Visual Robot Localization and Mapping based on Attentional Landmarks

Simone Frintrop

Comp. Science III, University of Bonn, Germany, frintrop@iai.uni-bonn.de

Abstract. In this paper, we present a system for simultaneous localization and map building of a mobile robot, based on an attentional landmark detector. A biologically motivated attention system finds regions of interest which serve as visual landmarks for the robot. The regions are tracked and matched over consecutive frames to build stable landmarks and to estimate the 3D position of the landmarks in the environment. Furthermore, matching of current landmarks to database entries enables loop closing and global localization. Additionally, the system is equipped with an active camera control, which supports the system with a tracking, a re-detection, and an exploration behaviour.

1 Introduction

One of the most important tasks of a mobile robot is to localize itself within its environment. This task is especially difficult if the environment is not known in advance. Within the robotics community, this problem is well known as *SLAM (Simultaneous Localization and Mapping)*. Currently, there has been special interest in *visual SLAM*, which uses cameras as main sensors since cameras are low-cost, low-power and lightweight sensors [1, 6, 7].

A key competence in visual SLAM is to choose useful visual landmarks which are easy to track, stable over several frames, and easily re-detectable when returning to a previously visited location. Here, we present a visual SLAM system based on an attentional landmark detector: the attention system VOCUS [2] detects regions of interest (ROIs) which are tracked and matched over consecutive frames. To improve the stability of the features, the ROIs are combined with Harris corners. When re-visiting a location after some time, knowledge about the appearance of expected landmarks is used to search in a top-down manner for expected features. Additionally, active camera control improves the quality and distribution of detected landmarks.

The novelty of the presented system in comparison to other approaches of visual SLAM – e.g., [1, 6, 7] – lies first, in the attentional feature detection in combination with Harris corners [4], second, in the top-down, target-directed feature computations to improve loop closing [3], and third, in the active camera control [5]. Here, we combine the results of these previous findings.

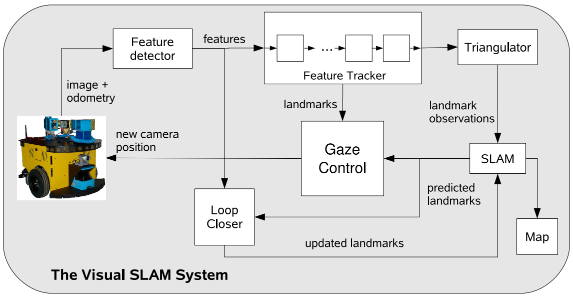

Fig. 1. The visual SLAM system builds a map based on image data and odometry

2 The Visual SLAM System

The visual SLAM architecture (Fig. 1) consists of a *robot* which provides camera images and odometry information, a *feature detector* to find ROIs in the images, a *feature tracker* to track ROIs over several frames and build landmarks, a *triangulator* to identify useful landmarks, a *SLAM module* to build a map of the environment, a *loop closer* to match current ROIs to the database, and a *gaze control module* to determine where to direct the camera to.

When a new frame from the camera is available, it is provided to the *feature detector*. This module finds ROIs based on the visual attention system VOCUS [2]. VOCUS computes a bottom-up saliency map, based on strong contrasts and uniqueness of the features intensity, orientation, and color. For each ROI, a feature vector is stored which is used for matching and top-down search. Since the shape of attentional ROIs differs sometimes in consecutive frames, the ROIs are combined with Harris corners to improve position stability [4]. A bottom-up saliency map and the corresponding ROIs are displayed in Fig. 2.

Next, the features are provided to the *feature tracker* which stores the last n frames, performs matching of ROIs and Harris corners in these frames and creates landmarks which are lists of features found in several frames. Matching of ROIs and Harris corners is based on proximity and similarity of the feature vector (ROIs) or a SIFT descriptor (Harris) [4]. The purpose of the buffer is to identify features which are stable over several frames and have enough parallax information for 3D initialization. These computations are performed by the *triangulator*. Selected landmarks are stored in a database and provided to the *SLAM module* which computes an estimate of the position of landmarks and integrates the position estimate into the *map* (details to SLAM module in [6]).

The task of the *loop closer* is to detect if a scene has been seen before. The SLAM module provides the loop closer with expected landmark positions and their feature descriptions. The attentional feature vector is used to search in a top-down manner for the expected landmarks. The result is a top-down saliency map which highlights regions which correspond to the target (cf. Fig. 2). The corresponding top-down ROIs are compared with the ROIs of the expected landmarks by comparing the similarity of their feature vectors. If two ROIs

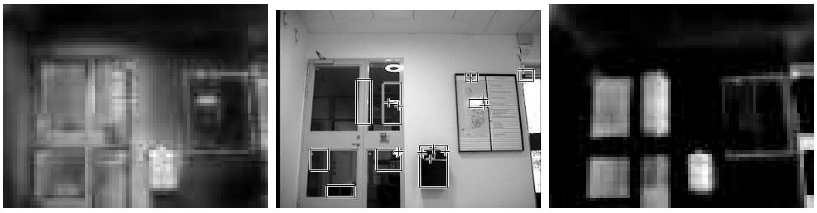

Fig. 2. Left: bottom-up saliency map. Middle: attentional ROIs (rectangles) and Harris corners (crosses). Right: top-down saliency map for target "wastebin" (black box).

match, this information is provided to the SLAM module to update the positions of the robot and the landmarks.

Finally, the *gaze control module* controls the camera actively with three behaviours: a *tracking* behaviour identifies the most promising landmarks and prevents them from moving out of the field of view. A *redetection* behaviour actively searches for expected landmarks to support loop-closing. Finally, an *exploration* behaviour investigates regions with no landmarks, leading to a more uniform distribution of landmarks. The process to decide which behaviour is activated is based on the amount of uncertainty about the current position and on the number of currently visible landmarks (details in [5]).

3 Experiments and Results

To illustrate the advantages of the presented visual SLAM system, we performed two experiments which show i) the advantages of the top-down attentional matching approach in loop closing situations, and ii) the advantages of active over passive camera control. In both experiments, the robot drove through a room in an office environment, through a corridor, and entered the room again. Here, it should detect that it closed a loop.

In the 1st experiment, we compared bottom-up matching of ROIs (VOCUS computes a bottom-up saliency map and the similarity of ROIs is compared based on a threshold) and top-down matching (VOCUS searched for the expected landmarks in the current frame and the resulting ROIs are compared afterwards) (Fig. 3, left). If only very few false matches are accepted, the bottom-up matching is better. But if more false matches are acceptable, we get a significantly higher number of correct matches (42% more). Note that this number of false matches is not the number of false matches reported to SLAM since several of the matched ROIs belong to the same landmark and we also use matching of Harris corners afterwards to reduce the number strongly (details in [4]). In the current example, only one false landmark match remained.

In the 2nd experiment, we compared passive with active camera control. The resulting maps are displayed in Fig. 3, middle/right. With active control, we achieve a better distribution of landmarks and more matches, e.g. loop closing takes places earlier and more reliably (details in [5]).

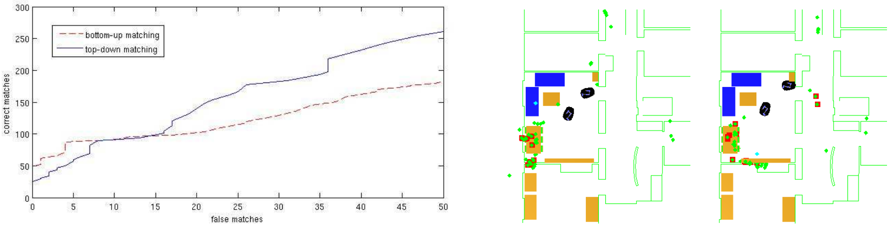

Fig. 3. Left: Experiment 1: Correct matches for bottom-up and top-down matching depending on the error rate: For a low number of false detections, bottom-up matching results in more correct matches. If more false matches are acceptable, top-down matching provides more correct matches. **Middle/Right: Experiment 2:** Two maps consisting of visual landmarks (green/cyan dots), created with passive (middle) and with active (right) camera control. Two robots in one image correspond to the robot at the beginning and at the end of the buffer, i.e., the robot further ahead on the path is the real robot, the one behind is the virtual robot position 30 frames later. Landmarks matched to database entries are displayed as large, red dots. Active control enables a better distribution of landmarks and more matches.

4 Conclusion

We have presented a visual SLAM system based on an attentional landmark detector. The attentional regions are especially useful landmarks for tracking and redetection; the loop closing is improved by using top-down guidance. Active camera control helps to achieve better, more stable landmarks, a better distribution of landmarks, and a faster and more reliable loop closing.

In future work, we plan to combine the method with other visual loop-closing techniques, for example by considering not only one expected landmark for matching, but all in the current field of view.

References

1. A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. of the ICCV*, oct 2003.
2. S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. PhD thesis, Bonn, Germany. Published 2006 in LNAI, Springer.
3. S. Frintrop and A. B. Cremers. Top-down attention supports visual loop closing. In *Proc. of ECMR (to appear)*, 2007.
4. S. Frintrop, P. Jensfelt, and H. Christensen. Pay attention when selecting features. In *Proc. of the 18th Int'l Conf. on Pattern Recognition (ICPR 2006)*, 2006.
5. S. Frintrop, P. Jensfelt, and H. Christensen. Attentional robot localization and mapping. In *ICVS Workshop WCAA*, 2007.
6. P. Jensfelt, D. Kragic, J. Folkesson, and M. Björkman. A framework for vision based bearing only 3D SLAM. In *Proc. of ICRA'06*, Orlando, FL, May 2006.
7. P. Newman and K. Ho. SLAM- loop closing with visually salient features. In *Proc. of the International Conference on Robotics and Automation, (ICRA 2005)*, 2005.