Published in Proceedings. 18th International Conference on Pattern Recognition (ICPR) 1, 695 - 698, 2006
which should be used for any reference to this work

Robot Navigation by Panoramic Vision and Attention Guided Features

Alexandre Bur[1], Adriana Tapus[2], Nabil Ouerhani[1], Roland Siegwart[2] and Heinz Hügli[1]
[1]Institute of Microtechnology (IMT), University of Neuchâtel, Neuchâtel, Switzerland
[2]Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland
{alexandre.bur,nabil.ouerhani,heinz.hugli}@unine.ch, adriana.tapus@ieee.org

Abstract

In visual-based robot navigation, panoramic vision emerges as a very attractive candidate for solving the localization task. Unfortunately, current systems rely on specific feature selection processes that do not cover the requirements of general purpose robots. In order to fulfill new requirements of robot versatility and robustness to environmental changes, we propose in this paper to perform the feature selection of a panoramic vision system by means of the saliency-based model of visual attention, a model known for its universality. The first part of the paper describes a localization system combining panoramic vision and visual attention. The second part presents a series of indoor localization experiments using panoramic vision and attention guided feature detection. The results show the feasibility of the approach and illustrate some of its capabilities.

1. Introduction

Vision is an interesting and attractive choice of sensory input, in the context of robot navigation. Specifically, panoramic vision is becoming very popular because it provides a wide field of view in a single image and the visual information obtained is independent of the robot orientation. Many robot navigation methods based on panoramic vision have been developed in literature. For instance, a model in [9] was designed to perform topological navigation and visual path-following. The method has been tested on a real robot equipped with an omnidirectional camera. Another model for robot navigation using panoramic vision is described in [1]. Vertex and line features are extracted from the omnidirectional image and tracked so that to determine the robot's position and orientation. In [8], the authors present an appearance-based system for topological localization. An omnidirectional camera was used. The resulting images were classified in real-time based on nearest-neighbor learning, image histogram matching and a simple voting scheme. Tapus et al. [7] have conceived a multi-modal, feature-based representation of the environment called a fingerprint of a place for localization and mapping. The multi-modal system is composed of an omnidirectional vision system and a 360 degrees laser rangefinder.

In these systems, the feature selection process is usually quite specific. In order to fulfill new requirements of versatility and robustness imposed to general purpose robot operating in wide varying environments, adaptive multi modal feature detection is required. Inspired from human vision, the saliency-based model of visual attention [3] is able to automatically select the most salient features in different environments. In [5], the authors presented a feature-based robot localization method relying on visual attention applied on conventional images and also showed its robustness. Applying the saliency-based model for feature detection provides automatic adaptation to different environments, like indoor and outdoor environments (Figure 1).

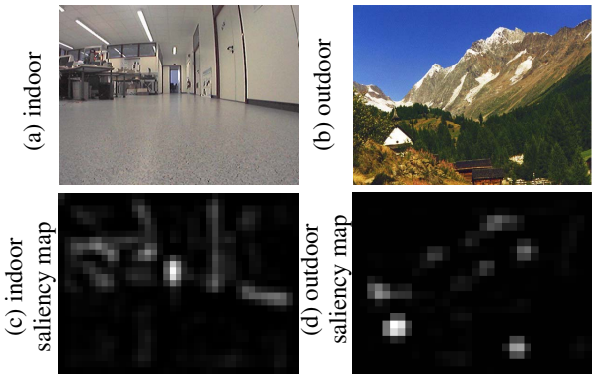

Figure 1. Adaptive behavior of the visual attention model for different environments

The purpose of this work is to get benefit of two main aspects: a) the omnidirectional vision for its independence of robot orientation and b) the visual attention-based feature extraction for its ability to cope with a wide varying environment.

The rest of the paper is structured as follows. Section 2 shows how visual attention applies to panoramic vision and how orientation independent robot localization is performed. Section 3 presents robot localization experiments and section 4 provides conclusions.

2. Visual Attention-based Navigation Using Panoramic Vision

2.1. Saliency-based Model of Visual Attention

The saliency-based model of visual attention, used for selecting the features of a scene, is composed of four main steps [3, 4], described as follows:
1) A number of cues are extracted from the scene by computing the so called feature maps F_j.
2) Each feature 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 feature, from their surrounding.
3) The conspicuity maps are integrated together, in a competitive way, into a *saliency map* $\mathcal{S}$ in accordance with:

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

where $\mathcal{N}()$ is the weighting operator for map promotion [3].
4) The features are derived from the peaks of the saliency map (Figure 1 c and d).

2.2. Visual Feature Detection in Panoramic Images

The saliency computation must be tuned to the specificities of panoramic images. As the features should also be detected in the full range of 360°, saliency computation algorithm must be adapted to the circularity of the input image. The circularity of the panoramic images allows to define the neighborhood on the borders, so that features on the image borders are also detected. Thus, the feature detection is obtained in the full panoramic range (Figure 2 b and c). In this paper, the saliency map is based on four different cues: image intensity, two opponent color components red/green and yellow/blue, and a corner-based cue (Harris method [2]).

Feature Characterization and Landmark Selection

Once detected, each feature O_n is characterized by its spatial position in the image $\mathbf{x}_{O_n} = (x_{O_n}, y_{O_n})$ and a visual descriptor vector $\mathbf{f}_{O_n}$, in which each component f_j holds the value of a cue at that location:

$$\mathbf{f}_{O_n} = (f_1, ..., f_j, ..., f_J)^T \quad with \quad f_j = F_j(\mathbf{x}_{O_n}) \tag{2}$$

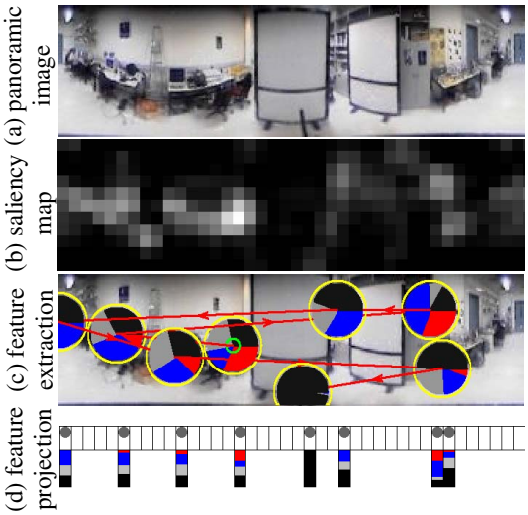

Figure 2. From the panoramic image to the horizontal feature projection

In order to take into account the spatial information of the features, an appropriate spatial representation is used: each set of features is represented on an horizontal one-dimensional space, by projection (Figure 2d).

Finally, an observation catched by a panoramic image is described by the set of features S_t (Figure 2c):

$$S_t = \{O_n\} \;\; with \;\; O_n = (\mathbf{x}_{O_n}, n_{x_{O_n}}, \mathbf{f}_{O_n}) \tag{3}$$

where $n_{x_{O_n}}$ is the index corresponding to the rank of the features spatially ordered in the x direction.

2.3. Map Building

The features detected during a learning phase are used as landmarks for localization during the navigation phase. In this work, a topological approach is used. A simple segmentation divides the path into equidistant portions E_q, each described by a configuration of landmarks named key-frame K_q.

Intrinsically, saliency provides a powerful adaptation to the robot environment. To provide a further adaptation, detected features are then chosen accordingly to their robustness. The step consists in tracking features along the environment [6] and to select as landmarks, the most persistent features, i.e. the ones with the longest tracking paths. A landmark is thus the representation of a robust feature that is persistent along the same portion E_q.

A key-frame K_q is a set of robust features named landmarks L_m, where each landmark is defined by the mean characteristics of the considered feature along the same

portion: its mean spatial position in the image $\overline{\mathbf{x}}_{L_m} = (\overline{x}_{L_m}, \overline{y}_{L_m})$, its index $n_{x_{L_m}}$, its mean descriptor vector $\overline{\mathbf{f}}_{L_m}$ and its standard deviation vector $\mathbf{f}_{\sigma_{L_m}}$:

$$K_q = \{L_m\} \;\; with \;\; L_m = (\overline{\mathbf{x}}_{L_m}, n_{x_{L_m}}, \overline{\mathbf{f}}_{L_m}, \mathbf{f}_{\sigma_{L_m}}) \quad (4)$$

2.4. Navigation Phase

As soon as the navigation map is available, the robot localizes itself by determining which key-frame K_q matches the best the robot's observation S_t at its current location.

2.4.1. Localization by Key-frame. The purpose is to match a set S_t of visual features with a set K_q of landmarks by measuring the visual and spatial similarity.

The visual landmark similarity: A landmark L_m and a feature O_n are said similar in terms of visual characterization if their Mahalanobis distance is inferior to a given threshold α:

$$\Delta\mathbf{f} = (\frac{f_{1_{L_m}} - f_{1_{O_n}}}{f_{1_{\sigma_{L_m}}}}, ..., \frac{f_{J_{L_m}} - f_{J_{O_n}}}{f_{J_{\sigma_{L_m}}}})^T \; and \; \|\Delta\mathbf{f}\| < \alpha \quad (5)$$

where $f_{J_{L_m}}$, $f_{J_{O_n}}$ and $f_{J_{\sigma_{L_m}}}$ are the *J* components of respectively $\mathbf{f}_{L_m}$, $\mathbf{f}_{O_n}$ and $\mathbf{f}_{\sigma_{L_m}}$.

The spatial similarity of landmark triplet: In this work, a comparison "feature group to landmark group" is used and the spatial similarity is measured by comparing the relative distances between each element of the group. Such a group matching strategy has the advantage to take into account the spatial relationships of each element of the group, which improves the matching quality. In this work, the groups contain three elements (triplet).

Formally, let $o = \{O_1, O_2, O_3\}$ be a set of three features compared with a set of three landmarks $l = \{L_1, L_2, L_3\}$. A triplet *o* is spatially similar to a triplet *l* if:

- the pairings (O_1;L_1), (O_2;L_2), (O_3;L_3) satisfy Eq.5.
- both sets are ordered with respect to their index $n_{x_{L_m}}$, $n_{x_{O_n}}$ under the principle of circularity.
- the absolute difference distances δ_{12} and δ_{23} are inferior to a threshold T_d:

$$\delta_{12}, \delta_{23} < T_d \quad (6)$$

$$where \;\; \delta_{12} = \mid (x_{O_2} - x_{O_1}) - (\overline{x}_{L_2} - \overline{x}_{L_1}) \mid \quad (7)$$

$$and \quad \delta_{23} = \mid (x_{O_3} - x_{O_2}) - (\overline{x}_{L_3} - \overline{x}_{L_2}) \mid \quad (8)$$

Given two spatial similar triplets, a function s_{c_i} not further defined here quantifies the overall similarity:

$$s_{c_i}(\Delta\mathbf{f}_1, \Delta\mathbf{f}_2, \Delta\mathbf{f}_3, \delta_{12}, \delta_{23}) \quad (9)$$

where $\Delta\mathbf{f}_k$ holds for the visual similarity of the pairing (O_k, L_k) and δ_{12}, δ_{23} for the spatial similarity.

Observation likelihood: Let n_{K_q} be the number of observation triplets that satisfy the landmark triplet similarity for the key-frame K_q. In order to define which key-frame K_q matches the best the observation, $S_{C_{(K_q)}}$ is computed as the sum of the similarity contribution of the n_{K_q} triplets:

$$S_{C_{(K_q)}} = \sum_{i}^{n_{K_q}} s_{c_i} \quad (10)$$

Thus, each key-frame receives several contributions, depending on the observation triplets that match the landmarks triplets. The measurement is then normalized in order to represent a probability distribution, called visual observation likelihood and formalized as $P(S_t|K_q)$:

$$P(S_t|K_q) = \frac{S_{C_{(K_q)}}}{\sum_n S_{C_{(K_n)}}} \quad (11)$$

$P(S_t|K_q)$ quantifies the likelihood of the observation S_t given the associated key-frame K_q. Simple localization is performed according to the maximum likelihood criterion:

$$q^* = arg \; max_q P(S_t|K_q) \quad (12)$$

2.4.2. Contextual Localization. To improve the robustness of the localization, the contextual information of the environment is taken into account. Thus, the visual observation likelihood $P(S_t|K_q)$ is integrated into a Markov localization framework. In this work, the states of the Markov model correspond to the portions E_q represented by its key-frame K_q and the state transition model is defined by $P(K_i, K_j)$, corresponding to the probability of the state transition from E_j to E_i.

Let $P_t(K_q)$ be the probabilistic estimation of its location at time t. $P_t(K_q)$ is computed in Eq.13 by fusing the prediction $P_{pred_t}(K_q = K_i)$ with the visual observation likelihood $P(S_t|K_i)$:

$$P_t(K_q = K_i) = \frac{1}{\alpha_t} P(S_t|K_i) \cdot P_{pred_t}(K_q = K_i) \quad (13)$$

$$P_{pred_t}(K_q = K_i) = \frac{1}{\beta_t} \sum_{K_j \in K^\star} P(K_i, K_j) \cdot P_{t-1}(K_q = K_j) \quad (14)$$

Note that α_t and β_t are normalization factors used to keep $P(K_t)$ a probability distribution.

3. Experiments

In the experiments, the robot acquires a sequence of panoramic images obtained from an equiangular omnidirectional camera, while moving along a path in a lab environment (Figure 2). The path of about 10 meters long gives rise to a sequence of 64 panoramic images. From this sequence, the navigation map is built in three different configurations:

(A) the map segmenting the path in 8 equidistant portions, (B) in 10 portions and (C) in 13 portions.

To quantify the localization, an approximate success rate *R* is defined. *R* corresponds to the percentage of approximate correct localization, which is considered as correct if the location with the maximum likelihood q^* (Eq.12) corresponds to $q_e \pm 1$, where q_e represents the exact location.

During the localization experiment, the observation S_t of each frame of the navigation sequence is computed and compared with the key-frames of the map. S_t contains the 8 most salient features of the current frame and the matching refers to all possible triplets of features and landmarks.

The value $R_{\overline{c}}$ measures the success rate of the simple context-free localization. The value R_c holds for the contextual localization with the Markov framework, where the initial estimation $P(K_{t=0})$ is set to 80% at the exact location and the other are uniformly distributed at the other locations. The state transitions $P(K_i, K_j)$ are modelled by a Gaussian distribution, i.e. transition to the neighboring portions is more likely than transition to distant portions.

The first experiment (Exp.1) tends to evaluate the quality of the visual landmarks. It uses the same sequence for map building and navigation. The second experiment (Exp.2) verifies the orientation independence of the proposed process. It uses three test sequences corresponding to rotated views of the original sequence by $90°$, $180°$ and $270°$ respectively to be matched with the original map.

Exp.1	8 KF (A)	10 KF (B)	13 KF (C)	mean
$R_{\overline{c}}[\%]$	87.5	82.8	79.7	83.3
$R_c[\%]$	98.4	96.9	96.9	97.4
Exp.2	8 KF (A)	10 KF (B)	13 KF (C)	mean
$R_{\overline{c}}[\%]$	80.2	80.2	78.1	79.5
$R_c[\%]$	94.8	98.4	98.9	97.4

Table 1. Localization Results

The results are presented in Table 1. For simple key-frame localization, the success rate $R_{\overline{c}}$ decreases as expected when the number of portions increases and experiment 1 provides an average rate of 83%. Contextual localization improves the performance further and provides an average rate R_c of 97%. Given the fact that the sequence of panoramic images provides only small changes, with key-frames representing small portions of about one meter length, the performance is considered as quite good. In Exp.2, the results are similar to Exp.1 and show the orientation independence of the localization method.

These results confirm the feasibility of the proposed approach and show the capacity of the system to catch robust discriminant features. The next step will be to evaluate the robustness of the method in presence of condition changes (luminosity, different robot navigation trajectories).

4. Conclusions

An original robot localization system was presented, that encompasses panoramic vision and attention guided feature detection. First, the multi-cue saliency-based model of visual attention was adapted to panoramic image sequences; a description for a feature set, as well as a suited feature set matching method were also proposed. Then, localization experiments were conducted using two simple methods. In a sequence of panoramic images showing only small changes, the rate of successful localization is typically 83% and 97% with the context-free and contextual methods respectively. Another experiment shows the orientation independence of the proposed processing. These results confirm the feasibility of the proposed approach and show the capacity of the system to catch robust discriminant features.

5. Acknowledgment

This work is partially supported by the Swiss National Science Foundation under grant SNSF-108060.

References

[1] M. Fiala and A. Basu. Robot navigation using panoramic landmark tracking. *Society of Manufacturing Engineers (SME). Article RPOS-100. USA, NRC 47136*, 2003.

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

[3] 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 (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.

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

[5] N. Ouerhani, A. Bur, and H. Hugli. Visual attention-based robot self-localization. *ECMR 2005, in Proc. of European Conference on Mobile Robotics, Italy, pp. 8-13*, 2005.

[6] N. Ouerhani and H. Hugli. A visual attention-based approach for automatic landmark selection and recognition. *WAPCV 04, in Lecture Notes in Computer Science, Springer Verlag, LNCS 3368, pp. 183-195*, 2005.

[7] A. Tapus and R. Siegwart. Incremental robot mapping with fingerprints of places. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Canada*, 2005.

[8] I. Ulrich and I. R. Nourbakhsh. Appearance-based place recognition for topological localization. *IEEE International Conference on Robotics and Automation (ICRA), USA*, 2000.

[9] N. Winters, J. Gaspar, G. Lacey, and J. Santos-Victor. Omnidirectional vision for robot navigation. *Proceedings of IEEE Workshop on Omnidirectional Vision (Omnivis00)*, 2000.