Pay Attention When Selecting Features

Simone Frintrop, Patric Jensfelt and Henrik I. Christensen
CVAP, CSC, Kungliga Tekniska Högskolan (KTH), Stockholm, Sweden
Email: {frintrop/patric/hic}@nada.kth.se

Abstract

In this paper, we propose a new, hierarchical approach to landmark selection for simultaneous robot localization and mapping based on visual sensors: a biologically motivated attention system finds salient regions of interest (ROIs) in images, and within these regions, Harris corners are detected. This combines the advantages of the ROIs (reducing complexity, enabling good redetactability of regions) with the advantages of the Harris corners (high stability). Reducing complexity is important to meet real-time requirements and stability of features is essential to compute the depth of landmarks from structure from motion with a small baseline. We show that the number of landmarks is highly reduced compared to all Harris corners while maintaining the stability of features for the mapping task.

1. Introduction

One of the currently most investigated topics in the field of robotics is *simultaneous localization and mapping (SLAM)*, in which a robot builds a map of the environment using sensor data [13, 1]. While SLAM is considered to be solved for small 2D scenarios when using range sensors, e.g., laser scanners, much current interest focuses now on *Visual SLAM* which uses cameras as external sensors [2, 5, 10]. Of special interest are systems using a single camera, because of the low costs of such systems. Here, one problem is that it is hard to estimate the depth of a region from a single frame. Several frames from multiple viewpoints are required to estimate the depth, similar to the *structure-from-motion* problem (SFM).

Since the depth of a landmark cannot be initialized from a single frame, it has to be tracked over several frames. If the camera is mounted on the robot which is moving along the optical axes, the baseline between consecutive frames is small, so the points for the depth estimation have to be very stable. Even an error of 2 or 3 pixels can result in a significant error in the depth estimate. Thus, the method relies on stable and reliable tracking of landmarks. Features which are usually used for tracking are corner points as Harris corners [8] or SIFT keypoints [7]. They are largely stable under image transformations and illumination changes. On the other hand, these detectors produce a huge amount of features, more than can be tracked and stored in a real-time robotic system for large environments [12]. To our knowledge, there is currently no strategy for selecting a subset of these features in an intelligent way.

Other approaches to determine regions of interest in images are computational visual attention systems [14, 6, 3]. They select regions that “pop out” in a scene due to strong contrasts and uniqueness, e.g., the famous black sheep in a white herd. The advantage of these systems is that they determine globally which regions in the image are worth investigating instead of locally detecting certain properties like corners. Some systems additionally integrate previous knowledge (top-down information) into the computations [9, 3]. This enables a better redetection of landmarks when presuming to revisit a known location.

In this paper, we present a new, hierarchical approach to landmark selection for visual SLAM: a combination of attentional ROIs and Harris corners. Using Harris corners directly results in a significant number of features which pose a challenge to real-time performance. Attention can select a smaller number of ROIs but they do not have the same positional stability. Through a combination of stable feature points within ROIs, a stable and more scalable representation is presented. For matching across frames, a SIFT descriptor is attached to each feature point. The integration into the SLAM framework is subject for future work; the contribution of this paper is a detailed investigation of the landmark selection and the introduction of a new selection strategy for landmarks.

In our experiments, we compare the number of features created by the different methods as well as the tracking accuracy. We show that the features from the combined approach meet our requirements best: the number of features is reduced to a feasible amount and the tracking accuracy is high. Since each Harris corner is related to a ROI with its own feature vector, this vector can later be used to actively redetect a region.

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

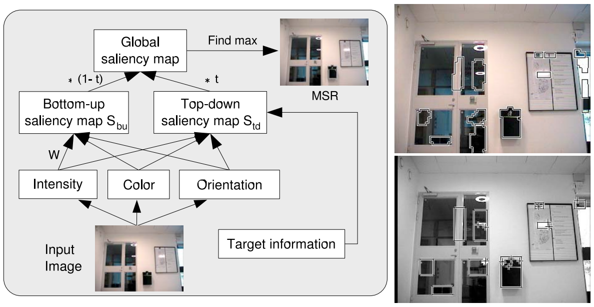

Figure 1. Left: The visual attention system VOCUS. Right: MSRs (top) and ROIs with Harris corners (bottom).

Using visual attention systems for robot localization was also done by [11]. As a topologically map was built and not a geometrical one, the stability of the regions is not as important. SLAM with features from a different saliency measure based on entropy was reported by [10]. However, here a laser-scanner was used additionally to a camera, so the approach is not dependent on extracting the depth of landmarks from image data.

2. The Visual Attention System VOCUS

The computational attention system VOCUS (Visual Object detection with a CompUtational attention System) [3] that we used here, differs from most other ones by the ability to consider target knowledge (top-down information) to enable goal-directed search. It consists of a bottom-up part and a top-down part; global saliency is determined from both cues (cf. Fig. 1).

The bottom-up part detects salient image regions by computing image contrasts and uniqueness of a feature, e.g., a red ball on green grass. The feature computations for the features intensity, orientation, and color are performed on 3 different scales with image pyramids. The feature intensity is computed by *center-surround mechanisms* [3]; on-off and off-on contrasts are computed separately. After summing up the scales, this yields 2 intensity maps. Similarly, 4 orientation maps ($0\,°, 45\,°, 90\,°, 135\,°$) are computed by Gabor filters and 4 color maps (green, blue, red, yellow) which highlight salient regions of a certain color (details in [3]). Each feature map X is weighted with the uniqueness weight $\mathcal{W}(X) = X/\sqrt{m}$, where m is the number of local maxima that exceed a threshold. This weighting is essential since it emphasizes important maps with few peaks, enabling the detection of *pop-outs*. After weighting, the maps are summed up first to 3 conspicuity maps I (intensity), O (orientation) and C (color) and finally, after again weighting for uniqueness, to the *bottom-up saliency map* $S_{bu} = \mathcal{W}(I) + \mathcal{W}(O) + \mathcal{W}(C)$ (cf. Fig. 2, top left).

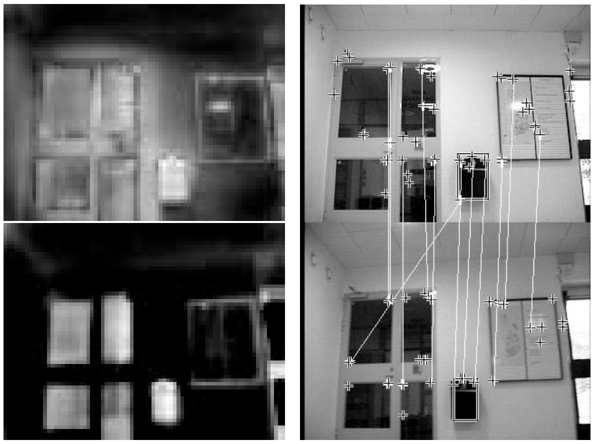

Figure 2. Loop closing example: top: scene at beginning of sequence. Bottom: revisited scene, 592 frames later, searching for the black waste bin with top-down attention. Left: saliency maps S_{bu} (top) and S_{td} (bottom). Right: the most salient ROI (top) and its redetection with top-down attention (bottom) as well as the matching of all Harris points. In our approach, we use only the points inside the ROIs what reduces the complexity strongly and eliminates false matches.

In top-down mode, VOCUS aims to redetect a target, i.e., input to the system is the image and some target information, provided as a feature vector $\vec{v}$ with 13 $(2+4+4+3)$ entries, one for each feature and conspicuity map. This vector is learned from a region which is provided manually or automatically; in this application it is determined automatically from a *most salient region (MSR)* of S_{bu} (see below). In *search mode*, VOCUS multiplies the feature and conspicuity maps with the weights of $\vec{v}$. The resulting maps are summed up, yielding the *top-down saliency map* S_{td} (cf. Fig. 2, bottom left). Finally, S_{bu} and S_{td} are combined by: $S = (1-t) * S_{bu} + t * S_{td}$, where t determines the contribution of bottom-up and top-down. Finally, the MSRs in S are determined by first finding local maxima in S (seeds) and second finding with *region growing* recursively all neighboring pixels over a saliency threshold (here: 25% of the seed) (Fig. 1, right top). For simplicity, we use the rectangle determined by height and width of the MSR as *region of interest (ROI)* (cf. Fig. 1, right bottom). Furthermore, we ignore MSRs on the borders of images, since these are not stable when the viewpoint changes.

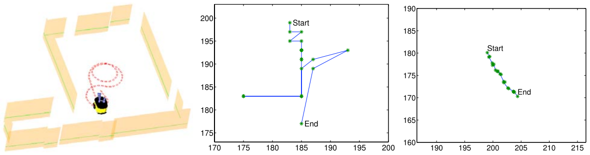

Figure 3. Left: the robot environment and the driven trajectory. Middle/right: movement of coordinates of two points (middle: center of attentional ROI. Right: Harris corner) tracked over a sequence of 13 images.

3. Stable Feature Detection and Tracking

The attentional ROIs are useful since they reduce complexity and enable active search for landmarks when returning to the same area again. However, we found that the position of the ROIs was roughly stable, but jumped from frame to frame sometimes by several pixels. Estimating the depth of a region by SFM, providing only a small baseline resulting from moving predominantly along the optical axis, requires precise measurements. Therefore, it is difficult to use the ROIs directly as landmarks in a geometric map constructed using a single camera.

We propose to use the ROIs to guide the search for more stable features. One type of features which is usually easy to localize exactly are corner features. We used the Harris-Laplace feature detector [8] – an extension of the Harris corner detector to Laplacian pyramids which enables scale invariance – and applied it inside the ROIs. This resulted in a few (average 1.6) points per ROI (cf. Fig. 1, right, bottom).

Fig. 3 compares the tracking accuracy of a ROI (middle) and a Harris corner (right). The robot moved smoothly during this sequence, so the image coordinates should lay on a smooth curve. This is the case for most coordinates of the ROI, but there are several strong outliers. The tracked Harris corner shows to be much more accurate.

To allow matching of points, a SIFT descriptor is computed for each detected corner [7]. In this approach, a 4×4 grid is placed on a point and a pixel gradient magnitude is calculated at $45°$ intervals for each of the grid cells. This yields an $4 \times 4 \times 8 = 128$ dimensional descriptor vector for each point.

To perform the tracking of ROIs and Harris corners between frames, we consider both proximity and similarity. For the Harris points we make use of the odometry to estimate how the camera moved between two frames and predict how the points should move assuming a certain depth (here 4m). We look for matching around the predicated position (proximity) and use the SIFT descriptor to find the best match (similarity). Although most points were only detected once (av. 1.5 frames), there were several points that could be tracked over a large number of frames, one was even tracked over 70 frames.

For VOCUS, we consider ROIs that are close (here: within 15 pixels) to the ROI of the previous frame (proximity) and have a local maximum in the saliency map (similarity). In current work [4], we have improved the matching by using odometry prediction and feature matching based on $\vec{v}$. There, we show that this approach gives good results in tracking as well as in loop closing situations.

Although the Harris corners outperform the ROIs in position accuracy, the ROIs showed to be tracked easier than Harris corners: On average, a ROI is tracked over 5.0 frames, the maximum was 110 frames. In the combined Harris/ROI approach, landmarks were tracked in average over 1.7 frames; the longest track contained 65 frames.

To sum up, the approach works as follows (cf. Fig. 1 and 2): VOCUS computes a bottom-up saliency map, the MSRs, and the ROIs. Then, Harris corners are determined inside the ROIs. From tracking the Harris corners and the ROIs over several frames, landmarks are computed and included into the environment map (for details cf. [4]). For each landmark, the attentional feature vector $\vec{v}$ is saved together with the SIFT descriptor of the Harris points. When closing a loop, i.e., returning to an already visited location, the estimated robot position provides a prediction of expected landmarks. The feature information $\vec{v}$ from the expected landmark is used to search actively for the landmark with top-down attention: a top-down saliency map is computed and the MSRs are determined. Harris corners are determined inside the corresponding ROI and finally the matching is performed between the expected landmark and the current ROI. Fig. 2 (right) shows the matching between all corners of two frames. From the 12 matches, 3 are false. Since we match the ROIs first and consider then only the Harris points inside the ROIs, no false matches occur and the matching complexity is highly reduced.

4. Experiments and Results

For our experiments, we used a sequence of 658 images, obtained in a hallway by a robot driving as shown in Fig. 3, left. The experiments consist of two parts: first, we investigated the complexity reduction of points and, second, we tested the stability of the different approaches.

The number of points/regions selected by the different approaches is shown here:

	Harris	ROI	Harris in ROI
# pts in all frames	21258	2871	4569
# landmarks	14036	570	2664

The first row shows that the number of points in all frames was 78% smaller in the Harris/ROI approach than when

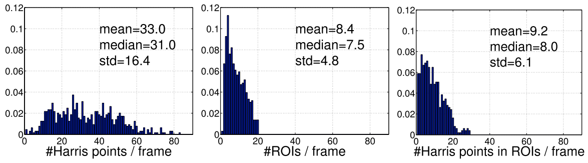

Figure 4. number of points per frame

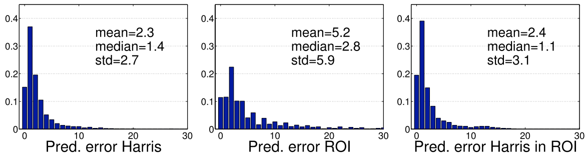

Figure 5. Prediction error distributions

computing all Harris corners. After tracking, the points were associated with 2664 landmarks, 81% less than in the Harris approach. The number of points per frame is depicted in Fig. 4. It clearly shows that in the new approach, significantly less points are selected per frame than when using only Harris corners.

To quantitatively evaluate the tracking of ROIs vs. Harris points, we computed the error between the measured points and the predicted position. To get the predicted position, we used the robot's odometry for motion prediction and laser scanner measurements for depth prediction. Note, that the laser values are only used for evaluation, the method itself relies only on camera data.

Fig. 5 shows histograms of the prediction errors for the for the Harris points, for the ROI center points, and for the Harris points in ROIs. It shows, that the prediction error for VOCUS is typically less then 3 pixels, but there are several outliers. The tracking of the Harris points gives much better results: most points have a prediction error of less than 2 pixels. The quality is about the same when the Harris points inside the ROIs are considered; the median is even slightly improved. Therefore, when restricting the Harris points to ROIs, we keep the quality but have the advantage of a reasonable sized subset of points for visual SLAM which are easily redetectable by the top-down part of VOCUS.

5. Conclusion

We have presented a new hierarchical landmark selection scheme for visual SLAM: an attention system provides regions of interest and inside these regions Harris corners are computed. We show that this approach reduces the number of detected points significantly, an important property for real-time performance, and that the points are highly stable, a condition for precise depth computations of landmarks. In future work, the landmarks will be integrated in our current SLAM architecture, to enable loop closing and active relocalization of landmarks utilizing the top-down saliency.

Acknowledgment

The present research has been sponsored by the European Commission through the project NEUROBOTICS (EC Contract FP6-IST-001917) and the Swedish Foundation for Strategic Research through its Center for Autonomous Systems. The support is gratefully acknowledged. Additional thanks to Maria Klodt for speeding up VOCUS to enable real-time performance.

References

[1] J. A. Castellanos and J. D. Tardós. *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach*. Kluwer Academic Publishers, 1999.

[2] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proc. of the ICCV*, oct 2003.

[3] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. PhD thesis, accepted Jul. 2005. in LNAI, Vol. 3899, Springer, 2006.

[4] S. Frintrop, P. Jensfelt, and H. Christensen. Attentional landmark selection for visual slam. 2006. submitted.

[5] L. Goncavles, E. di Bernardo, D. Benson, M. Svedman, J. Ostrovski, N. Karlsson, and P. Pirjanian. A visual front-end for simultaneous localization and mapping. In *Proc. of ICRA*, pages 44–49, apr 2005.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Trans. on PAMI*, 20(11):1254–1259, 1998.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of ICCV*, pages 1150–57, 1999.

[8] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages 525–531, 2001.

[9] V. Navalpakkam, J. Rebesco, and L. Itti. Modeling the influence of task on attention. *Vis. Res.*, 45(2):205–231, 2005.

[10] P. Newman and K. Ho. SLAM-loop closing with visually salient features. In *Proc. of ICRA'05*, pages 644–651, 2005.

[11] N. Ouerhani, A. Bur, and H. Hügli. Visual attention-based robot self-localization. In *Proc. of ECMR*, pages 8–13, 2005.

[12] R. Sim, P. Elinas, M. Griffin, and J. J. Little. Vision-based SLAM using the Rao-Blackwellised Particle Filter. In *Proc. IJCAI Workshop RUR*, 2005.

[13] S. Thrun, D. Fox, and W. Burgard. A probabilistic approach to concurrent mapping and localization for mobile robots. *Autonomous Robots*, 5:253–271, 1998.

[14] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.