LOCALIZATION AND POSE ESTIMATION OF TEXTURELESS OBJECTS FOR AUTONOMOUS EXPLORATION MISSIONS

Nicolai Wojke, Frank Neuhaus, and Dietrich Paulus

Active Vision Group, University of Koblenz-Landau
56070 Koblenz, Germany

ABSTRACT

In this paper we describe an approach for detection and pose estimation of colored objects with only few or no textural features. The approach consists of two separate stages. First, we perform vision-based object detection and hypothesis filtering. Then, we estimate and validate the object's pose in 3-D laser scans. For object detection we integrate image segmentation results from multiple viewpoints in a set-theoretical filter that provides a probabilistically sound estimate of the number of objects and their respective locations. For validation and pose estimation we search for the best pose by sampling from a geometric measurement model. The system has been validated during autonomous exploration missions in unstructured and space-like environments.

Index Terms— Object Localization, Pose Estimation, Hierarchical Sampling, Robot Vision

1. INTRODUCTION

The *DLR SpaceBotCamp 2015*[1] was a national challenge organized by the German Aerospace Center (DLR) to test robotic space exploration, object discovery, and manipulation capabilities. Though the method that we present here has been developed with this application scenario in mind, it can be applied to a broader range of applications. In particular, we target any application that requires vision-based localization of colored objects in far range and precise pose estimation in close range. Further, we focus on objects with few or no textural features. This lack of unique visual cues poses a challenge for the recognition system. With color being the primary feature, appearance is subject to varying lighting conditions. Consequently, day-time and view-dependent appearance changes as well as high clutter rates are characteristic challenges of the given scenario.

Recent progress in object detection and pose estimation is largely due to the development of stable keypoint descriptors that allow for robust matching between images (e.g., [1, 2]) and/or object geometry (e.g., [3, 4]). Usually, these methods use gradient or histogram information to describe the local neighborhood of distinct keypoints and, consequently, they are less applicable for geometrically simple, textureless objects. Similarly, performance of methods based on global point pair feature descriptors [5] usually degrades for small and geometrically simple objects. Several extensions have been proposed to overcome this issue by taking into account additional information such as color [6, 7]. If an initial estimate of the object pose is available, there exist several methods for pose registration (e.g., [8, 9, 10]).

On the other hand, pixel-wise image segmentation has seen significant progress in recent years. It has become increasingly popular to formulate segmentation in a conditional random field (CRF) framework to incorporate structural information into the segmentation process and there is an ongoing effort to increase model complexity (e.g., [11, 12, 13]) and efficiency of associated inference algorithms (e.g., [14, 15]). Recently, it has also been shown that deep neural networks can be used for image segmentation [16, 17].

In this paper, we integrate image segmentation results from multiple viewpoints in a set-theoretical filter that estimates the number of objects and their respective locations in a 2.5-D terrain grid (Section 2). For pose estimation and validation, we evaluate a geometric measurement model using a hierarchical particle filter that allows for efficient sampling when initial uncertainty is high (Section 3). In Section 4 we evaluate the approach with respect to pose estimation accuracy and overall system performance. In Section 5 we present our conclusion.

2. OBJECT LOCALIZATION

In this section, we describe our vision-based approach to object localization. First, we perform color image segmentation to find object candidates. Then, we use an online filtering framework on a discretized height map to integrate segmentation results from different viewpoints.

2.1. Image Segmentation

For color image segmentation we use a pairwise conditional random field (CRF) model that is defined on graph $G(\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ corresponds to an image

[1] `http://spacebotcup.uni-koblenz.de`, `http://www.spacebotcup.de`

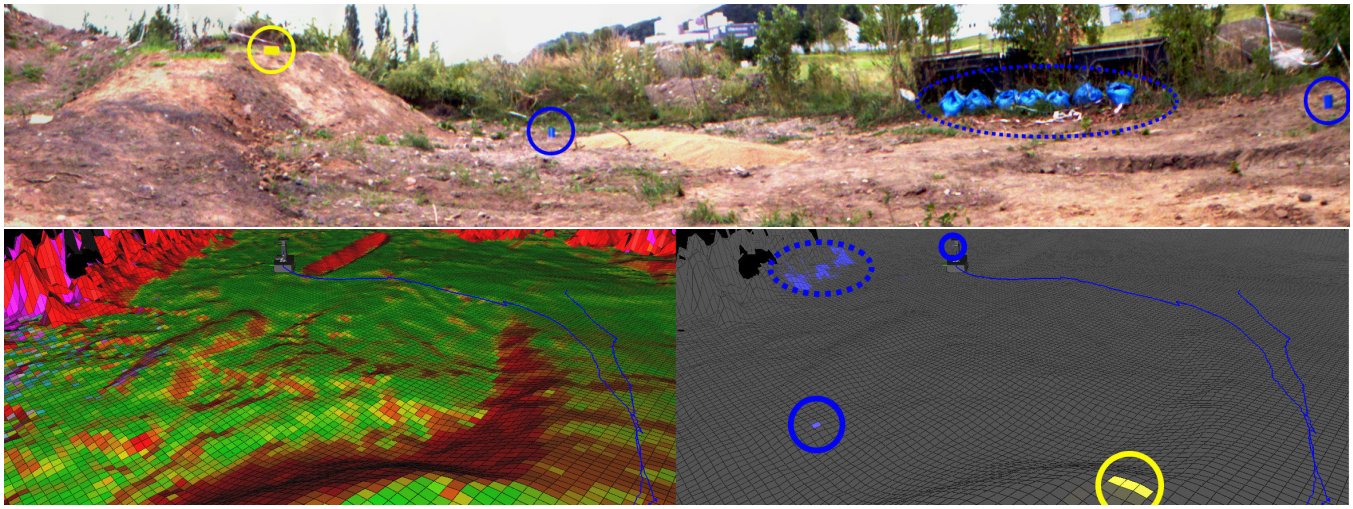

Fig. 1: Top: Panorama of test arena. Bottom left: Height map visualization. Bottom right: Candidate object map (stronger colors indicate higher probability). Circular markers highlight true object locations and clutter.

pixel and an edge $(i, j) \in \mathcal{E}$ between nodes i and $j \in \mathcal{V}$ is added if their corresponding locations are adjacent. Let $\boldsymbol{y} = (y_1, \ldots, y_{|\mathcal{V}|})^{\mathrm{T}}$ denote an image labeling that assigns a class to each pixel and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|\mathcal{V}|})^{\mathrm{T}}$ denote the observed data, in our case pixel colors in Lab space. The CRF energy function of this model can be written as

$$E_{\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{X}) = \sum_{i \in \mathcal{V}} D_i(y_i, \boldsymbol{X}) + \sum_{(i,j) \in \mathcal{E}} V_{ij}(y_i, y_j, \boldsymbol{X}),$$

where D_i is the unary term and V_{ij} is the pairwise term. This energy is conditional on CRF parameters $\boldsymbol{w}$ and the optimal labeling is the one that minimizes this cost. For the application at hand, we create binary foreground-background segmentations for each class of objects individually and, consequently, using a submodular energy for the pairwise term, inference in this model can be carried out exact and efficiently using graph cuts [14].

As we expect bright objects with only few textural cues and heterogeneous, cluttered background, our unary term is a linear model based on color histogram backprojection and Gabor filter responses. More specifically, for the unary term we write

$$D_i(y_i, \boldsymbol{X}) = \boldsymbol{w}_{\mathrm{D}} \cdot \gamma_i(\boldsymbol{x}_i),$$

where feature function γ_i returns the following six entries: (1) results of histogram backprojection for a foreground and background color model and (2) the response of a four dimensional Gabor filter bank that contains edge filters of different scale and rotation. We use the Lab color space for histogram backprojection and apply Gabor filters on the luminance channel. For the neighborhood term we use a submodular energy based on color difference in Lab space:

$$V_{ij}(y_i, y_j, \boldsymbol{X}) = \begin{cases} 0 & \text{if } y_i = y_j, \\ w_V \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 & \text{otherwise.} \end{cases}$$

During training, we first learn foreground and background color models for histogram backprojection on a set of training images. Then, we learn the 7-dimensional CRF parameters $\boldsymbol{w} = (\boldsymbol{w}_{\mathrm{D}}, w_{\mathrm{V}})$ in a max-margin framework using a structured support vector machine [15]. At test time, we use graph cuts [14] for inference.

2.2. Discrete Filtering

In order to deal with errorneous segmentations and clutter, we integrate image segmentation results from different viewpoints. For this purpose, we keep a discretized height map representation of the environment (c.f. Figure 1) that is built incrementally from 3-D scans using a graph-based SLAM algorithm. At each timestep, we obtain an image segmentation for each object class as well as the most current height map. In order to integrate segmentation results over time, we first compute the set of visible cells by reprojecting their respective centers into the image, carefully taking care of occlusions using a depth buffer. Then, we compute a cell-specific probability of detection based on the distance to the camera. For filtering, we apply a discrete-space implementation of the Probability Hypothesis Density filter [18] to obtain an estimate of the number of objects occupying each cell. By thresholding, we extract cells that are likely to contain objects and use those cells for subsequent pose estimation and validation. The concrete details of this method are beyond the scope of this paper and we refer the reader to [18, 19] for a comprehensive introduction to this topic.

3. POSE ESTIMATION

The final step of our object localization pipeline is pose estimation. For this purpose, we acquire a high resolution 3-D

laser scan in close range to the candidate object location and find the set of foreground measurements using the image segmentation method described in Section 2. Then, we search for the full 6-D object pose by sampling from a geometric measurement model.

3.1. Measurement Likelihood

Let $\boldsymbol{x}$ denote the 6-D object pose and let $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ denote the set of point measurements obtained from a 3-D laser range finder that have been classified as object surface readings based on color image segmentation. Further, let $\mathcal{M}$ denote the object model. Then, we compute the measurement likelihood as follows. First, we calculate the closest distance of each reading to the object surface $d_{\mathcal{M}}(\boldsymbol{z}_i, \boldsymbol{x})$ under the current pose transformation $\boldsymbol{x}$ to compute the fraction of inliers, i.e., points on the object surface, wrt. some fixed surface width ϵ:

$$r_\epsilon(\boldsymbol{x}) = \frac{1}{N}|I_\epsilon(\boldsymbol{x})|,$$

where $I_\epsilon(\boldsymbol{x}) = \{\boldsymbol{z}_i \in Z \mid d_{\mathcal{M}}(\boldsymbol{z}_i, \boldsymbol{x}) \leq \epsilon\}$. Then, we model the measurement likelihood as normal distribution on the fraction of inliers

$$p_\epsilon(Z \mid \boldsymbol{x}) \propto \exp\left(-\frac{(r_\epsilon(\boldsymbol{x}) - \mu)^2}{2\sigma^2}\right),$$

with mean μ and standard deviation σ. The parameters have been found through experimental evaluation. In this model, the object surface width ϵ has a smoothing effect. When the surface width is too narrow, the measurement likelihood becomes extremely peaked as small deviations from the true object pose lead to a rapid decline in the number of inliers. A large surface width on the other hand relaxes the model, allowing deviations from the true object pose without applying any penalty. It is well known that, in practice, peaked measurement likelihoods can lead to problems in probabilistic localization approaches as the likelihood function does not provide enough guidance towards its peak (e.g., [20]). Consequently, we make explicit use of the surface width to guide our sampling-based inference method.

3.2. Hierarchical Pose Sampling

Using the image segmentation and filtering approach of Section 2, we have a rough estimate of the object's position. However, without any information about the object's orientation, information about the full 6-D pose are highly uncertain. In order to efficiently sample from the measurement likelihood under this conditions, we use a hierarchical particle filter that gradually anneals the surface width ϵ from an initially high value to a small target value. The *Scaling Series* algorithm [21] has previously been used in a similiar situation to localize vehicles in two dimensional scans of a laser range finder [22, 23] and proceeds as follows: Given an initial guess for the object pose and an initial (large) value for the surface width, a set of samples is drawn from a hypersphere in parameter space surrounding the inital guess and evaluated according to the measurement likelihood. After resampling, the surface width is reduced according to an annealing schedule and the volume of the hypersphere surrounding each sample is halfed. Then, a new set of samples is drawn using rejection sampling. In this way, the algorithm produces a gradually more informative proposal distribution for the final importance sampling step, where samples are gathered around areas of high probability mass.

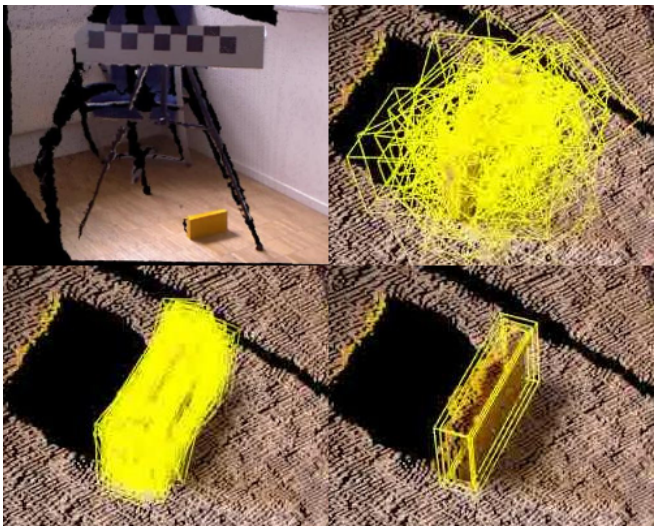

Fig. 2: Particle visualization at different iterations of the Scaling Series algorithm [21]. The object surface is gradually decreased (left to right).

Figure 2 shows an application of this algorithm for the purpose of estimating the pose of a box-shaped object. In early iterations, samples are located around the true object position with no distict orientation. With increasing number of iterations, the measurement model becomes more informative and samples are distibuted around the true object pose.

3.3. Pose Registration and Validation

We further refine the output of our sampling-based method using the Iterative Closest Point (ICP) algorithm [8], where we use the highest ranked particle's pose as initial guess. Then, we validate the final pose hypothesis by thresholding the fraction of foreground readings that fall onto the object surface. Consquently, object hypotheses that have passed the validation criterion are known to comply with the color as well as the geometric object model.

	Run-time (s)		Mean Error	
	Mean	Std	Position (cm)	Orientation (°)
Battery	2.17	0.44	0.52	1.50
Cup	0.72	0.01	0.36	5.92

Table 1: Run-time and accuracy of pose estimation evaluated on 20 scenes in total, captured with a Microsoft Kinect.

	Pixel-level		Object-level		
	Precision	Recall	TP	FP	FN
Battery	0.92	0.60	9	0	4
Cup	0.73	0.55	26	6	6

Table 2: Image segmentation results at pixel and object level (TP=true positives, FP=false positives, FN=false negatives).

4. EXPERIMENTS

We have evaluated our method with respect to accuracy of pose estimation as well as overall recognition performance. All experiments have been run on a desktop computer with an Intel(R) Core(TM) i7-2760QM CPU.

4.1. Accuracy of Pose Estimation

The run-time and accuracy of our pose estimation algorithm has been evaluated for a box-shaped object of dimension $20\,\text{cm} \times 4\,\text{cm} \times 10\,\text{cm}$ (*battery*) and a cylindrically shaped object of height 12 cm and radius 8 cm (*cup*), both of which have been used for manipulation tasks during the DLR SpaceBotCamp 2015. For the box-shaped object the initial surface width was set to 10 cm such that in early iterations the entire inner volume is occupied by the surface boundary. In a total of 16 iterations the surface width was shrinked to a target value of 3 cm. For the cylindrically shaped object the initial surface width was set to 8 cm and gradually annealed down to a target value of 3 cm in 12 iterations. For each object we have taken 10 scenes from different viewpoints in range between 0.5 m and 1 m. The ground truth pose has been established through manual annotation.

Results of our evaluation are shown in Table 1. Run-times differ between the two objects. In particular, they are considerably higher for the box-shaped object than for the cylindrically shaped object. This is due to two reasons. First, the larger number of iterations in the annealing schedule requires more time. Second, the object geometry is more complex. In particular, for the box-shaped object the hierarchical sampling procedure follows multiple modes in early iterations whereas there is a rapid focus on one mode for the cylindrically shaped object. In terms of accuracy, we archieved mean error in translation below 6 mm and mean error in orientation below 6°. Higher error rates for the cylindrically shaped object are due systematically missing and misplaced surface readings on the back side, which sometimes caused slight misalignment in orientation. All poses obtained during our experiment were suitable for object gripping.

4.2. Overall Recognition Performance

We have further evaluated the system in a realistic exploration scenario: The robot has been placed in an outdoor test arena that contained two cups and one battery. In addition, a set of blue plastic bags served as clutter. During exploration, the robot took a total of 10 laser scans. For object localization, 80 images, captured while robot was driving, have been segmented and fed into the filtering framework. A panorama image built from some of these images as well as the generated terrain height map are shown in Figure 1.

Table 2 shows results of our image segmentation method, where we count object detections based on a 30% overlap with the true image location. For both objects, we got relatively low recall rates at pixel-level and a considerable amount of false negatives at object-level. We found that, in practice, color-based image segmentation worked well in distances of up to 5–8 m. Most missed detections are at greater distance or are due to strong color distortions caused by direct sunlight. Further, the intentionally placed blue bags lead to a considerable drop in cup detector precision. A visualization of the filtered object candidate map is shown in Figure 1 (bottom right). Due to integration of consecutive segmentation results, all three objects are clearly visible in the final map. Due to similarity in color, the blue plastic bags are also found as candidate locations. However, during our test run, they have been rejected at geometrical validation after pose estimation. Using this approach, we have successfully located all objects during the SpaceBotCamp 2015.

5. CONCLUSION

We have presented an object recognition pipeline where we generate candidates using a color-image segmentation approach and where we perform pose estimation in close range using 3-D laser scans. In order to deal with errorneous segmentations and clutter, we integrate segmentation results from multiple viewpoints in a probabilistic filter. Final candidate validation is based on a geometric measurement model.

Based on our experiments, accuracy of pose estimation is high. Further, the presented method is general and can be adapted to specific application requirements by changing parts of the overall recognition pipeline. For example, one may integrate higher-oder potentials into the CRF framework when complex appearance models require more structural information. Further, the geometric measurement model could be implemented in a way that allows for more generic object geometries. Possibly, this could be done using OpenGL with custom shaders by rendering the object at different scales.

6. REFERENCES

[1] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Und.*, vol. 110, no. 3, pp. 346–359, 2008.

[3] Radu Bogdan Rusu, Gary R. Bradski, Romain Thibaux, and John M. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 2155–2162.

[4] Aitor Aldoma, Markus Vincze, Nico Blodow, David Gossow, Suat Gedikli, Radu Bogdan Rusu, and Gary R. Bradski, "CAD-model recognition and 6DOF pose estimation using 3D cues," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 585–592.

[5] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 998–1005.

[6] Eunyoung Kim and Gérard G. Medioni, "3d object recognition in range images using visibility context," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2011, pp. 3800–3807.

[7] Changhyun Choi and Henrik I. Christensen, "3d pose estimation of daily objects using an RGB-D camera," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 3342–3349.

[8] Paul J. Besl and Neil D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[9] Samuel Dambreville, Romeil Sandhu, Anthony J. Yezzi, and Allen Tannenbaum, "Robust 3D pose estimation and efficient 2d region-based segmentation from a 3D shape prior," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 169–182.

[10] Victor Adrian Prisacariu and Ian D. Reid, "PWP3D: real-time segmentation and tracking of 3D objects," *Int. J. Computer Vision*, vol. 98, no. 3, pp. 335–354, 2012.

[11] Pushmeet Kohli, Lubor Ladicky, and Philip H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.

[12] Xuming He, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 695–702.

[13] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems*, 2011, pp. 109–117.

[14] Yuri Boykov and Vladimir Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, 2004.

[15] Martin Szummer, Pushmeet Kohli, and Derek Hoiem, "Learning crfs using graph cuts," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 582–595.

[16] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.

[17] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation," *Computing Research Repository (CoRR)*, vol. abs/1505.04366, 2015.

[18] Ronald. P. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, 2003.

[19] Ronald. P. Mahler, *Statistical multisource-multitarget information fusion*, Artech House, Inc., 2007.

[20] Patrick Pfaff, Christian Plagemann, and Wolfram Burgard, "Improved likelihood models for probabilistic localization based on range scans," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 2192–2197.

[21] Anna Petrovskaya, Oussama Khatib, Sebastian Thrun, and Andrew Y. Ng, "Bayesian estimation for autonomous object manipulation based on tactile sensors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2006, pp. 707–714.

[22] Anna Petrovskaya and Sebastian Thrun, "Model based vehicle detection and tracking for autonomous urban driving," *Auton. Robots*, vol. 26, no. 2-3, pp. 123–139, 2009.

[23] Nicolai Wojke and Marcel Häselich, "Moving vehicle detection and tracking in unstructured environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 3082–3087.