

A Probabilistic Approach to Simultaneous Segmentation, Object Recognition, 3D Localization and Tracking using Stereo

Georg von Wichert

Siemens AG, Corporate Technology, Information and Communications, 81730 Munich, Germany

Abstract. Vision systems for service robotics applications have to cope with varying environmental conditions, partial occlusions, complex backgrounds and a large number of distractors (clutter) present in the scene. This paper presents a new approach targeted at such application scenarios that combines segmentation, object recognition, 3D localization and tracking in a seamlessly integrated fashion. The unifying framework is the probabilistic representation of various aspects of the scene. Experiments indicate that this approach is viable and gives very satisfactory results.

1 Introduction

Vision systems for service robotics applications have to cope with varying environmental conditions, partial occlusions, complex backgrounds and a large number of distractors (clutter) present in the scene. Systems mounted on mobile platforms additionally have to incorporate ego-motions and therefore have to solve the real-time tracking problem. Conventional vision systems generally perform the necessary segmentation and recognition as well as possibly 3D-localization and tracking in a pipelined, sequential way, with a few exceptions like [7] who integrate recognition and segmentation in a closed loop fashion.

However, one can imagine several ways in which different parts of the image processing pipeline could profit from each other. Service robots for example will in most cases be allowed observe their environment for short time periods to take advantage of the information present in image streams. Moving scenes will show the objects under observation from changing view points and this can be exploited to maintain and improve existing object recognition and localization hypotheses, provided that those are tracked over time. In a similar way, object recognition can profit from successful segmentation and segmentation in turn can benefit from depth information, if available. These possible synergies are currently not exploited by most systems.

This paper presents a new approach that combines segmentation, recognition, 3D-localization and tracking in a seamlessly integrated fashion. We aim at developing vision algorithms which will reliably work in real everyday environments. To reach this goal it is mandatory to take advantage of the synergies between the different stages of the conventional processing pipeline. Our approach is to eliminate the pipeline structure and simultaneously solve the segmentation, object recognition, 3D localization and

tracking problems. The unifying framework is the probabilistic representation of various aspects of the scene.

2 Method

Our method, as currently implemented and described in section 2.3, takes advantage of previous work in two major areas. First of all object recognition methods based on probabilistic models of the objects appearance have recently been presented by many research groups (see [9, 11] among many others) and have shown promising results with respect to robustness against varying viewpoints, lighting and partial occlusions. In addition these models can be trained from demonstrated examples. Thus, the model acquisition process does not necessarily require a skilled operator, which is a nice property for the application of such a system in the service robotics field.

The second major contribution onto which the current implementation of our approach is built is the condensation algorithm by Isaard and Blake [4]. It is used for the probabilistic representation and tracking of our object hypotheses (recognition and localization) over time. A short overview of probabilistic object recognition and the condensation algorithm is given in sections 2.1 and 2.2. As mentioned before, our goal is to integrate segmentation, object recognition, 3D localization and tracking in order to take advantage of synergies among them. The probabilistic approach to object recognition is the framework that enables us to do so.

2.1 Probabilistic object recognition

Probabilistic approaches have received significant attention in various domains of robotics reaching from environmental map building [6] to mobile robot localization [1, 13] and active view planning [10, 14, 8]. This is mainly due to the absolute necessity for explicit models of environmental uncertainty as a prerequisite for building successful robot systems.

For a probabilistic recognition of an object o from a image measurement m , we are interested in the conditional probability $p(o|m)$. This is called posterior probability for the object o given the measurement m . The optimal decision rule for deciding whether the object is present, is to decide based on which probability is larger $p(o|m)$ or $p(\bar{o}|m) = 1 - p(o|m)$ with \bar{o} referring to the absence of the object. This decision rule is optimal in the sense of minimizing the rate of classification errors [2].

It is practically not feasible to fully represent $p(o|m)$, but using the Bayes rule we can calculate it according to

$$p(o|m) = \frac{p(m|o)p(o)}{p(m)} \quad (1)$$

with

- $p(o)$ the a priori probability of the object o
- $p(m)$ the a priori probability of the measurement m

- $p(m|o)$ the conditional probability of the measurement m given the presence of the object o

These probabilities can be derived from measurement histograms computed from training data. In case of a simple object detection problem, the decision rule $p(o|m) > p(\bar{o}|m)$ can be rewritten as a likelihood test. We decide that the object is present if

$$\frac{p(m|o)}{p(m|\bar{o})} > \frac{p(\bar{o})}{p(o)} = \lambda \quad (2)$$

Here λ can be interpreted as a detection threshold, which in many cases will be set arbitrarily, since the a priori probabilities of the objects depend upon the environmental and application context and are difficult to obtain. Having k *independent* measurements m_k (e.g. from a region R belonging to the object) the decision rule becomes

$$\frac{\prod_k p(m_k|o)}{\prod_k p(m_k|\bar{o})} > \frac{p(\bar{o})}{p(o)} = \lambda \quad (3)$$

If the m_k are measurement results of local appearance characteristics like color or local edge energy (texture), the resulting recognition systems tend to be comparatively robust to changes in viewing conditions [9, 11, 8]. However, this is not the main subject of this article.

The appearance based object recognition approach solves only one part of our problem. Furthermore, it theoretically requires, that the scene is properly segmented since equation 3 assumes that all measurements m_k come from the same object. A proper segmentation of complex cluttered scenes is known to be a difficult task. Depth information would be extremely helpful. Object boundaries generally will be easier to detect in depth images, but these are computationally expensive when computed over full frames. The depth recovery could be done more efficiently, if we already had a position hypothesis \underline{x} for the object to be recognized at the current time step t . Section 2.3 will show how all these synergies can be exploited, without the need for extensive stereo correspondence search.

2.2 A Short Introduction to the Condensation Algorithm

The goal of using the condensation algorithm is to efficiently compute and track our current belief $p(\underline{x}, t)$ of the object position \underline{x} , i.e. the probability that the object is at \underline{x} at time t . While one could represent this belief distribution in a regular grid over the 3D-space of interest as done by Moravec [5], it is evident that this requires huge amounts of memory and is computationally expensive. Additionally, in the context of object pose tracking most of the space is *not* occupied by the object(s) being tracked. Therefore those grid cells will have uniformly low values. Closed form (e.g. Gaussian) unimodal representations of the belief distribution are generally not suitable in cluttered scenes, since the belief will be multi-modal. Isard and Blake propose factored sampling to represent such non-gaussian densities. They developed the Condensation Algorithm for tracking belief distributions over time, based on a dynamic model of the process and observations from image sequences. Details of their method can be found in [4].

For the context of this paper it is important, that it is an iterative method in which the density is represented as a weighted set $\{s^{(n)}\}$ of N samples from the state space of the dynamic system (in our case from the 3D space of possible position hypotheses). The samples are drawn according to the prior belief distribution $p(\underline{x})$ and assigned weights $\pi^{(n)} = p(z|\underline{x} = s^{(n)})$ and z being an observation. The $\pi^{(n)}$ are normalized to sum up to 1. The subsequent distribution is predicted using a dynamic model $p(\underline{x}_t|\underline{x}_{t-1})$. The weighted set $\{s^{(n)}, \pi^{(n)}\}$ represents an approximation of the posterior density $p(\underline{x}|z)$ which is arbitrarily close for $N \rightarrow \infty$. Accordingly the observation density has to be evaluated at $\underline{x} = s^{(n)}$ only. Thus, only small portions of the images have to be processed. The number of samples N determines the computational effort required in each time step.

2.3 Simultaneous Solution to the Segmentation, Object Recognition, 3D Localization and Tracking Problems

Our approach is to use Condensation for representing and tracking the belief $p(\underline{x})$ of the object of interest being at \underline{x} . The state space used in our implementation is the three-dimensional vector \underline{x} describing the object position in space. The current system does not model object rotations. As can be seen from the short description of the Condensation algorithm given in section 2.2 we have to model our objects dynamics $p(\underline{x}_t|\underline{x}_{t-1})$. The "dynamic" model currently used is a simple three-dimensional random walk with a Gaussian conditional density. Of course, this will be extended in the future. Having specified the very simple dynamic model, we have to define the evaluation procedure of the observation density $p(z|\underline{x})$, i.e. the probability for the observation z , given that the object is at \underline{x} .

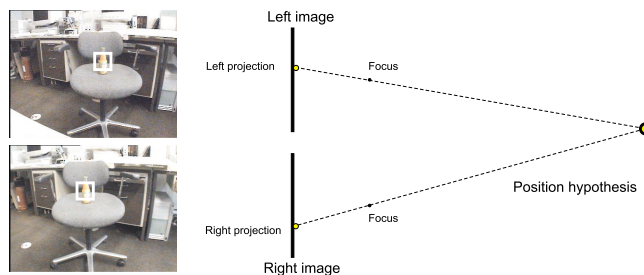


Fig. 1. A simple stereo setup.

We use a calibrated stereo setup. Therefore we are able to project each position hypothesis \underline{x} into both images (see figure 1). A hypothesis is probable, if the local image patches centered at the projected points in both images satisfy two constraints:

Consistency Constraint: First of all both patches have to depict the same portion of the object, i.e. look nearly the same after being appropriately warped to compensate

for the different viewpoints. This is implemented by computing the normalized correlation of both patches. The constraint is satisfied if the correlation result exceeds an empirically determined threshold. The satisfaction of this constraint implicitly contains a range based segmentation of the scene.

Recognition Constraint: In addition both patches should depict portions of the object the system is looking for. Here we use a probabilistic recognition system that currently takes color as the local appearance feature. Our measurements m_k used by equation 3 are 16 different colors obtained by an unsupervised vector quantization (using the K-Means algorithm on characteristic images) of the UV-sub-space of the YUV-color-space. The constraint is satisfied if the likelihood ratio of equation 3 is bigger than the detection threshold $\lambda = 150$. The recognition part of our algorithm is thus similar to those presented in [12] and [3].

If both constraints are satisfied, the observation density $p(z|\underline{x})$ is computed as follows

$$p(z|\underline{x}) = \prod_j p(m_j(\underline{x}_j^l)|o) \prod_k p(m_k(\underline{x}_k^r)|o)$$

where $m_k(\underline{x}_i^{l,r})$ is the local appearance measurement computed at the projection $x_i^{l,r}$ of the position hypothesis \underline{x} into the left and right images. The object position is computed as the first moment of $p(z|\underline{x})$, which in turn is approximated by the weighted sample set¹.

3 Experimental results

The combined object recognition, localization and tracking method presented in the previous sections has been implemented and evaluated using real world data. The task in the experiments was to discover and track a known object (bottle of orange juice) under realistic environmental conditions (complex background, dynamic scene). Object model ($p(m_k|o)$) and background model ($p(m_k|\bar{o})$) were estimated from training images. The position belief distribution $p(\underline{x})$ was initialized to be uniform inside the field of view up to a distance of $5m$.

Fig. 2 shows that only 14 iterations are required to initially recognize and localize the object in a distance of $1.5m$. It is obvious, that this initial phase could not be represented using a uni-modal Gaussian model of the belief distribution. The approximation used by the Condensation Algorithm is able to represent the belief with only 1000 samples. After the initial discovery of the object, the samples are concentrated in a very small portion of the search space and cover only one quarter of the image. The computation of the feature values has to be performed in this portion of the images only. It has to be noted, that there is *no* expensive search for stereo correspondence in the disparity space.

The experimental setup selected for this paper ensures that the object recognition part of the problem can reliably be solved, the features used are "appropriate" for the

¹ This is based on the assumption of a single object being tracked.

problem. The experiment shows, that if this is the case, the integrated approach presented in this paper can segment the scene as well as recognize, localize and track the object of interest.

The image sequence shown in fig. 2 has a duration of 1.5s after which the object was found and localized. On our Pentium II/400 computer we can compute around 10 iterations per second and our implementation still has room for significant optimizations.

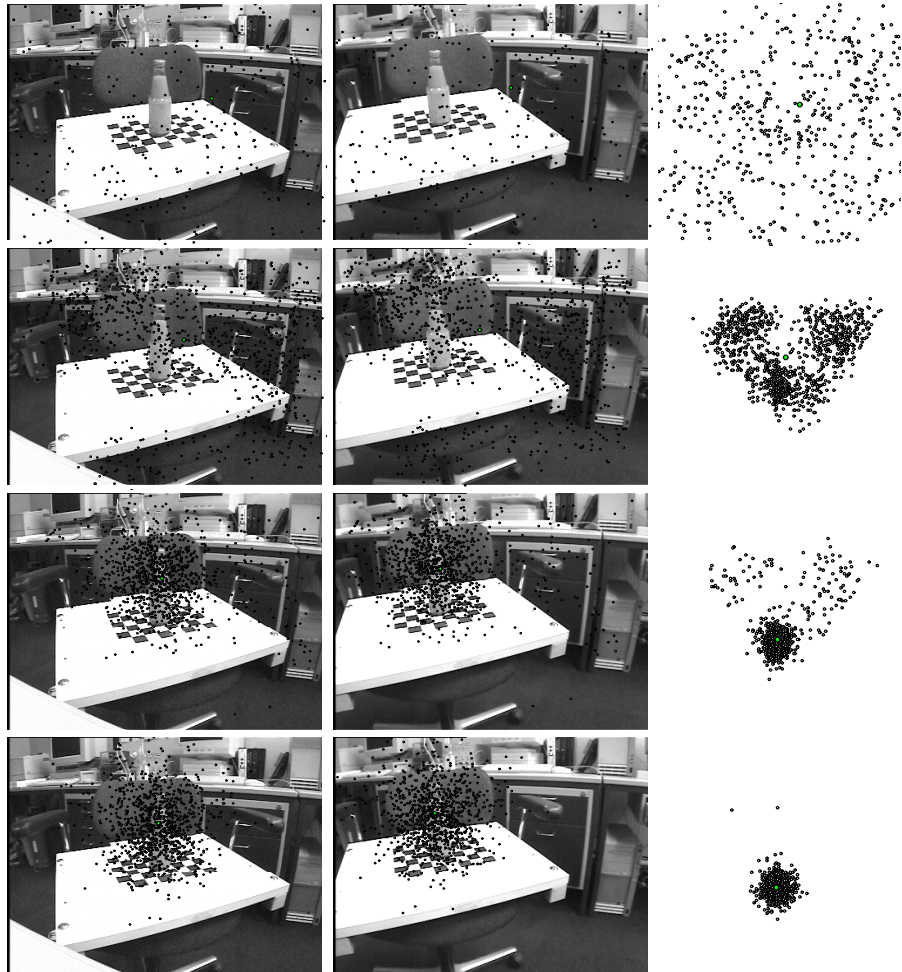


Fig. 2. The convergence of sampled belief approximation after 0, 5, 10 and 14 iterations on a static scene. The images show the sample ($N = 1000$) set projected onto left and right image and the floor plane.

After the convergence of the belief distribution, the object can reliably be tracked by the system. Fig. 3 depicts the trace of a tracking experiment, where the bottle was manually moved, (approximately) on a straight line. The images show the scene before the author’s hand enters it. The small dots depict the center of mass of the belief distribution, estimated from the samples at each iteration during the experiment. Even with our slow cycle of only $10Hz$ the tracker locks onto the object robustly. This is due to the tight coupling with the object recognition, which from the tracking point of view provides a sharp object-background separation.

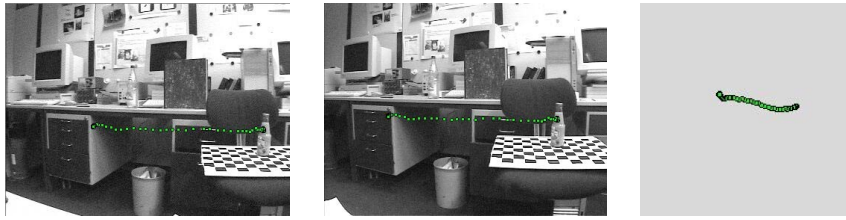


Fig. 3. The trace of a short tracking experiment.

4 Conclusions

We have presented a novel and very efficient approach for a probabilistic integration of recognition, 3D-localization and tracking of objects in complex scenes. Experiments indicate that this approach is viable and gives very satisfactory results. It is important to note, that the single components of our system are still very simple. The probabilistic model does not account for spatial dependencies among the image measurements m_k (and thus prohibits the estimation of object rotations), the color features we use will not be sufficient for more complex objects and finally the underlying dynamic model of the condensation tracker is extremely limited. But these limitations can easily be overcome using more advanced and well known algorithms especially for feature extraction and object modeling (see for example [11] for wavelet features incorporating spatial dependencies). Future work will focus on these improvements and other new features as the incorporation of multiple objects. In addition the system will be integrated on our mobile manipulation test-bed MobMan [15].

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) under contract no. 01IL902D0 (MORPHA).

References

- [1] Wolfram Burgard, Dieter Fox, Daniel Henning, and Timo Schmidt. Estimating the absolute position of a mobile robot using position probability grids. Technical report, Universität Bonn, Institut für Informatik III, 1996.

- [2] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing. Academic Press, Inc., 2 edition, 1990.
- [3] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.
- [4] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Intern. Journal on Computer Vision*, 29(1):5–28, 1998.
- [5] Hans P. Moravec. Robot spatial perception by stereoscopic vision and 3d evidence grids. Technical Report CMU-RI-TR-96-34, Carnegie Mellon University, Robotics Institute, Pittsburgh, USA, 1996.
- [6] Hans P. Moravec and Alberto Elfes. High resolution maps from wide angle sonar. In *Intern. Conf. on Robotics and Automation*, pages 19–24, 1985.
- [7] J. Peng and B. Bhanu. Closed-loop object recognition using reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(2):139–154, 1998.
- [8] M. Reinhold, F. Deinzer, J. Denzler, D. Paulus, and J. Pösl. Active appearance-based object recognition using viewpoint selection. In B. Girod, G. Greiner, H. Niemann, and H.-P. Seidel, editors, *Vision, Modeling, and Visualization 2000*, pages 105–112. infix, Berlin, 2000.
- [9] Bernt Schiele and James Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *Proc. of the Intern. Conf. on Pattern Recognition (ICPR'96)*, pages 50–54, 1996.
- [10] Bernt Schiele and James Crowley. Where to look next and what to look for. In *Proc. of the Conf. on Intelligent Robots and Systems (IROS'96)*, pages 1249–1255, 1996.
- [11] Henry Schneiderman and Takeo Kanade. A statistical model for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2000.
- [12] M. J. Swain and D. H. Ballard. Color indexing. *International Journal on Computer Vision*, 7(1):11–32, 1991.
- [13] S. Thrun, D. Fox, and W. Burgard. Monte carlo localization with mixture proposal distribution. In *Proc. of the Seventh National Conference on Artificial Intelligence (AAAI)*, 2000.
- [14] Sebastian B. Thrun. A bayesian approach to landmark discovery and active perception in mobile robot navigation. Technical Report CMU-CS-96-122, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 1996.
- [15] Georg von Wichert, Thomas Wösch, Steffen Gutmann, and Gisbert Lawitzky. MobMan – Ein mobiler Manipulator für Alltagsumgebungen. In R. Dillmann, H. Wörn, and M. von Ehr, editors, *Autonome Mobile Systeme 2000*, pages 55–62. Springer, 2000.