2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

A Gaussian Measurement Model for Local Interest Point Based 6 DOF Pose Estimation

Thilo Grundmann Wendelin Feiten Georg v. Wichert

***Abstract*— One of the main challenges for service robots during operation lies in the handling of unavoidable uncertainties which originate from model and sensor inaccuracies and which are characteristic for realistic application scenarios. Robustness under real world conditions can only be achieved when the dominant uncertainties are explicitly represented and purposefully managed by the robot's control system. We therefore adopt a probabilistic approach in which perception is regarded as a sequential estimation process and follow a Bayesian filtering methodology. Under these assumptions probabilistic models of the robot's perception systems are key.**

In this paper we shortly describe a model based object recognition and localization system. However, we do not not focus on the 6D pose estimation procedure itself, but on the method to quantify and compute the uncertainty associated with it. We construct a Gaussian approximation of the resulting pose error using the implicit function theorem. It is then used as a proposal density for importance sampling. Our goal is to sample from the measurement model describing 6D object localization based on local features in a Bayesian filtering context.

I. Introduction

Precise object localization in all six Cartesian dimensions is essential to all service robotic scenarios in which the robot interacts with and purposefully manipulates the environment.

We regard perception as a sequential estimation process, because in realistic application scenarios, we expect that isolated sensor measurements rarely provide information precise enough for e.g. grasping an object in a cluttered scene. This is due to unavoidable uncertainties in sensors and models. We therefore follow a Bayesian filtering methodology to fuse multiple sensor readings over time.

Equation (1) describes the recursive estimation of the belief $Bel(\omega_t)$ at time t over the world state ω_t, z_t denotes the sensor measurements and u_t represents the actions that modify the overall world state ω_t (here η is a normalization constant to be chosen such, that the resulting belief distribution integrates to 1).

$$Bel(\omega_t) = \eta \; p(z_t|\omega_t) \int p(\omega_t|u_t, \omega_{t-1}) Bel(\omega_{t-1}) d\omega_{t-1} \tag{1}$$

In this paper we present the details of the sensor model $p(z_t|\omega_t)$ used in this filtering process. The focus of this paper is *not* on the rather obvious 6D object pose estimation procedure itself, but on the method to quantify and compute the uncertainty associated with this procedure. We show how to sample directly from the observation model using an intermediate Gaussian approximation of the pose reconstruction errors as a proposal for importance sampling.

Thilo Grundmann, Wendelin Feiten and Georg v. Wichert are with Corporate Technology, Intelligent Systems & Control, Siemens AG, D-80200 Munich, Germany

Fig. 1. A precise model of the errors that occur during 6D object localization is essential for many robotic tasks.

We use Lowe's SIFT method [1] for determining local, scale-invariant features in images. Doing this on both left and right images from a calibrated stereo camera with consecutive matching of the detected feature points from both images allows for the reconstruction of 6D object poses. Our appearance- and model-based approach consists of two separate stages: *Model generation* and *Object recognition and pose estimation*.

Model generation is an off-line process, where the object database is generated from training data. Object pose estimation at run-time is based on matching local features extracted from the current camera images with features from the previously built object database.

The underlying object recognition system [2] has been developed for the anthropomorphic service robot shown in Fig. 1. It is the basis for various other research in the area of object manipulation [3][4], probabilistic perception planning [5] and scene analysis [6].

The remainder of the paper is organized as follows: Section II outlines current state of the art approaches to model-based object recognition, model generation and probabilistic models for stereo vision. In Section III our model generation procedure and the method for object 6D pose estimation is described. Section IV explains the probabilistic sensor model which is required for multi view fusion. In order to sample from this sensor model we follow the importance sampling idea using a Gaussian proposal which is described in Section V. Finally, in Section VI, we report on experiments which demonstrate the proposed theoretical concepts on real data.

978-1-61284-385-8/11/$26.00 ©2011 IEEE

II. Related work

Object recognition and localization has attracted interest of the scientific community for a long time. Early attempts restricted the pose dimension to one [7] or three [8][9].

Point correspondence from wire-frame model corners were used by Kragic [10] to find full 6D poses using a mono camera and POSIT [11]. This particular approach is restricted to objects that can be modeled by wire frames, but the underlying concept could be generalized by using interest points like SIFT [12] or SURF [13] that can be found in large numbers on textured objects.

All approaches to 6D pose estimation that use interest points belong to the class of model based recognition methods and therefore need a method to create the required models. Gordan and Lowe [14] proposed a system that constructs 3D model of SIFT features from arbitrary object images using bundle adjustment off-line and determines the 6D pose through the use of RANSAC [15] and the Levenberg-Marquardt algorithm. Azad et. al. presented a method stereo vision based [16] for full 6D pose retrieval of textured objects using classic SIFT interest points. The method requires the objects to possess flat surfaces for the stereo recognition and no empirical evaluation of the accuracy of the pose is given.

Recently Collet et al. presented an object localization system [17] using a monocular camera, based on SIFT features, which uses RANSAC and mean shift clustering to generate object pose hypotheses. They also describe an almost fully automatic process for the model generation and give some experiments with four objects where measured poses are evaluated against ground truth. The error is described by two histograms over the translational and rotational error.

Our localization approach is mostly comparable to the method of [17], using a comparable 3D model which is not restricted in the shape and is also based on SIFT features. By using stereo image pairs, and thus conceptually different methods for the pose determination, a higher accuracy in pose estimation is achieved. None of the other localization methods mentioned above have proposed a model to describe the uncertainty of the resulting pose estimate.

III. Stereo-based 6D object pose estimation

Robotic manipulation in realistic scenes imposes high demands on the precision of 6D object localization, since objects will in general be close to each other, without large clearance between them. We chose a stereo-based approach, which for obvious geometric reasons can be expected to deliver superior pose accuracy compared to monocular approaches [18]. In this section we first describe the model generation process and the reconstruction of the 6D object pose using local features.

A. Model Generation

In the model generation step we compute models of object (class) appearance and geometry based on a set of training data. The KIT object modeling center IOMOS [19] is used to acquire stereo images and a precise 3D surface point cloud for each of the objects to be modeled.

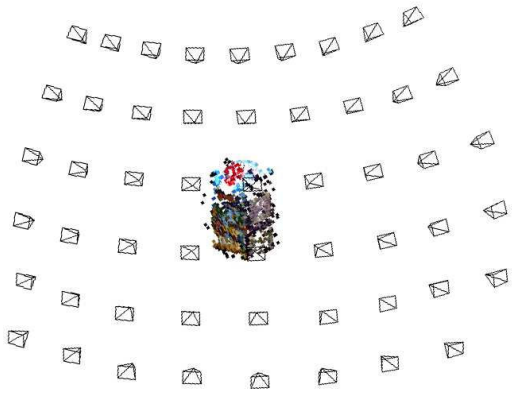


Fig. 2. Bundler result: 3D points and camera poses

For the construction of 3D object models we employ bundle adjustment techniques. Bundle adjustment [20] aims at finding camera poses and 3D model points simultaneously by optimizing the re-projection error of the object model points into the training images.

We use the Bundler software [21] to compute the camera poses and 3D locations (Fig. 2) for all matching interest points in conjunction with a list of v corresponding SIFT descriptor vectors per point, where v is the number of images in which the respective feature was detected and matched. A standard 2D SIFT feature point is described by its 2D location in the image, its scale s and orientation o, and the 128 dimensional descriptor d. The 3D SIFT object model M consists of a set of K 3D interest points $\mathbf{m}_k$ with $k \in [0, K]$, which have a 3D location $\mathbf{x} = \{x, y, z\}$ with a covariance matrix $\Sigma_{\mathbf{x}}$, a list of v lines of sight $\tilde{\mathbf{x}}^v$, a scale s and a classic SIFT descriptor d.

$$\mathbf{m}_k = (\mathbf{x}, \Sigma_{\mathbf{x}}, \tilde{\mathbf{x}}^v, s, d). \tag{2}$$

The 3D location $\mathbf{x}$ is computed using Bundler, the corresponding location covariance matrix is approximated using the covariance matrix of the minimal distances between the calculated 3D feature point location $\mathbf{x}$ and the corresponding lines of sight. The descriptor is the mean value of all v descriptors which contributed to the 3D point. The list of lines of sight consists of normalized vectors from the 3D interest points to the v camera poses and represents, in conjunction with the averaged scale, the range of viewing directions from where the interest point can be detected. Only 3D interest points with $v > 5$ sightings are used in the object models to filter out spurious 3D points. The full model for one object needs about 3% of storage of the initial sift features. This enables fast recognition since the database can be held in RAM completely.

B. 6D Object Localization

For object localization based on a pair of stereo camera images, we compute SIFT interest point locations and feature vectors on both images from the two cameras $j \in [L, R]$. From feature points $\hat{\mathbf{z}}_{i,j}$ with feature vectors matching across both images, we reconstruct a set $\hat{\mathbf{x}}_{\mathbf{i}}$ of 3D interest points

Fig. 3. a) 3D surface dot cloud, b) 3D model (colored) generated from training images by means of sparse bundle adjustment (SBA), c) SBA model overlaid with surface dot cloud obtained using a commercial 3D scanning system, d) real object

and store them along with their corresponding sightings $\hat{\mathbf{z}}_{i,j}$ in both images. We then use the corresponding SIFT feature vectors d to associate them with feature points $\mathbf{m}_k$ stored in our object database, and cluster them to form sets, which are consistent spatially and with respect to their object classes. From the resulting point sets, each of which implicitly represents an object class and pose hypothesis we compute a initial pose estimate $\hat{\omega}$ using a least squares fit with the corresponding object model point cloud, giving an initial pose estimate.

This process, which is described in detail in [2], delivers only a subset of the available features for several reasons: Firstly, all feature matchings are performed using only approximate nearest neighbor search techniques to reduce the computational effort. Secondly, the stereo approach requires a SIFT point to appear consistently in both images. Under partial occlusion valuable matches between a 3D model point m and observations $\hat{\mathbf{z}}$ from only one of the images are not taken into account. To take maximal advantage of all observations, we refine the initial pose estimate in a second step. After obtaining the initial hypothesis the set of possible feature matches to be examined can be reduced significantly compared to the unconstrained search above. This is done by choosing only a subset of interest points from both images and only a subset of the 3D interest points of the object model according to the following criterion: Image feature candidates are selected using a dilated bounding volume which is projected onto the images according to the object hypothesis. As model feature candidates we choose all 3D model points $\mathbf{m}$, which have a scale that is low enough to be seen at the distance constituted by the hypothesis, which also have modeled viewing directions $\tilde{\mathbf{x}}$ that are close to the current viewing direction, and that's projection lies on one image sensor.

The error model described in the next section is then used as objective function for an optimization of the pose minimizing the re-projection error in both image planes.

IV. Probabilistic Sensor Model

As already pointed out in Section I, the sensor model $p(\mathbf{z}_t|\omega_t)$ plays an important role in the sequential estimation process that models the perceptual activity of our system[1]. It is the basis for the fusion of sensor readings over time into a consistent model of the robot's environment.

Assuming correct correspondences between image and model interest points, as provided by the method described in Section III-B, the dominant sources of pose estimation errors are the localization of the interest points in both images and the uncertainty of the 3D interest points in the model database.

For an arbitrary object pose hypothesis ω in 6D camera coordinates C_j the projected image coordinates $\mathbf{z}_{i,j}$ in the image $j \in [L, R]$ can be computed assuming a standard pin hole camera model $h_j(\omega, \mathbf{x}_i)$ as

$$\mathbf{q}_{i,j} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} T_{C_j}(\omega)\ \mathbf{x}_i \tag{3}$$

$$\mathbf{z}_{i,j} = h_j(\omega, \mathbf{x}_i) = \frac{1}{q_{j\,z}} \begin{bmatrix} q_{j\,x} \\ q_{j\,y} \end{bmatrix} \tag{4}$$

Here f is the focal length of the camera and T_{C_j} is the homogeneous transformation from object to camera coordinates of the left and right images. For the set $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ of N interest point correspondences, $\mathbf{z}_n = \mathbf{z}_{i,j}$ with $n \in [1, N]$ and $N <= I \times J$ and an arbitrary object pose hypothesis ω in 6D camera coordinates we compute the observation density $p(\mathbf{z}|\omega)$ assuming a mutually uncorrelated detection of the interest points $\mathbf{z}_{i,j}$.

$$p(\mathbf{z}|\omega) = \prod_{i}^{I} p(\mathbf{z}_{i,L}|\omega) p(\mathbf{z}_{i,R}|\omega) = \prod_{n}^{N} p(\mathbf{z}_n|\omega) \tag{5}$$

We assume the location uncertainty of the 3D model database points $\mathbf{m}_i$ to be Gaussian with a covariance Σ_i and the detection of the interest point locations $\mathbf{z}_n$ also to be affected by Gaussian error with a covariance Σ_j. Σ_i is determined during step 4 of the model generation process,

[1]While in general the world state comprises the poses of several objects, we consider only a single object in this section, and use ω to refer to its 6D pose vector, also leaving out the temporal index t for notational simplicity.

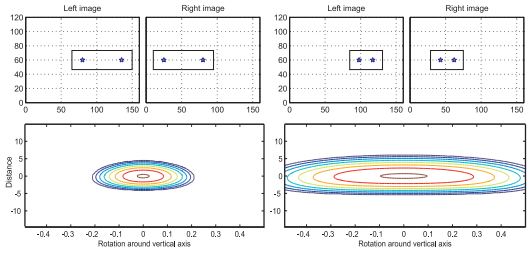


Fig. 4. Comparison of synthetic sensor model results for two objects of different width, represented each by two 3D feature points. Both objects are at the same distance from the stereo camera system. The upper plots show simulated camera images, the lower plots show the measurement uncertainties for both situations. Here we plot $p(\mathbf{z}|\omega)$ as a function of ω in the distance vs. vertical rotation space. As expected, the rotational component of the smaller object's pose-estimate has a significantly higher uncertainty.

see Section III-A. Σ_j has been empirically determined. We linearize the camera model to project the Gaussian model error Σ_i at the interest point location $\mathbf{x}_i$ into the image planes of the two cameras. Under these assumptions $\mathbf{z}_n$ is approximately normally distributed according to

$$J_{h_j} = \partial h_j(\omega, \mathbf{x}_i)/\partial \mathbf{x}_i \quad (6)$$

$$\Sigma_{i,j} = \Sigma_j + J_{h_j} \Sigma_i J_{h_j}^T \quad (7)$$

$$p(\mathbf{z}_n|\omega) = \mathcal{N}(h_j(\omega, \mathbf{x}_i), \Sigma_{i,j}) \quad (8)$$

where J_{h_j} is the Jacobian of the perspective projection (4) with respect to the 3D interest point locations $\mathbf{x}_i$. This is used in (5).

The sensor model captures the dominant uncertainties of the pose reconstruction process based on local point features. Especially the effect of the geometric configuration of the detected 3D feature points on the estimated object orientation is correctly covered as shown in Fig. 4 using a synthetic example. Narrowly spaced feature points lead to significantly increased angular uncertainty in the pose estimate, compared to spatially wider distributed feature points. Narrow or partially occluded objects frequently cause such problems in our household scenarios (Fig. 1) leading to significant pose-estimation errors in reality. If such uncertainties are correctly represented, they can be compensated by performing additional sensor measurements [5] in order to reach the precision required e.g. for grasping objects in cluttered scenes.

V. Gaussian proposal for $p(\omega|\hat{\mathbf{z}})$

The measurement model (8) can be used to find the weight in an importance sampling step for a given particle set. Since the sharpness of the measurement model is very high when numerous interest points have been found the convergence speed of the particle filter will degrade. A common solution to this is the usage of mixed proposals, where samples are also drawn from the current measurement and weighted according to the state transition model.

Since to our knowledge up to now, no method for sampling directly from the distribution of the state ω corresponding to a given measurement z is available, one way of generating samples from (8) for a given z and ω is to use importance sampling. The practicability here depends on the usage of a suitable proposal density.

Since the form of the true density depends on the measurement, no constant proposal seems feasible.

The basic idea is to use a Gaussian approximation as proposal which is derived from the measurements and their distribution given in (8).

Let us recall that the pose estimate ω is based on the actual measurement vector $\hat{\mathbf{z}} = \{\hat{\mathbf{z}}_1, \ldots..., \hat{\mathbf{z}}_N\}$. Now $\mathbf{x}$ is determined by minimizing the weighted back-projection errors of the model points with respect to ω. We use the distribution for the image error from (8), renaming the predicted image points $\mathbf{z}_{i,j}(\omega) =: \mathbf{z}_n(\omega)$ and $\Sigma_{i,j} =: \Sigma_n$

$$\mathbf{d}_n(\hat{\mathbf{z}}_n, \omega) := \mathbf{z}_n(\omega) - \hat{\mathbf{z}}_n \quad (9)$$

$$f_{err}(\hat{\mathbf{z}}, \omega) := \sum_{n=1}^{N} \mathbf{d}_n^T \cdot (\Sigma_n)^{-1} \cdot \mathbf{d}_n \quad (10)$$

In general, if there are enough point correspondences, this error function assumes a local minimum, i.e. the gradient g of the error function w.r.t. the pose ω is 0. The function

$$g(\hat{\mathbf{z}}, \omega) := \frac{\partial f_{err}}{\partial \omega}(\hat{\mathbf{z}}, \omega) = 0 \quad (11)$$

therefore provides an implicit definition of the non-linear function $f(\hat{\mathbf{z}}) = \omega$ that finds the optimal pose for a given measurement $\hat{\mathbf{z}}$. This function is not explicitly known, however it is implemented in the localization method described in Section III-B and can be evaluated for a given measurement vector.

We can now approximate the covariance matrix for the random variable ω as a non-linear function of the random variable $\hat{\mathbf{z}}$ by using the linear approximation, the Jacobian

$$J_f = \frac{\partial\, f}{\partial\, \hat{\mathbf{z}}}(\hat{\mathbf{z}})\,. \quad (12)$$

This Jacobian is calculated by applying the method of implicit differentiation to g. The partial derivatives for one measurement $\hat{\mathbf{z}}_\mathbf{n}$

$$J_{g,\omega,n} = \frac{\partial\, g}{\partial\, \omega}(\hat{\mathbf{z}}_\mathbf{n}, \omega) \quad , \quad J_{g,\hat{\mathbf{z}}_\mathbf{n}} = \frac{\partial\, g}{\partial\, \hat{\mathbf{z}}_\mathbf{n}}(\hat{\mathbf{z}}_\mathbf{n}, \omega) \quad (13)$$

are explicitly calculated from the closed form for h_j using a computer algebra system. These yield the Jacobians for the entire measurement vector $\hat{\mathbf{z}}$

$$J_{g,\omega} = \sum_{n=1}^{N} J_{g,\omega,n} \quad , \quad J_{g,\hat{\mathbf{z}}} = [J_{g,\hat{\mathbf{z}}_1} \cdots J_{g,\hat{\mathbf{z}}_N}] \quad (14)$$

Using these partial derivatives, we obtain

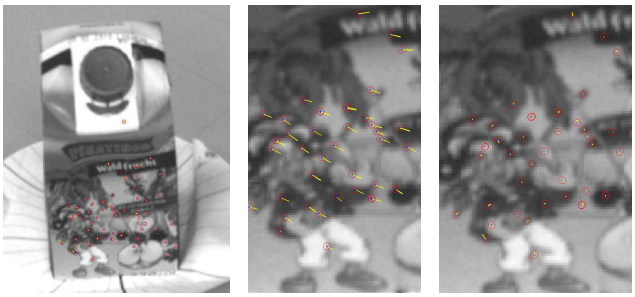
$$J_f = -J_{g,\omega}^{-1}\, J_{g,\hat{\mathbf{z}}} \quad (15)$$

With the $2\,N \times 2\,N$ block diagonal covariance matrix $\Sigma_{\hat{\mathbf{z}}}$ of the actual measurement $\hat{\mathbf{z}}$, this yields the desired Gaussian model for the pose ω:

$$p(\omega|\hat{\mathbf{z}}) \approx \mathcal{N}(f(\hat{\mathbf{z}}), J_f\, \Sigma_{\hat{\mathbf{z}}}\, J_f^T) \quad (16)$$

VI. Experimental validation

The vision system on our experimental robot (Fig. 1) consists of two AVT Pike F-145C cameras with a resolution of 1388×1038 pixels each, equipped with 8.5mm lenses and mounted with a disparity of roughly $0.12m$. Precise intrinsic and stereo calibration of the cameras is essential for our algorithms, so they were carried out with the Camera Calibration Toolbox for MATLAB, using about 60 stereo image pairs of a custom made highly planar checker board calibration pattern.

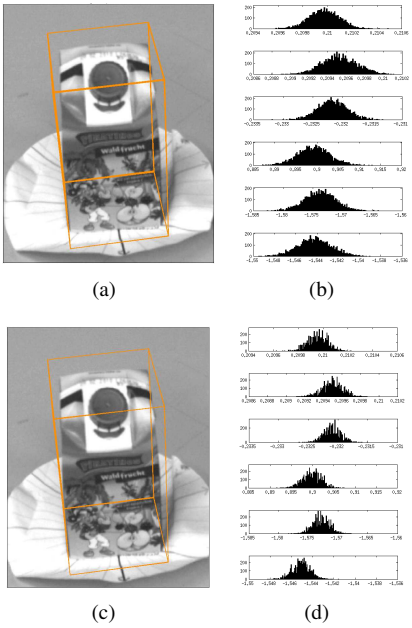


(a) (b) (c)

Fig. 5. An object from the database (a). The details show the result of the initial pose determination procedure (b) and the result of the optimization step (c).

For the following experimental evaluations, the Gaussian error in the SIFT feature point localization Σ_s was empirically determined to have a standard deviation of 1 pixel in both directions. Using the figures 5 to 8 we provide a walk-through of our method: The lower left part of Fig. 5 shows the result of the initial pose determination procedure. It depicts the result of step 5 in Section III-B. The short yellow lines depict the individual feature points reconstruction error, i.e. the distance between the feature points detected in this image and their locations predicted using the 3D object model from Section III-A, the initial pose estimate ω and the camera model in equation (4). The small ellipses (67%) represent the covariance matrix Σ^n_{ges} from (8). The lower right part of the figure shows the result of the optimization (step 7 in Section III-B). As can be seen, the reprojection error is significantly reduced by the optimization resulting in a much better pose estimate. This improvement is of importance, since the essential assumption behind (11) is, that the estimated pose minimizes the error function.

Based on the optimized pose estimate, we now construct the proposal density as described in Section V. The result is shown in Fig. 6 (a). Due to the high resolution images and the large number of feature points detected (70 points in the left and 84 in the right camera image), the pose estimate is obviously very accurate. The histograms of the right side show the marginal densities for all six degrees of freedom. Fig. 6 (b) shows samples drawn from the measurement model (8) obtained through importance sampling using the constructed proposal. The sensor model captures the true, non-Gaussian uncertainty more accurately than the proposal. However, the Gaussian proposal, while less peaked at least



(a) (b) (c) (d)

Fig. 6. (a) The constructed measurement proposal visualized by drawing samples from the model and plotting the corresponding bounding box of the object for each of the samples into the image (b) Samples drawn from the actual measurement model by means of importance sampling using the constructed proposal. The apparent misalignment of this special objects bounding box is due to some inaccuracies in its construction procedure. However, the plotted bounding boxes for the samples give a visual impression of the pose variance.

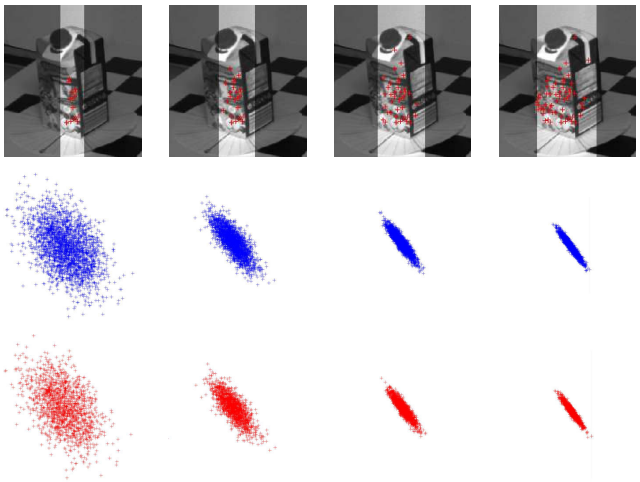
in this case, is very close to the true distribution and thus well suited for its purpose. The following experiment features a series of scenes where the object is artificially occluded. Thus, the amount of interest points that contribute to the pose estimate varies. In Fig. 7, four scenes are shown along with the corresponding proposal and the re-sampled distribution. The *effective sample size* S_{eff} [22], which measures the suitability of the proposal is given for all test-cases, S being the number of samples.

Finally Fig. 8 shows the evaluation of the sensor model for the complex scene of Fig. 1. Depending on the number and geometric configuration of the detected features, the accuracy of the pose reconstruction varies significantly.

VII. Summary

Interest point based methods have become widely popular in the field of object recognition and 6D localization. In general, their accuracy is high, however it varies strongly with the number and spatial distribution of the interest points used for pose reconstruction. Under realistic assumptions for actual applications this can and will be an issue for real systems, due to varying environmental conditions influencing sensor measurements and other real-world effects, like e.g. occlusions in cluttered scenes.

Our work aims at an overall perception architecture explicitly representing the actual uncertainties in a probabilistic framework. This enables a task-oriented assessment of the quality of the available knowledge and allows for the active refinement of the belief state of the system over time [5].



(a) $\frac{S_{eff}}{S} = 0.84$ (b) $\frac{S_{eff}}{S} = 0.79$ (c) $\frac{S_{eff}}{S} = 0.98$ (d) $\frac{S_{eff}}{S} = 0.70$

Fig. 7. Test with varying number of interest points. The artificial occlusion reduced the number of interest points that are used in the proposal (red). The second row shows the corresponding proposals (translational part). The third row shows the final distribution after weighting and re-sampling.

Fig. 8. Evaluated measurement models for a complex scene containing multiple objects.

Sensor models accurately modeling the actual measurement uncertainty are key in this context.

Besides our object modeling and pose reconstruction procedures, we presented such a measurement model for interest point based pose estimation, that explicitly models the actual sources of error. We used Gaussian models for the error in the 2D interest point localization as well as in the 3D models of the objects which are known to the system. In order to sample from the measurement model we adopted an importance sampling approach using a Gaussian proposal distribution constructed using the implicit function theorem. Our method enables us to realistically quantify the pose reconstruction uncertainty in all six degrees of freedom with high fidelity.

VIII. Acknowledgement

This work was made possible by funding from the ARTEMIS Joint Undertaking as part of the project R3-COP and from the German Federal Ministry of Education and Research (BMBF) under grant no. 01IS10004E.

References

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

[2] T. Grundmann, R. Eidenberger, M. Schneider, M. Fiegert, and G. Wichert v., "Robust high precision 6d pose determination in complex environments for robotic manipulation," in *ICRA 2010 Workshop: Best Practice in 3D Perception and Modeling for Mobile Manipulation*, 2010.

[3] T. Grundmann, R. Eidenberger, M. Schneider, and M. Fiegert, "Robust 6d pose determination in complex environments for one hundred classes," in *Proceedings of the 7th International Conference On Informatics in Control, Automation and Robotics*, 2010.

[4] Z. Xue, J. Marius Zoellner, and R. Dillmann, "Grasp planning: Find the contact points," in *IEEE International Conference on Robotics and Biomimetics*, 2007.

[5] R. Eidenberger, T. Grundmann, and R. Zoellner, "Probabilistic action planning for active scene modeling in continuous high-dimensional domains," *IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 2009*, 2009.

[6] T. Grundmann, M. Fiegert, and W. Burgard, "Probabilistic rule set joint state update as approximation to the full joint state estimation applied to multi object scene analysis," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010.

[7] S. Nayar, S. Nene, and H. Murase, "Real-time 100 object recognition system," in *Proc. IEEE International Conference on Robotics and Automation*, vol. 3, 22–28 April 1996, pp. 2321–2325.

[8] J. Zhang, R. Schmidt, and A. Knoll, "Appearance-based visual learning in a neuro-fuzzy model for fine-positioning of manipulators," in *ICRA*, 1999.

[9] J. A. Walter and B. Arnrich, "Gabor filters for object localization and robot grasping," in *ICPR*, 2000, pp. 4124–4127.

[10] D. Kragic, A. T. Miller, and P. K. Allen, "Real-time tracking meets online grasp planning," in *Proc. of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, Seoul, Republic of Korea, 2001, pp. 2460–2465.

[11] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision, Springer Netherlands*, vol. Volume 15, Numbers 1-2, pp. 123–141, 1995.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 1150–1157.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," vol. 110, no. 3. New York, NY, USA: Elsevier Science Inc., 2008, pp. 346–359.

[14] I. Gordon and D. G. Lowe, "What and where: 3d object recognition with accurate pose," *Toward Category-Level Object Recognition*, vol. Lecture Notes in Computer Science, pp. 67–82, 2006.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6d object localization for grasping with humanoid robot systems," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, San Diego, CA, USA, 2007.

[17] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *ICRA 09*, 2009.

[18] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based vs. monocular 6-dof pose estimation using point features: A quantitative comparison," in *Autonome Mobile Systeme 2009*, ser. Informatik aktuell. Karlsruhe: Springer, 2009.

[19] Z. Xue, A. Kasper, J. Zoellner, and R. Dillmann, "An automatic grasp planning system for service robots," in *14th International Conference on Advanced Robotics (ICAR)*, June 22nd - 26th, 2009.

[20] M. A. Lourakis and A. Argyros, "SBA: A Software Package for Generic Sparse Bundle Adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 1–30, 2009.

[21] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[22] J. S. Liu, "Metropolized independent sampling with comparisons to rejection sampling and importance sampling," *Statistics and Computing*, vol. 6, pp. 113–119, 1996.