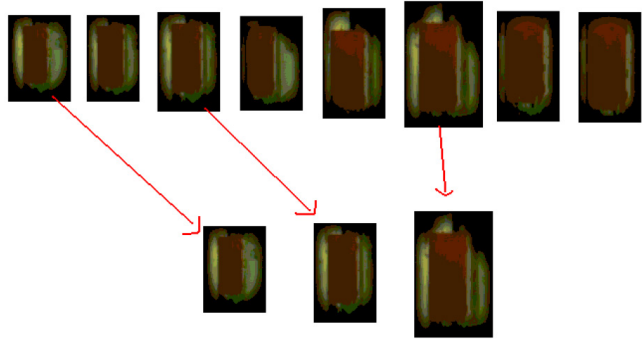


Figure 2. Views selection. From the eight different views available for the same object, only three are selected to be stored.

uses a likelihood function that is based on a color similarity measure in RGB space and a propagation function that implements a stationary gaussian function around the previous object location, with a variance of 25 pixels. The tracking result determines the center location of each landmark in each of the frames.

Using the tracking results, we apply each of the available segmentations algorithms around the area of the new location. The result of these segmentations is considered a new view of the corresponding landmark, and not a new landmark. Additionally, the area of the new view is not considered for new landmark selections in the corresponding frame, in order to avoid storing repeated landmark sequences. In this way, we build a sequence of views that represent a single landmark, which is a natural way to represent objects that can be seen from different points of view.

As many of the views of the object can be very similar, we only store in the database SIFT point descriptions from views that are substantially different. The selection of these views is performed in the following way. The first view, that corresponds to the original landmark selection, is always kept. Subsequent views are only kept if the SIFT points similarity between successive views exceeds the threshold used for landmark recognition in the original implementation (see [6]). Figure 2 shows an example where there are eight views of a particular landmark, but the system decides to keep in the database just the SIFT points of three of these views.

To decide if a new candidate landmark corresponds to an existing one, we use the new landmark representation to compare the new candidate with the sequence of views of the existing landmarks. In order to build a comparison metric between a new potential landmark and the sequence of views of a previously stored landmark, we built a training set using stored landmark sequences and new views of potential landmarks. We then performed a cross validation routine to determine the optimal number of neighbors and the optimal detection threshold used to specify a new recognition.

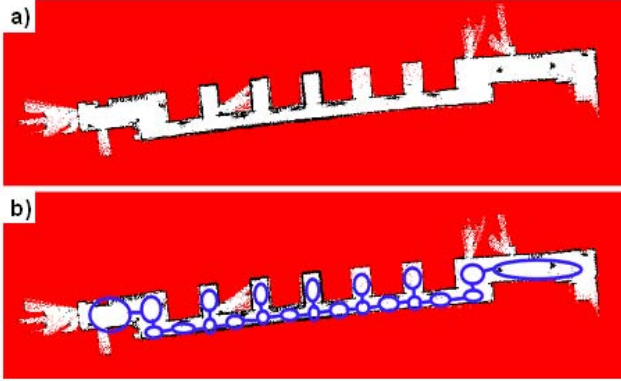


Figure 3. Map representations of our Computer Science Department. a) Metric map. b) Topological map superimposed over the metric map.

5. Results

In order to test the impact of adding the tracking method to the original implementation, we now compare the results of our system with the results obtained with the original implementation.

To perform our experiments, we use a dataset that consists of 10 video sequences, with an average number of 1500 image-pairs per sequence, captured by a stereo vision system in an indoor office environment using a resolution of 320x240 pixels. The images were automatically obtained by a mobile robot navigating inside our Computer Science Department (DCC), a typical office building environment. Figure 3-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 3-b shows the corresponding topological map, superimposed over the metric map. Each sequence corresponds to the trip of the robot crossing the long corridor displayed in the map of figure 3. The images of this dataset do not feature relevant illumination changes, and do not correspond to any of the datasets used in the original implementation.

Comparison results show that our method gives similar results to the original approach for most of the existing landmarks, however, it does present an increase in the recall rate of landmarks that, despite of being distinctive features in the environment, are not consistently recognized as such by the visual attention mechanism. This is the case of objects that are seen from several side views, instead of frontal views, making the visual attention mechanism to recognize it only under certain angles. Using the tracking method, we can estimate the object position in all subsequent frames and represent the object in a better way. An example of an object with two different selected views can be seen in figure 4. This object presents a poor recall rate of 30% in the original implementation. Using the tracking method and the se-



Figure 4. Example of a landmark with two selected views.

Table 1. Average Recall Performance of the Original Approach vs the Proposed Approach

Dataset	Original	Proposed	Ideal recall
Office	7.2	8.6	10

quence of views, the recall rate for the same object increases to 60%.

Overall results show an increase in performance using our method when compared to the original implementation. Given that in each sequence the robot visits each place only once, we can state that the ideal average recall is 10. Table 1 shows the average recall performance of the proposed system with the tracking method and the original system without the tracking method.

In terms of the number of stored landmarks, the tracking method allows the overall system to store a fewer number of landmarks. This is because in the original implementation, if a landmark recently selected is not recognized in a posterior frame, it is stored as a new landmark. Using tracking, the object is more robustly detected in posterior frames so it is not stored repeatedly as a different landmark, but only as a different view. Despite of this, the amount of stored data is similar in both implementations, as in our approach each landmark stores more information than in the original implementation.

In terms of search efficiency, our method is also similar in terms of computational cost when compared with the original implementation. This is because our method reduces the number of comparisons by storing fewer landmarks, but increases the cost of each comparison by including several views and a nearest neighbor method.

In terms of selection efficiency, our method adds complexity to the method, as it incorporates the tracking routine, that has a computational cost of $O(Np * No * Is)$, where Np is the number of samples in the particle filter algorithm, No is the number of objects currently being tracked, and Is is the average size of the images representing the objects.

6. Conclusions and Future Work

In this work, we extend our previous work on visual landmark selection and recognition by adding a visual tracking step for new selected landmarks. Our results indicate that we achieved two main goals. First, we obtained an improved landmark representation by storing a sequence of views of each selected landmark, which is a natural way of representing objects that can be seen from different positions. Second, we significantly improved the recall rate of landmark recognition, which is an important fact for mobile robot navigation tasks.

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. The tracking process can also be used to help in this task, as we can discard objects that are originally selected, but can not be tracked, considering that these objects may correspond to noisy or moving objects.

References

- [1] A. Aranedo, 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. 6
- [2] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188, 2002. 3
- [3] D. Comaniciu and V. Ramesh. Mean shift and optimal prediction for efficient object tracking. In *Proc. Intl. Conf. on Image Processing*, 2000. 3
- [4] J. Duncan. Integrated mechanisms of selective attention. *Current Opinion in Biology*, 7:255–261, 1997. 2
- [5] A. Elfes. *Occupancy grids: A Probabilistic Framework for Robot Perception and Navigation*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1989. 1
- [6] P. Espinace, D. Langdon, and A. Soto. Unsupervised identification of useful visual landmarks using multiple segmentations and top-down feedback. *Robotics and Autonomous Systems (to appear)*. 1, 3, 4, 5
- [7] R. Gonzalez and R. Woods. *Digital Image Processing, 2nd ed.* Addison-Wesley, 2002. 2
- [8] C. Harris and M.J. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conferences*, pages 147–152, 1988. 3
- [9] 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. 2, 3
- [10] 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. 3
- [11] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 3, 4, 5
- [12] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. *Lecture Notes in Computer Science*, LNCS 2350:661–675, 2002. 3
- [13] A. Rabinovich, A. Vedaldi, and S. Belongie. Does image segmentation improve object categorization? Technical Report CS2007-0908, University of California, San Diego, 2007. 2
- [14] V. Roth and B. Ommer. Exploiting low-level image segmentation for object recognition. In *Pattern Recognition, Symposium of the DAGM, LNCS 4174*, 2006. 2
- [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. 3
- [16] Y. Sun and R. Fisher. Hierarchical selectivity for object-based visual attention. In *Proc. 2nd Biologically Motivated Computer Vision Workshop, BMCV*, 2002. 2
- [17] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146:77–123, 2003. 2
- [18] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. Cambridge University Press, New York, 2006. 1, 2
- [19] T. Tomasz Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference (BMVC)*, 2007. 2
- [20] J. K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nufflo. Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995. 2
- [21] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):37–154, 2004. 1
- [22] 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. 3
- [23] C. Wang, C. Thorpe, M. Thrun, S. and Hebert, and H. Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. *The International Journal of Robotics Research*, 26(9):889–916, 2007. 3
- [24] C. Wang, C. Thorpe, and S. Thrun. Online simultaneous localization and mapping with detection and tracking of moving objects: theory and results from a ground vehicle in crowded urban areas. In *In Proc. IEEE/RSJ Intl. Conf. on Robotics and Automation, ICRA*, 2003. 3
- [25] B. Williams, G. Klein, and I. Reid. Real-time SLAM relocation. In *Proc. Intl. Conference on Computer Vision, ICCV*, 2007. 3
- [26] C. Yang, R. Duraiswami, and L. Davis. Efficient mean-shift tracking via a new similarity measure. In *Proc. Intl. Conf. on Computer Vision and Pattern Recognition CVPR*, 2005. 3