Published in Proceedings.
6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), 309-314, 2005
which should be used for any reference to this work

Robot Self-localization Using Visual Attention

Nabil Ouerhani and Heinz Hügli
Institute of Microtechnology
University of Neuchâtel
Rue Breguet 2
CH-2000 Neuchâtel, Switzerland
{*Nabil.Ouerhani,Heinz.Hugli*}*@unine.ch*

***Abstract*— This paper presents a robot self-localization method based on visual attention. This method takes advantage of the saliency-based model of attention to automatically learn configurations of salient visual landmarks along a robot path. During navigation, the visual attention algorithms detect a set of conspicuous visual features which are compared with the learned landmark configurations in order to determine the robot position on the navigation path. More specifically, the multi-cue attention model detects the most salient visual features that are potential candidates for landmarks. These features are then characterized by a visual descriptor vector computed from various visual cues and at different scales. By tracking the detected features over time, our landmarks selection procedure automatically evaluates their robustness and retains only the most robust features as landmarks. Further, the selected landmarks are organized into a topological map that is used for self-localization during the navigation phase. The self-localization method is based on matching between the currently detected visual features configuration and the configurations of the learned landmarks. Indeed, the matching procedure yields a probabilistic measure of the whereabout of the robot. Thanks to the multi-featured input of the attention model, our method is potentially able to deal with a wide range of navigation environments.**

I. Introduction

It is generally agreed that vision is one of the richest source of information for humans but also for machines that need to interact with their operating environment. Therefore, vision is becoming a more and more indispensable component of autonomous robot navigation systems. Particularly, the landmark-based navigation paradigm makes extensive use of the visual information about the navigation environments.

The earliest works that introduced vision into landmark-based robot navigation used, essentially, artificial landmarks which are easily recognizable by the robot. The work presented in [1], for example, used black rectangles with white dots as landmarks. It is obvious that this approach requires a modification of the environment which is not always feasible or desirable. More recent works introduced novel approaches that use more natural landmarks in order to solve the problem of robot localization. Fluorescent tubes, for instance, have been used as landmarks in [2], whereas the work presented in [3] used posters and doorplate for the same purpose. These approaches require, however, precise knowledge about the environment and are too specific to the considered environment.

More recently, more general approaches have been proposed. They are based on the idea that robots should find the landmarks by themselves [4], [5]. That is the robot explores its navigation environment and automatically selects a set of features that can be considered as robust but also distinctive landmarks. Numerous works have dealt with the feature extraction problem with the aim to derive appropriate landmarks. In [6], [7] the authors used intensity patches that are unique in the environment as landmarks, whereas vertical edges have been used in [8]. Others used color interest operator for the same purpose [9]. The Scale Invariant Feature Transform (SIFT) that extracts features from grey-scale images at different scales has been used in [10]. The work presented in [11] uses the fingerprint concept for selecting landmarks, that is a collection of features that are unique in the navigation environment. One of the most used feature detector, however, is the corner interest operator [12], [13].

It is noteworthy that most of the proposed feature detection methods for landmarks selection apply on gray-scale images and only few of them have an adaptive behavior. With adaptive behavior is meant, here, the ability of a method to automatically choose the feature detector (corner, color patches, intensity patches, ...) most appropriate to the considered environment for the landmark selection process. For instance, to use corners and vertical lines in indoor environment and color and intensity patches outdoor. Since adaptive behavior is one of the strengthes of biological vision systems, biologically inspired computational models of vision could be potential solutions to build adaptive landmark selection algorithms. Particularly, bio-inspired saliency-based visual attention models, which aim to automatically select the most salient and thus the most relevant information of complex scenes, could be useful in this context. Note that visual attention has been used to solve numerous other problems related to computer vision, like image segmentation [14], object tracking in dynamic scenes [15] and object recognition [16]. The usefulness of attention in real world applications is further strengthened by the recent realization of a real time visual attention system [19].

This paper reports a novel method for robot localization based on visual attention. This method takes advantage of the saliency-based model of visual attention at two different phases as shown in Figure 1. During a learning phase, the attention algorithms automatically select the most visually salient features along a navigation path, using various cues like color, intensity and corners. These features are

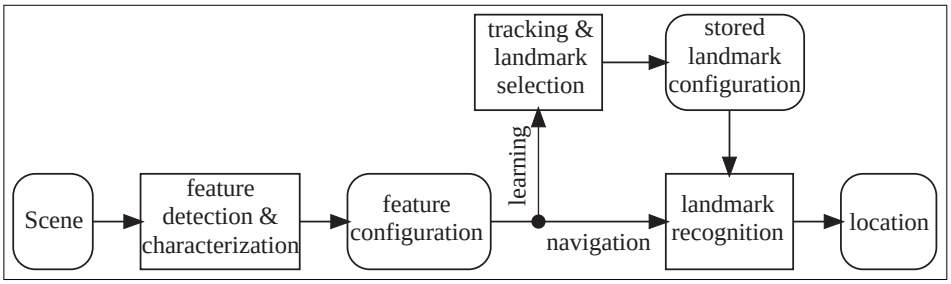


Fig. 1. Overview of the attention-based method for robot localization.

characterized by a descriptor vector whose components are computed from the considered cues and at different scales. They are then tracked over time in order to retain only the most robust of them as the representative landmarks of the environment. These landmarks are then used to build a topological map of the environment associated to the robot path. During a navigation phase the same attention algorithms computes visual features that are compared with the learned landmarks in order to compute a probabilistic measure of the robot position within the navigation path.

The remainder of the paper is organized as follows. Section II describes the landmark selection procedure that is based on the visual attention algorithms. In Section III the mapping process consisting in representation as well as the organization of the selected landmarks into a topological map is presented. The landmark recognition algorithms and the robot localization approach are described in Section IV. Section V reports some experimental results that show the potential of our method. Finally, conclusions and some future works are stated in Section VI.

II. Attention-based landmark selection

In the context of robot navigation reliable landmarks must satisfy two major conditions: uniqueness and robustness. On one hand, the landmarks must be unique enough in the environment so that the robot can easily distinguish between different landmarks. On the other hand, landmarks must be robust to conditions changes like illumination and view angle. We intend to solve the uniqueness condition by using an extended version of the saliency-based model of visual attention, whereas the robustness condition is provided by a persistency test of the landmarks based on a tracking procedure. These two solutions are described in the sections below.

A. Feature detection and characterization

In order to detect robust features, we use an extended version of the saliency-based model of visual attention. The saliency-based model of attention has been firstly reported in [17] and gave rise to numerous soft and hardware implementations [18], [19].

The standard model of attention computes a saliency map, that encodes the conspicuousness of image locations, according to the following scheme.

1) First, a number of visual cues are extracted from the scene by computing the cue maps F_j. The cues most used in previous works are intensity, color, and orientation. The use of these cues is motivated by psychophysical studies on primate visual systems. In particular, the authors of the model used two chromatic channels that are inspired from human vision, namely the two opponent colors red/green (RG) and blue/yellow (BY).
2) In a second step, each map F_j is transformed in its conspicuity map C_j. Each conspicuity map highlights the parts of the scene that strongly differ, according to a specific visual cue, from their surroundings. This operation that measures, somehow, the uniqueness of image locations is usually achieved by using a *center-surround*-mechanism which can be implemented with multiscale *difference-of-Gaussian*-filters. It is noteworthy that this kind of filters have been used by D. Lowe for extracting robust and scale-invariant features (SIFT) from grey-scale images for object recognition, stereo matching but also for robot navigation [10], [20].
3) In the third stage of the attention model, the conspicuity maps are integrated together, in a competitive way, into a *saliency map* $\mathcal{S}$ in accordance with equation 1.

$$\mathcal{S} = \sum_{j=1}^{J} \mathcal{N}(\mathcal{C}_j) \tag{1}$$

where $\mathcal{N}()$ is a normalization operator that promotes conspicuity maps in which a small number of strong peaks of activity are present and demotes maps that contain numerous comparable peak responses [18]. In fact $\mathcal{S}$ encodes the saliency and ,thus, the uniqueness of image locations according to used visual cues.
4) Finally the most salient parts of the scene are derived from the saliency map by selecting the most active locations of that map. The automatically selected locations are designated, henceforth, as **features**.

In a recent work [21], we extended the basic model of visual attention to consider also a corner-based cue computed according to the Harris approach [22] for saliency computation. This extension yielded a more unique and more robust features.

Once selected, each feature P_i is characterized by its spatial position $\mathbf{x}_i = (x_i, y_i)$ and a visual descriptor vector $\mathbf{f}_i$:

$$\mathbf{f}_i = \begin{pmatrix} f_1^i \\ .. \\ f_J^i \end{pmatrix} \quad (2)$$

where J is the number of the considered visual cues in the attention model and f_j^i refers to the contribution of the cue j to the detection of the feature P_i. Formally, f_j^i is computed as follows:

$$f_j^i = \frac{\mathcal{N}(C_j(\mathbf{x}_i))}{\mathcal{S}(\mathbf{x}_i)} \quad (3)$$

Note that $\sum_{j=1}^{J}(f_j^i) = 1$.

B. Feature tracking and landmark selection

In order to automatically select robust landmarks from the set of features computed above, the features undergo a persistency test. This test consists in tracking the features over an extended portion of the navigation path and only those features that have been successfully tracked long enough are considered as robust landmarks.

The basic idea behind the proposed algorithm is to build a trajectory T for each tracked feature. Each point of the trajectory memorizes the spatial information and the visual descriptor of the tracked feature at a given time.

Specifically, given the M features computed from the first frame, the tracking algorithm starts with creating M initial trajectories, each of which contains one of the M initial features. The initial features represent also the head elements of the initial trajectories. A new detected feature P_i is either appended to an existing trajectory (and becomes the head of that trajectory) or gives rise to a new trajectory, depending on its similarity with the head elements P_h of already existing trajectories. The dissimilarity between P_i and P_h is quantified by a distance $\mathbf{d}(P_i, P_h)$ that takes into account the spatial and the descriptors distances between the two features. Formally $\mathbf{d}(.)$ is defined in accordance with Equation 4.

$$\mathbf{d}(P_i, P_h) = \frac{\|\mathbf{f}_i - \mathbf{f}_h\|}{f_{norm}} + \frac{\|\mathbf{x}_i - \mathbf{x}_h\|}{x_{norm}} \quad (4)$$

where f_{norm} and x_{norm} are two normalization factors that can be determined empirically or learned from a set of image sequences.

Given the results of the tracking procedure, the next step of the method consists in selecting the most robust features as landmarks L. The basic idea is that the cardinality ($Card(T)$) of a trajectory i.e. its length directly determines whether the corresponding features are robust landmarks. Results reported in [21] shows the efficiency of this criteria to select robust landmarks.

III. Mapping

Once selected, the landmarks should be then represented in an appropriate manner in order to best describe the navigation environment along the robot path. In this work, a navigation path is divided into representative portions E_q. Each path portion E_q is represented by a key frame K_q which is described by a configuration of the landmarks as showed in Figure 2.

Three attributes are assigned to the landmarks of a key frame:

- the horizontal spatial order of the landmarks x_index_L
- the mean height $\overline{y}_L$ of each landmark
- the corresponding maximum deviation Δy_L,
- the mean descriptor vector $\overline{\mathbf{f}}_L$ of each landmark L
- the corresponding standard deviation $\mathbf{\Sigma}_L$.

Note that these attributes are computed within the corresponding path portion E_q. formally, a key frame K_q is defined as:

$$K_q = \{L_m \,|L_m \text{ appears in } E_q\} \text{ with } L_m = \begin{pmatrix} x\_index_{L_m} \\ \overline{y}_{L_m} \\ \Delta y_{L_m} \\ \overline{\mathbf{f}}_{L_m} \\ \mathbf{\Sigma}_{L_m} \end{pmatrix} \quad (5)$$

It is to emphasize that the mapping method, at its current version, does not consider the spatial relationships between individual path portions E_q and does not, thus, use contextual information about the navigation environment.

IV. Self-localization

The localization procedure aims, here, to find the current position of the robot during navigation, by determining the key frame that is most likely to harbor the robot. To do so, the robot computes the set of salient features from its current position and compares them with the learned landmark configurations. The matching determines the similarity of the current features with each key frame and therefore the likelihood of the current position. In this work, we use a voting technique to compute this likelihood.

A. Landmark recognition

Our landmark recognition method is based on the spatial and visual similarity between features detected during navigation and landmarks learned during learning. Further, the method uses the spatial relationships between features / landmarks as an additional constraint. Specifically, a set of three features $p = \{P_1, P_2, P_3\}$ ($P_i = (\mathbf{f}_i, \mathbf{x}_i)^T$) is compared with a set of three landmarks $l = \{L_1, L_2, L_3\}$. The feature set p matches the landmark set l if the single features P_i visually and spatially (according to height y) match the single landmarks L_i as described in Equation 4 and if, additionally, the horizontal spatial order of the three features is the same as that of the three landmarks. Formally:

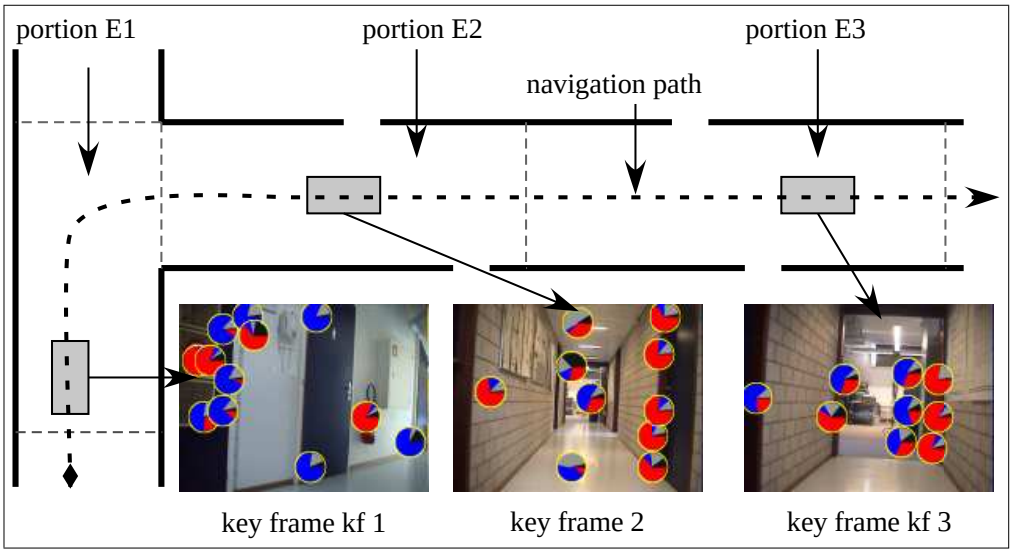


Fig. 2. Mapping. The navigation path is divided into representative portions E_q, represented by key frames KF_q, which are described by the corresponding landmarks.

$$\begin{aligned} match(p,l) &= True;\ \ if \\ \frac{\|\mathbf{f}_i - \overline{\mathbf{f}}_{L_i}\|}{f_{norm}} + \frac{|y_i - \overline{y}_{L_i}|}{y_{norm}} &< \sigma_i\ \forall i \in \{1,2,3\}\ \& \\ OrderX(p) &= OrderX(l) \end{aligned} \quad (6)$$

where f_{norm} and y_{norm} are two normalization factors determined as for Equation 4, σ_i is a combination of the height variation Δy_{L_i} and the descriptor vector standard deviation $\mathbf{\Sigma}_{L_i}$, and $OrderX()$ sorts a list of features or landmarks according to their x-coordinates.

B. Voting procedure

In order to determine which key frame is most likely to be the current position of the robot, the detected features vote for key frames which contain landmarks that match these features. Given the set of all features $P^\star = \{P_1, ..., P_m\}$ detected from the current position and the set of all key frame $K^\star = \{K_1, ..., K_n\}$, the voting procedure is achieved as follows. For each key frame K_i, each triplet $p = \{P_a, P_b, P_c\} \subset P^\star$ of features is compared to each triplet of landmarks $l = \{L_q, L_r, L_s\} \subset K_i$ belonging to K_i. If the matching between the features/landmarks triplets is correct, then a votes accumulator $A[i]$ corresponding to K_i is incremented by one vote. These steps are formalized in Algorithm 1.

The number of votes represents a measurement of the likelihood of a key frame to be the current position of the robot. This measurement is called, henceforth, localization score.

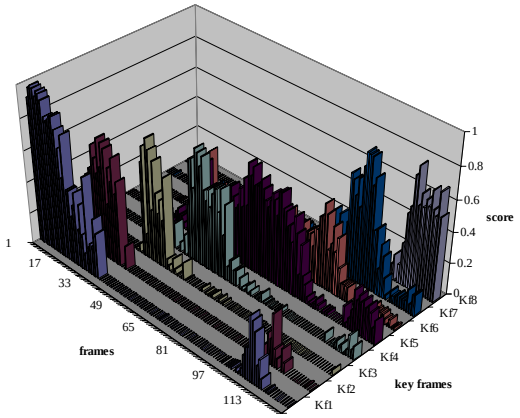
Algorithm 1 Voting procedure

Set of features $P^\star = \{P_1, ..., P_m\}$
Set of key frames $K^\star = \{K_1, ..., K_n\}$
Accumulator $A[n]$ /* votes accumulator for each key frame */
Initialize $A[i] = 0\ \forall i \in [0, n]$
for all $p = \{P_a, P_b, P_c\} \subset P^\star$ **do**
  for $i = 1\,..\,n$ **do**
    for all $l = \{L_q, L_r, L_s\} \subset K_i$ **do**
      if $(match(p, l) == TRUE)$ **then**
        A[i]++
      end if
    end for
  end for
end for

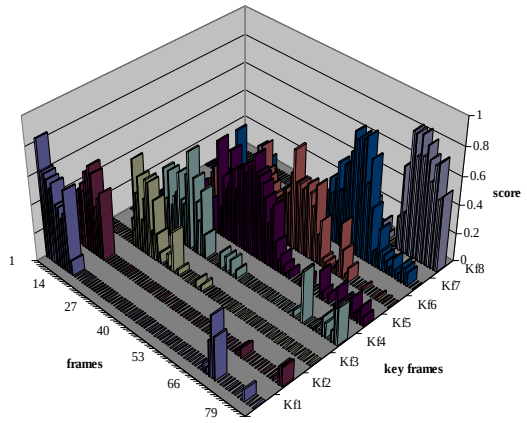
V. Experiments

This section reports some experiments that aim at evaluating the presented localization method. The experiments consists first in learning visual landmarks from a reference sequence of color images acquired by the robot while navigating along a certain path. These landmarks are organized into key frames. Then, a test sequence is acquired while the robot follows a similar path and a localization score is computed for each key frame. Since the robot starts almost from the same position for both sequences (reference and test), there exists an approximate timing between the two sequences. The distribution of the localization scores over the key frames as the robot moves forward is an indicator of success/failure of the localization method. That means that the first key frames must have the highest localization score at the beginning of the test sequence while the last ones must have lower localization scores (and vis-versa at the end of the test sequence).

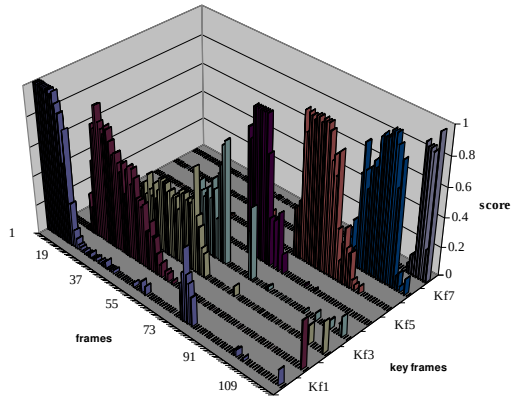
In our experiments, we consider two different navigation paths: path 1 where the robot moves along a corridor and path 2 chosen in a lab environment. For each path,



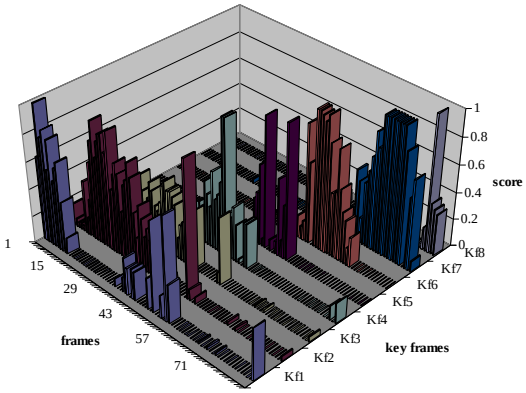
(a) Ref1–Ref1



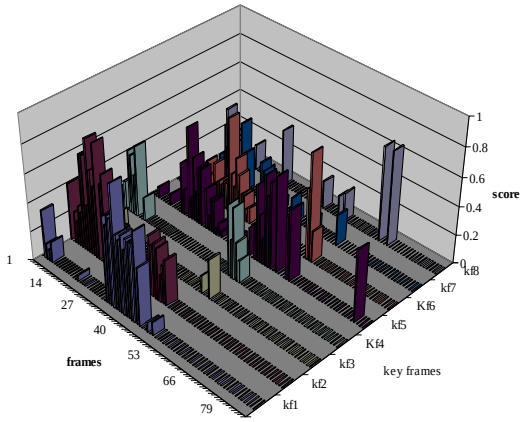
(b) Ref1–Test1



(c) Ref2–Ref2



(d) Ref2-Test2



(e) Ref1–Test2

Fig. 3. Experimental results (for more details, see text).

we acquired two different sequences, a reference sequence (Ref1/Ref2) and a test sequence (Test1/Test2). For both paths, the landmarks are learned from the reference sequences and are organized into 8 key frames. As test sequences, we used both the reference sequence itself and the test sequence.

Figure 3 summarizes the results of these experiments. (a) illustrates the localization score for path 1 when using the reference sequence as a test sequence (Ref1–Ref1). The localization scores for the same path but with Test1 as a test sequence are represented in (b). (c) and (d) of the same figure illustrate the same results with navigation path 2. For comparison purposes, the localization score using Ref1 from path 1 as reference sequence and Test2 from path 2 as test sequence is illustrated in (e). In this latter case the training and test sequences come from different environments. Note that at each frame the detected features are compared to the full set of key frames.

It can be seen that, for the four cases ((a)..(d)), the localization scores tend to form a diagonal distribution, which indicates the reliability of the localization method. Note that the distribution of the localization scores does not show any diagonal tendency when the reference and the test sequences stem from different navigation environments ((e)). It must be emphasized that no contextual information about the navigation environment is used for these experiments.

VI. Conclusions

This paper presents a visual attention-based method for robot self-localization. Using a saliency-based model of visual attention, the method automatically selects the most conspicuous and thus the most unique visual landmarks of the navigation environment. In order to retain only the most robust landmarks, the proposed method visually characterizes and tracks the landmarks over time and uses the quality of the tracking results as a robustness criterium. The so selected landmarks are organized into a topological map of the navigation path. During navigation, the method automatically detects the most salient features of the environment and compares them to the learned landmarks. The number of correct matching between the features and the landmarks is used as a score that measures the probability of the location of the robot. The effectiveness of our method is demonstrated by various experiments involving different navigation environments. In future work, we intend to extend our self-localization method to consider contextual information about navigation environments, which is expected to enhance the localization results. Further, our method will be integrated into a more complete stochastic navigation framework.

Acknowledgment

This work was partially supported by the CSEM-IMT Common Research Program

References

[1] J. Borenstein. *The Nursing Robot System*. PhD thesis, Technion Haifa, Israel, 1987.

[2] F. Launay, A. Ohya, and S. Yuta. Autonomous indoor mobile robot navigation by detecting fluorescent tubes. *International Conference on Advanced Robotics (ICAR '01), pp. 664-668*, 2001.

[3] J.B. Hayet, F. Lerasle, and M. Devy. A visual landmark framework for indoor mobile robot navigation. *International Conference on Robotics and Automation (ICRA), pp. 3942-3947*, 2002.

[4] S. Thrun. Finding landmarks for mobile robot navigation. *International Conference on Robotics and Automation (ICRA), pp. 958-963*, 1998.

[5] Y. Takeuchi and M. Hebert. Evaluation of image-based landmark recognition techniques. *Technical report CMU-RI-TR-98-20, Robotics Institute, Carnegie Mellon University*, 1998.

[6] Ulrich I. and I. Nourbakhsh. Appearance-based place recognition for topological localization. *International Conference on Robotics and Automation (ICRA), Vol. 2, pp. 1023-1029*, 2000.

[7] S. Thompson, T. Matsui, and A. Zelinsky. Localization using automatically selected landmarks from panoramic images. *Australian Conference on Robotics and Automation (ACRA),*, 2000.

[8] N. Ayache. *Artificial Vision for Mobile robots - Stereo-vision and Multisensor Perception*. MIT-Press, 1991.

[9] Z. Dodds and G.D. Hager. A color interest operator for landmark-based navigation. *AAAI/IAAI, pp. 655-660*, 1997.

[10] S. Se, D. Lowe, and J. Little. Global localization using distinctive visual features. *Intrenational Conference on Intelligent Robots and Systems, IROS, pp. 226-231*, 2002.

[11] A. Tapus, N. Tomatis, and R. Siegwart. Topological global localization and mapping with fingerprint and uncertainty. *International Symposium on Experimental Robotics*, 2004.

[12] A.J. Davison. *Mobile Robot Navigation Using Active Vision*. PhD thesis, University of Oxford, UK, 1999.

[13] A.A. Argyros, C. Bekris, and S. Orphanoudakis. Robot homing based on corner tracking in a sequence of panoramic images. *Computer Vision and Pattern Recognition Conference (CVPR), pp. 11-13*, 2001.

[14] N. Ouerhani and H. Hugli. MAPS: Multiscale attention-based presegmentation of color images. *4th International Conference on Scale-Space theories in Computer Vision, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2695, pp. 537-549*, 2003.

[15] N. Ouerhani and H. Hugli. A model of dynamic visual attention for object tracking in natural image sequences. *International Conference on Artificial and Natural Neural Network (IWANN), Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2686, pp. 702-709*, 2003.

[16] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and Ch. Koch. Attentional selection for object recognition - a gentle way. *2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), pp. 472-479*, 2002.

[17] Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.

[18] L. Itti, Ch. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.

[19] N. Ouerhani and H. Hugli. Real-time visual attention on a massively parallel SIMD architecture. *International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196*, 2003.

[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, Vol. 60, (2), pp. 91-110*, 2004.

[21] N. Ouerhani, H. Hugli, G. Gruener, and A. Codourey. Attentirobot: A visual attention-based landmark selection approach for mobile robot navigation. *International Workshop on Attention and Performance in Computational Vision (WAPCV 04), pp. 83-89*, 2004.

[22] C.G. Harris and M. Stephens. A combined corner and edge detector. *Fourth Alvey Vision Conference, pp. 147-151*, 1988.