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 salient features along a navigation path, using various cues like color, intensity and corners. 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 acquire the landmarks by themselves [4, 5]. That is the robot explores its navigation environment and automatically acquires a set of features that can be considered as robust but also distinctive 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. 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 most appropriate to the considered environment for the landmark selection process. Since adaptive behavior is one of the strengths 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 [14, 15], which aim to automatically select the most salient and thus the most relevant information of complex scenes, could be useful in this context [16]. Note that visual attention has been used to solve numerous other problems related to computer vision, like image segmentation, object tracking in dynamic scenes [17] and object recognition [18]. The usefulness of attention in real world applications is further strengthened by the recent realization of a real time visual attention system [17].

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 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 compute visual features that are compared with the learned landmarks in order to compute a probabilistic measure of the robot location within the navigation path. Further, this localization measure is integrated into a more general contextual localization framework based on a Markov model in order to take into account structural constraints of the environment.

The remainder of the paper is organized as follows. Section 2 describes the landmark selection procedure that is based on the visual attention algorithms. In Section 3, 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 contextual localization approach are described in Section 4. Section 5 reports some experimental results that show the potential of our method. Finally, conclusions and future works are stated in Section 6.

2. 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 viewpoint. 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.

2.1. Feature detection using saliency measure

In order to detect robust features, we use an extended version of the saliency-based model of visual attention [19]. The saliency-based model of attention has been firstly reported in [20] and gave rise to numerous soft and hardware implementations. For more details on the saliency-based attention model, the reader is referred to [21, 17].

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

1. First, a number $J$ of visual cues are extracted from the scene by computing the cue maps $F_j$. The cues used in this work are: 1) image intensity, 2) two opponent colors red/green ($RG$) and blue/yellow ($BY$), and 3) a corner-based cue computed according to the Harris approach [22], which leads to $J = 4$.

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) for robot navigation [10]. Unlike our approach, SIFT is limited to grey-scale images.

3. In the third stage of the attention model, the conspicuity maps are integrated together, in a competitive way, to form a saliency map $S$ in accordance with equation 1.

$$S = \sum_{j=1}^{J} N(C_j)$$  (1)
where $\mathcal{N}(\cdot)$ 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 [21]. 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. The total number of features can be either set interactively or automatically determined by the activity of the saliency map. For simplicity, the number of features is set to eight in our implementation.

2.2. Feature characterization and landmark selection

Once selected, each feature $S_i$ is characterized by its spatial position in the image $x_i = (x_i, y_i)$ and a visual descriptor vector $f_i$:

$$f_i = \begin{pmatrix} f_i^1 \\ \vdots \\ f_i^J \end{pmatrix}$$

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

$$f_i^j = \frac{\mathcal{N}(C_j(x_i))}{\mathcal{S}(x_i)}$$

Note that $\sum_{j=1}^J f_i^j = 1$.

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.

3. 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. In our experiments, a key frame contains about a dozen of landmarks.

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

- the horizontal spatial order of the landmark $x_{\text{index}}L_m$,
- the mean vertical position $y_{L_{m}}$ of each landmark,
- the corresponding maximum deviation $\Delta y_{L_{m}}$,
- the mean descriptor vector $\bar{f}_{L_{m}}$ of each landmark, and
- the corresponding standard deviation $\Sigma_{L_m}$.

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

$$K_q = \{L_m | L_m \text{ appears in } E_q \} \text{ with } L_m = \left( x_{\text{index}}L_m, \bar{f}_{L_{m}}, \Delta y_{L_{m}}, \bar{y}_{L_{m}}, \Sigma_{L_m} \right)$$

4. SELF-LOCALIZATION

The localization procedure aims, here, to determine 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 location 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 location. In this work, we use a voting technique to compute this likelihood. Further, this likelihood is integrated into a more general contextual localization framework.

4.1. Landmark recognition

Our landmark recognition method is based on the spatial and visual similarity between features detected during navigation and landmarks acquired during learning. Further, the method uses the spatial relationships between features / landmarks as an additional constraint. Specifically, a set of three features $s = \{S_1, S_2, S_3\}$ ($S_i = (f_i, x_i)^T$) is compared with a set of three landmarks $l = \{L_1, L_2, L_3\}$. The feature set $s$ matches the landmark set $l$ if the single features $S_i$ visually and spatially (according to height $y$) match the single landmarks $L_i$ and if, additionally, the horizontal spatial order of the three features is the same as that of the three landmarks. Formally:

$$\text{match}(s, l) = \begin{cases} True; & \text{if} \\ \frac{\|f_i - \bar{f}_{L_{i}}\| + |y_i - \bar{y}_{L_{i}}|}{\|f_i\|_{\text{norm}} + |y_i|_{\text{norm}} < \sigma_i \forall i \in \{1, 2, 3\} \&} \\
\end{cases}$$

$$\text{OrderX}(s) = \text{OrderX}(l)$$

where $f_{\text{norm}}$ and $y_{\text{norm}}$ are two normalization factors, $\sigma_i$ is a combination of the height variation $\Delta y_{L_i}$ and the descriptor vector standard deviation $\Sigma_{L_i}$, and $\text{OrderX()}$ sorts a list of features or landmarks according to their x-coordinates.
4.2. Voting procedure

In order to determine which key frame is most likely to be the current location of the robot, the detected features vote for key frames which contain landmarks that match these features. Given the set of all features \( S_t = \{ S_1, ..., S_m \} \) detected at time \( t \) and the set of all key frame \( K^* = \{ K_1, ..., K_n \} \), the voting procedure is achieved as follows. For each key frame \( K_i \), each triplet \( s = \{ S_a, S_b, S_c \} \subset S_t \) of features is compared to each triplet of landmarks \( l = \{ L_a, L_b, L_c \} \subset K_i \). 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.

The number of votes associated to a key frame \( K_i \) measures how likely the computed features \( S_t \) stem from that key frame (location). This measurement is normalized so that it becomes a probability distribution over the space of the key frames and called, henceforth, visual observation likelihood and formalized as \( P(S_t | K_i) \).

4.3. Contextual localization

In order to take into account the topological structure of the environment, we integrate the attention-based visual observation likelihood \( P(S_t | K_i) \) computed above into a Markov localization framework [23, 24]. Though the contextual navigation is not the main contribution of this work, we want to show that our attention-based landmark recognition method can be integrated into a more general navigation framework. To do so, we construct a rather simple Markov model of the environment where the states of the model correspond to the key frames (locations) and the transitions between the states simulate the displacement of the robot from a key frame to another. During navigation, the robot carries a probabilistic measure of its location i.e. the location likelihood \( P(K_t) \). \( P(K_t) \) is updated whenever the robot undertakes a displacement \( d \) or gets a visual observation. Formally, let \( P(K_t = K_i) \) be the probability that the actual location of the robot is key frame \( K_i \), \( P(K_t | d, K_j) \) be the probability that the robot moves from key frame \( K_j \) to key frame \( K_i \) when the displacement \( d \) is undertaken, and \( P(S_t | K_i) \) be the visual observation likelihood at time \( t \). As mentioned above, \( P(K_t) \) is updated in two different cases:

- when the robot undertake a displacement \( d \),
  \[
  P(K_t = K_i) = \frac{1}{\beta_t} \sum_{K_j \in K^*} P(K_t | d, K_j) \cdot P(K_{t-1} = K_j)
  \]
  \[\text{(6)}\]

- when the robot gets a visual observation,
  \[
  P(K_t = K_i) = \frac{1}{\alpha_t} P(S_t | K_i) \cdot P(K_{t-1} = K_i)
  \]
  \[\text{(7)}\]

Note that \( \alpha_t \) and \( \beta_t \) are normalization factors used to keep \( P(K_t) \) a probability distribution.

Fig. 2. Localization results. At each frame \( t \), the location likelihood \( P(K_t = K_i) \) is computed. In (a) the reference sequence is used as the navigation sequence. In (b) the test sequence is used as the navigation sequence (see text).
5. EXPERIMENTS

This section reports some experiments that aim at evaluating the presented localization method. The experiments consist first in learning visual landmarks from a reference sequence (Ref) of 120 color images acquired by the robot while navigating along a certain path of about 10 meters in a lab environment. These landmarks are organized into 8 key frames. Then, a test sequence (test) of about 80 frames is acquired while the robot follows a similar path. Since the robot starts almost from the same position for both sequences (reference and test), there exists an approximate timing between the two sequences. This timing is useful for the evaluation the localization results (a kind of ground truth). Regarding the Markov model, the number of states is 8 which corresponds to the number of key frames. The transition between the states is modelled by a gaussian distribution e.i. transitions between two neighboring key frames is more likely than transitions between distant key frames. Finally, the initial location likelihood $P(K_{t=0})$ is set to 80% at the real starting position, the other 20% are uniformly distributed over the other locations. During navigation, the location likelihood $P(K_t)$ is computed at each frame. As it is shown below, both sequences (Ref and Test) have been used as navigation sequence, e.i. the sequence used to compute the visual observation likelihood. Two types of results are presented in this section: qualitative results and quantitative ones.

The qualitative results are illustrated in Figure 2. The figure shows the evolution of the location likelihood $P(K_t)$ over time while the robot moves forward. (a) represents this evolution when the navigation sequence is the reference sequence itself (Ref-Ref), while in (b) the test sequence is used as the navigation sequence (Test-Ref). It can be seen in both cases that $P(K_t)$ is quasi mono-modal at each frame, with a quasi diagonal distribution of its highest value over time and location space.

The quantitative results are summarized in Table 1. In order to evaluate the performance of our localization method quantitatively, we introduce a metric called the approximative success rate (ASR). ASR is defined as the percentage of approximative correct localization (ACL). Note that a localization is considered as approximatively correct (ACL) if the maximum of the location likelihood $P(K_t)$ appears at key frames $K_c \pm 1$, where $K_c$ corresponds to the real location (according to ground truth data) of the robot.

The metric is computed for both situations as above: (Ref-Ref) and (Test-Ref). In both cases, it can be seen that the localization quality is quite high. More specifically, a localization score of over 99% when using the reference sequence as navigation sequence (Ref-Ref) stresses the distinctiveness or uniqueness of the visual landmarks automatically selected by the visual attention model. Further, the localization score (97.3%) when the test sequence (which is not acquired in exactly the same conditions as the learning sequence) is used as navigation sequence (Test-Ref) speaks for the robustness of the attention-based visual landmarks.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>ASR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref-Ref</td>
<td>99.2</td>
</tr>
<tr>
<td>Test-Ref</td>
<td>97.3</td>
</tr>
</tbody>
</table>

Table 1. Localization Results. As navigation sequences we used both: reference (Ref-Ref) and test (Test-Ref).

To summarize, the experimental results clearly show that the visual attention-based landmark selection and recognition method can be integrated into a more general contextual navigation framework. Further, the quantitative results speak for the uniqueness and robustness of the selected landmarks. Given the intrinsic capability of visual attention to adapt to the environment, the proposed attention-based localization method has high potential in developing navigation systems that have to operate in different environments.

6. CONCLUSIONS

This paper presents a robot localization method that combines visual attention-based landmark selection and recognition and contextual modeling of navigation environments. Using a saliency-based model of visual attention, the method automatically acquires the most salient and thus the most unique visual landmarks of the navigation environment. These landmarks are then organized into a topological map. During navigation, the method automatically detects the most conspicuous visual features and compares them to the learned landmarks. The result of the matching is then integrated into a more general contextual localization framework implemented with a Markov model. The fusion of the visual observation and the contextual constraints yields a probabilistic measure of the robot location. The Experimental results clearly demonstrate the effectiveness of the proposed method in a lab environment. Given the intrinsic capability of visual attention to adapt to the environment, our approach is expected as a generic method suited for all environments and capable of adaptation.

Acknowledgment

This work was partially supported by the CSEM-IMT Common Research Program. The authors highly appreciate the discussions with Adriana Tapus (Autonomous Systems Lab, EPFL) on some aspects of this work.
7. REFERENCES


