Strategies for Multi-Modal Scene Exploration

Jeannette Bohg, Matthew Johnson-Roberson, Mårten Björkman and Danica Kragic

***Abstract*— We propose a method for multi-modal scene exploration where initial object hypothesis formed by active *visual* segmentation are confirmed and augmented through *haptic* exploration with a robotic arm. We update the current belief about the state of the map with the detection results and predict yet unknown parts of the map with a Gaussian Process. We show that through the integration of different sensor modalities, we achieve a more complete scene model. We also show that the prediction of the scene structure leads to a valid scene representation even if the map is not fully traversed. Furthermore, we propose different exploration strategies and evaluate them both in simulation and on our robotic platform.**

I. Introduction

The ability to interpret the environment, detect and manipulate objects is at the heart of autonomous robot systems, [1], [2]. These systems need to represent known and unknown objects for generating task-relevant actions. In this paper, we present strategies for autonomously exploring a scene containing unknown objects. Our robotic setup consists of a vision system that generates initial object hypotheses using active visual segmentation, [3], [4]. Thereby, large parts of the scene are explored in a few glances. However, without significantly changing the viewpoint, areas behind objects are occluded. Having a complete scene representation is essential for finding suitable grasps. To achieve that, parts of the scene that are not visible to the vision system are actively explored by the robot using a hand with tactile sensors. Compared to a gaze shift, moving the arm is expensive in terms of time and gain in information. Therefore, the next best measurement has to be determined to explore the unknown space efficiently.

As exploration strategies, we adapt two approaches from the area of mobile robotics. First, we use *Spanning Tree Coverage* (STC) that is optimal because every place in the scene is explored just once [5]. Secondly, we extend the approach presented recently in [6] where unexplored areas are predicted from sparse sensor measurements by a *Gaussian Process* (GP). Exploration then aims at confirming this prediction and reducing its uncertainty. The resulting scene model is multi-modal in the sense that it i) generates object hypotheses emerging from the integration of several visual-cues, and ii) fuses visual and haptic information. This model then forms the basis for *interactive perception* and a general symbol grounding problem [7].

This work was supported by EU through the project PACO-PLUS, IST-FP6-IP-027657, and GRASP, IST-FP7-IP-215821 and the Swedish Foundation for Strategic Research. The authors are with the Centre for Autonomous Systems and Computational Vision and Active Perception Lab, School of Computer Science and Communication, KTH, Stockholm, Sweden. bohg,mattjr,celle,danik@csc.kth.se

Fig. 1. Top Left: ARMAR III robot head. Top Right: Kuka arm with the Schunk hand. Middle Left: Peripheral view of a typical experimental scene. Middle Right: Foveal view of the same scene.

Our experimental platform includes the Armar III robotic head with a foveal and peripheral stereo camera. Attention is used in the peripheral view to direct fixation of the foveal cameras at regions of interest [8]. Object manipulation is done using a 6DoF Kuka arm equipped with a Schunk Dextrous Hand 2.0. Fig. 1 shows the hardware, an example view of each camera and a typical table top scene.

The contributions of this paper are i) strategies for active exploration of a predicted map, ii) a quantitative comparison with coverage based exploration and iii) a multi-modal scene representation that integrates data from a state-of-the-art vision system with haptic data.

II. Related Work

Interactive perception has gained considerable interest in the last years. In [9], [10], the robot pushes objects to gain more information about the objects or the scene. In that work the assumption is initial object hypotheses are given. The problem of choosing an action for forming these hypotheses is circumvented.

In this paper, we want to actively choose measurements from a latent function, in our case the scene, for approximating it. This is related to *active learning* in the field of machine learning. In [11], this is studied in an object classification task. Specific training examples are selected for querying their ground truth label to improve the estimated decision boundary. A common approach is to choose a measurement action that maximizes the expected information gain. In [12], such a method is applied to C-space exploration with a robotic arm. In [13], entropy is considered for reinforcement learning. Training data for learning the policy model is chosen based on an utility function that considers both the expected reward and expected information gain.

Algorithm 1: Pseudo Code for Scene Exploration

Data: Segmented point cloud $\mathcal{S}$ from active segmentation
Result: Fully explored $\mathcal{P}$
begin
  $t = 0,\ j = 0$
  $\mathcal{P} = \texttt{project}(\mathcal{S})$
  while $|\mathcal{P}_u| > 0$ **do**
    $\hat{\mathcal{P}} = \texttt{predict}(\mathcal{P})$
    $p_j = \texttt{planNextMeasurements}(\mathcal{P}, \hat{\mathcal{P}})$
    repeat
      $t++$
      $z_t = \texttt{observe}(\mathcal{P}, p_j)$
      $\mathcal{P} = \texttt{update}(\mathcal{P}, z_t)$
    until $z_t \neq occ$
    $j++$
  end
end

Such combined utility functions are also common in mobile robotics [14], [15], [16]. Here, the information gain of a specific point in the commonly used 2D map is traded off with the distance to travel there.

Other exploration strategies apply algorithms that systematically explore the space. We employ Spanning Tree Coverage [5] in which every grid cell is guaranteed to be measured only once. In this paper, we compare this exploration strategy with an information-theoretic approach. Specifically, we want to analyze how these strategies influence the quality of the scene estimation over the whole exploration process.

III. Scene Representation

As a scene representation, we choose a traditional 2D *occupancy grid* (OG) [17]. It is well suited for integrating measurements from different sources. The grid which is aligned with the table top, uniformly subdivides the scene into N cells C_i with coordinates (w_i, v_i). Each cell has a specific state $s(C_i)$. For simplicity, we will refer to it as s_i. It is defined over a binary random variable with two possible values: occupied (*occ*) or empty (*emp*). It holds that $P(s_i = occ) + P(s_i = emp) = 1$. We define $\mathcal{P} = \{C_i \mid 0 < i < N\}$, as the whole grid. Our goal is to estimate $P(s_i = occ \mid \{z\}_t)$, the probability for each cell C_i to be occupied given a set of sensor measurements $\{z\}_t$ up to point t in time. Let $\mathcal{P} = \mathcal{P}_k \cup \mathcal{P}_u$ where $\mathcal{P}_k$ is the set of cells whose state has already been estimated based on observations. $\mathcal{P}_u$ is the set of cells that has not been observed yet. Each cell is initialized with a prior probability $P(s_i = occ) = 0.5$. Our approach for scene exploration is summarized in Algorithm 1.

Initially, we `project` the stereo reconstructed point cloud $\mathcal{S}$ of the scene on the grid as follows. Disparity maps are gathered from several views of the robot head on the scene. They are converted into 3D points and projected into a common reference frame for all observations. Once aggregated, the whole point cloud is cleaned to remove outliers. The labeling from the 3D object segmentation (discussed in Section IV-A) is applied to the remaining points identifying objects from the background. These object points are placed into a voxel grid. This voxelized representation is projected down into a 2D occupancy grid $\mathcal{P}$ for planning. Fig. 2 displays this process and the resulting 2D map.

The initial $\mathcal{P}$ contains a lot of unknown space that needs to be explored with the hand. The single steps in the main loop of Algorithm 1 will be explained in the following sections.

IV. Scene Observation

A. Visual Observation

For 3D object segmentation we use a recent approach [4]. It relies on three possible hypotheses: figure, ground and a flat surface. It is assumed that most objects are placed on flat surfaces thereby simplifying segregation of the object from its supporting plane.

The segmentation approach is an iterative two-stage method that first performs pixel-wise labeling using a set of model parameters and then updates these parameters in the second stage. This is similar to Expectation-Maximization with the distinction that instead of enumerating over all combinations of labelings, model evidence is summed up on a per-pixel basis using marginal distributions of labels obtained using belief propagation.

The model parameters consists of the following three parts, corresponding to the foreground, background and flat surface hypothesis:

$$\theta_f = \{p_f, \Delta_f, c_f\},\ \theta_b = \{d_b, \Delta_b, c_b\},$$
$$\theta_s = \{\alpha_s, \beta_s, \delta_s, \Delta_s, c_s\},$$

p_f denotes the mean 3D position of the foreground. d_b is the mean disparity of the background, with the spatial coordinates assumed to be uniformly distributed. The surface disparities are assumed to be linearly dependent on the image coordinates, i.e. $d = \alpha_s x + \beta_s y + \delta_s$. All these spatial parameters are modeled as normal distributions, with Δ_f, Δ_b and Δ_s being the corresponding covariances. The last three parameters, c_f, c_b and c_s, are represented by color histograms expressed in hue and saturation space.

For initialisation, there has to be some prior assumption of what is likely to belong to the foreground. In our system, we have a fixating system and assume that points close to the center of fixation are most likely to be part of the foreground. For the flat surface hypothesis we apply RANSAC to find the most likely plane. The remaining points are initially labeled as background points.

B. Haptic Observation

For haptic exploration we use the two sensor matrices padding one finger of the robotic hand. Our goal is to decide when the hand is in contact with an object and when it is not. For this purpose, we compute a noise profile of the sensor matrices H_p and H_d. We model the distribution of these random variables as multivariate normal distributions $H_p \sim N(\mu_p, \Sigma_p)$ and $H_d \sim N(\mu_d, \Sigma_d)$ for which means and covariance matrices are computed from a number of non-contact measurements. A contact with an object can then be seen as a multivariate *outlier*. For outlier detection, we compute the Mahalanobis distance between the current measurement $h_{t|p,d}$ and the respective mean $\mu_{p,d}$.

$$d(h_t) = \sqrt{(h_t - \mu)^T \Sigma^{-1} (h_t - \mu)} \tag{1}$$

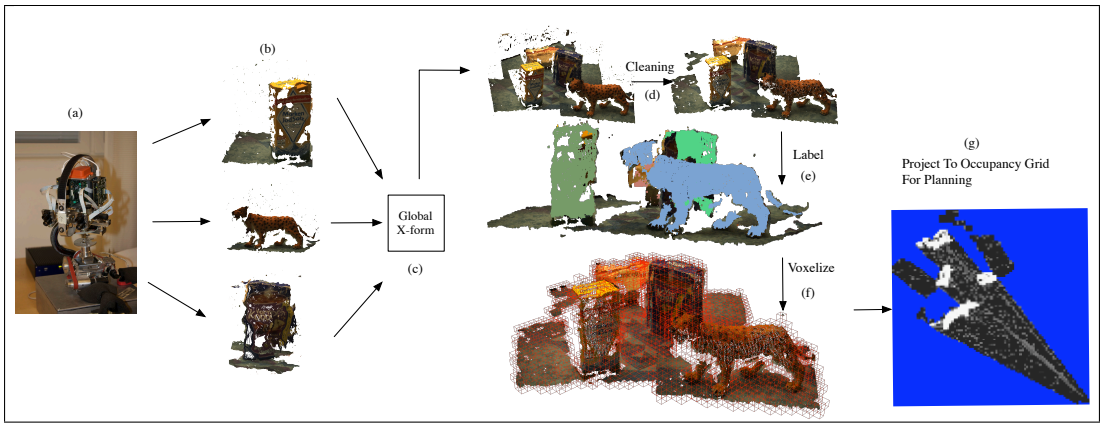


Fig. 2. Generation of an occupancy grid from individual views. (a) ARMAR robot head that (b) gathers several views. (c) Views are projected into a common reference frame and (d) cleaned to remove noise. (e) Points are labeled according to the 3D object segmentation (Sec.IV-A). (f) Scene is voxelized. The voxels that belong to objects are projected down into the map (g). Blue labels are unseen cells and gray levels correspond to occupancy probability.

Note that the subscripts p and d are skipped in this equation for simplicity. If $d(h_t)$ is greater than a threshold ϕ then $z_t = contact$, otherwise $z_t = \neg contact$.

V. Map Update

For each movement of the haptic sensors along the planned path, we are receiving a measurement z_{t+1}. Based on this and the current estimate $P(s_i = occ \mid \{z\}_t)$ of the state of each cell s_i in the occupancy grid, we want to estimate

$$P(s_i = occ \mid \{z\}_{t+1}) = \frac{P(z_{t+1} \mid s_i = occ)\ P(s_i = occ \mid \{z\}_t)}{\sum_{s_i} P(z_{t+1} \mid s_i) P(s_i \mid \{z\}_t)} \quad (2)$$

In this recursive formulation, the resulting new estimate $P(s_i = occ \mid \{z\}_{t+1})$ is stored in the occupancy grid. $P(z_{t+1} \mid s_i)$ constitutes the haptic sensor model. As described in the previous section, we model the haptic measurements as random variables with a multivariate Gaussian distributions. The case $s_i = emp$ is related to Eq. 1 as follows

$$P(h_{t+1} \mid s_i = emp) = \exp(-\frac{1}{2}\ d(h_{t_1})) = \exp(-\frac{1}{2}\sqrt{(h_{t+1} - \mu)^T \Sigma^{-1} (h_{t+1} - \mu)}) \quad (3)$$

VI. Map Prediction

In the traditional occupancy grid, cells that have not been observed yet will have a probability $P(s_i = occ \mid \{z\}_t) = 0.5$, i.e. there is no information available about the state of these cells. However, we know that cells in the grid that are close to occupied spaces but are due to occlusions not directly observable, are likely to be part of the occluding object. By modeling this spatial correlation, we can predict unobserved places from observed ones. Instead of exploring the whole environment exhaustively, we want to confirm the predicted map at specifically uncertain places. Recently, it was proposed that the spatial correlation in a 2D occupancy grid can be modeled with a Gaussian Process [6]. The assumption of independence of neighbouring cells in a traditional occupancy grids is removed.

A GP is used to fit a likelihood function to training data. In our case this is the set of cells $C_r \in \mathcal{P}_k$ and the estimate of their state s_r. Given the estimated continuous function over the occupancy grid, we can then estimate the state of the cells $C_j \in \mathcal{P}_u$ that have not been observed yet. We will briefly introduce GPs. For a more detailed explanation, we refer to [18]. A GP is defined as a collection of a finite number of random variables with a joint Gaussian distribution. In our case, the set of random variables is $C_i \in \mathcal{P}$. A GP can be seen as a distribution over functions with a mean μ and covariance Σ. Given a matrix of M already observed 2D grid cells $\mathbf{x} = \{C_r\}_M = \{(w_r, v_r)\}_M$ and a vector $\mathbf{y} = \{s_r\}_M$ of state labels, we want to query the state s_j of cell C_j then $f(C_j) \sim N(\mu, \Sigma)$ where

$$\mu = k(C_j, \mathbf{x})^T [K(\mathbf{x}, \mathbf{x}) + \sigma_M^2 I]^{-1} \mathbf{y} \quad (4)$$

$$\Sigma = k(C_j, C_j) - k(C_j, \mathbf{x})^T [K(\mathbf{x}, \mathbf{x}) + \sigma_M^2 I]^{-1} k(C_j, \mathbf{x}). \quad (5)$$

σ_M^2 is the variance of the noise on the target values. The entries of the covariance matrix $K(\mathbf{x}, \mathbf{x})_{u,v}$ at row u and column v are defined based on a covariance function $k(C_u, C_v)$ with some hyperparameters θ. We use the squared exponential covariance function

$$k(C_u, C_v) = \sigma_l^2 \exp(-((C_u - C_v)^T L^{-1} (C_u - C_v))/2) \quad (6)$$

where the hyperparameters are σ_l, the signal variance, and L, the identity matrix multiplied with the length scale l.

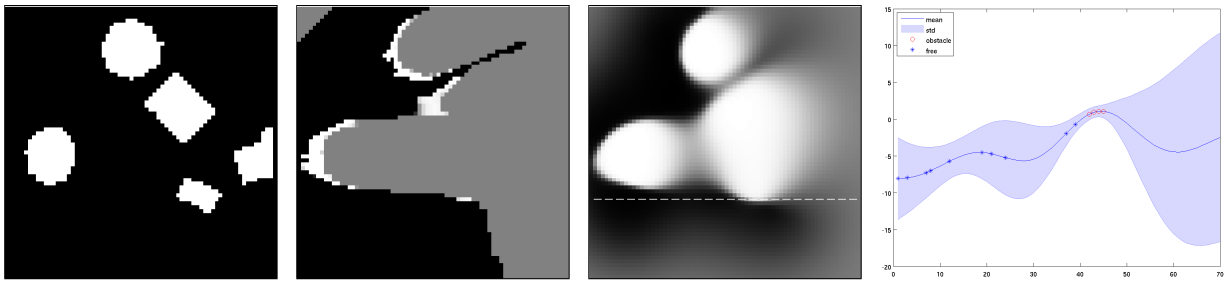
To compute $P(s_j = occ \mid \{z\}_t)$, we *squash* $f(C_j)$ through the cumulative Gaussian function.

$$P(s_j = occ \mid \{z\}_t) = 1/2 \cdot (1 + \mathrm{erf}(f(C_j)/\sqrt{2})) \quad (7)$$

An example for this prediction given a 2D map of partially explored scene is given in Fig. 3. In Sec. VIII, we will show quantitatively on synthetic data that a GP predicted map is a reasonable estimate of the ground truth.

VII. Action Selection for Exploration

Given a partial map of the environment, we want to efficiently explore the remaining unknown parts with haptic sensors, i.e., we want to minimize the amount of measurement



(a) Ground truth 2D occupancy grid

(b) Occupancy grid after camera Measurements taken from the left side.

(c) Likelihood of occupancy for each cell predicted by GP. 50th row labeled.

(d) Predictive mean and variance for 50th row of occupancy grid along with training points.

Fig. 3. Example for the prediction of a 2D map from camera measurements using GPs.

actions needed to reach a sufficient scene understanding. We will present two algorithms for planning a measurement path in the given map. First, we will use Spanning Tree Covering [5]. Second, we propose an active learning scheme based on the predicted map.

A. Spanning Tree Covering

STC tackles the *covering problem* that can be formulated as follows. Given the haptic sensor of size d and a planar work-area $\mathcal{P}_u$, the sensor has to be moved along a path such that every point in $\mathcal{P}_u$ is covered by it only once.

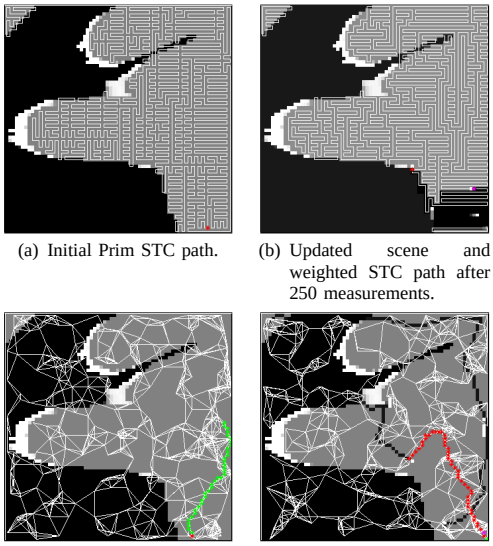
STC first defines a graph $G(V, E)$ on $\mathcal{P}_u$ with cells of size $2d$, the double tool size. In our case where $G(V, E)$ has uniform edge weights, Prim's algorithm can be used to construct a *Minimum Spanning Tree* (MST) that covers every vertex V in G at minimum cost regarding the edges E [19]. The haptic measurement path is defined on the original grid with cells of size d such that the MST is circumnavigated in counterclockwise direction. This circular path starts and ends at the current arm position. In case an obstacle is detected along the path, a new spanning tree has to be computed based on the updated grid. An example for such a path is shown in Fig. 4(a) and 4(b)

B. Active Learning

Our goal is to estimate the scene structure early in the whole exploration process without exhaustive observation. Thus, we want to support the map prediction by selecting most informative observations. Let us consider a set of measurements made along an MST as described above. These measurements will tend to be very close to each other without leaving any unobserved holes in the map. The GP prediction of the map based on these measurements will not be significantly different from the prediction based on only half of it. By using a GP and thereby exploiting spatial correlation in a map, the probability for a cell to be occupied can be inferred from its neighbors without explicitly observing it.

We will present exploration strategies that follow an active learning paradigm of selecting new measurement points that maximize the expected information gain. As it has been shown in [20], this is equivalent to minimizing the predictive variance Σ from Eq. 5.

$$C^* = \arg\max_{P_u} U_1(C_i) \text{ with } U_1(C_i) = \Sigma_i \tag{8}$$



(a) Initial Prim STC path.

(b) Updated scene and weighted STC path after 250 measurements.

(c) First measurement path along PRM.

(d) Updated scene and measurement path after 250 measurements.

Fig. 4. Examples for potential measurement paths generated with different exploration strategies. Red stars label current and previous traversed arm positions, respectively.

It can occur that the hand is detecting an obstacle along the chosen measurement path. It then has to re-plan and would potentially never reach the initially selected optimal observation point. Instead of considering the predictive variance only, the expected gain in information of the whole measurement path has to be taken into consideration. We propose two different utility functions that are dependent on both predictive variance and distance of a specific cell. The first one is

$$U_2(C_i) = \alpha\Sigma_i - (1-\alpha)\ d(C_s, C_i) \tag{9}$$

where Σ_i is the predictive variance of the cell C_i and $d(C_s, C_i)$ is any distance function of the current position of the arm C_s and cell C_i. The parameter $0 < \alpha < 1$ is user determined. The closer it is to 1 the more important becomes the value of the predictive variance.

The second utility function uses a discount factor δ.

$$U_3(C_i) = \sum_{r=1}^{R} \delta^r \Sigma_{p(r)} \tag{10}$$

where R is the number of measurement that are needed to reach the final cell C_i along the path $p = [C_{s+1} \dots C_i]$. Not

just the final cell C_i is considered. Instead, the predictive variance of all the cells along the path contributes to the utility value of C_i. The parameter δ has to be chosen by the user. It steers how steep the decrease of influence of the cells in the path are dependent on their distance from the current arm position.

To find the global maxima of Eq. 8 we have to maximize over all cells C_i in $\mathcal{P}_u$ and over all the paths through which a cell can be reached from C_s. Since this is a prohibitive number of possibilities to compute, we use sampling techniques to find a local maxima of the utility functions. We are building a *probabilistic road map* (PRM) in the two dimensional C-space of the occupancy grid [21].

A set T of cells from $\mathcal{P}$ is sampled according to their predictive variance and connected to the PRM. The *Dijkstra* algorithm is used to compute the shortest path from the current arm position to each cell in T through the PRM. The result is used to compute the utility of each cell in T. An example for the PRM and therefore the possible paths to traverse is shown in Fig. 4(c) and 4(d).

VIII. Experiments

We evaluate the proposed exploration strategies quantitatively on synthetic data and demonstrate their feasibility in a real-world scenario.

A. Synthetic Data

1) Data Set and Measure of Comparison: We generated 50 different 2D occupancy grids (70×70 cells) an example appears in Fig. 3(a). Every scene contains ten objects that can either be of circular, elliptical or rectangular shape with a random size, aspect ratio, orientation and position. Overlappings are allowed so that fewer than ten connected components can occur as well as more complex contours.

For every scene, we simulated three camera observations made from a fixed position on their left side with a random direction. As a sensor model, we used a beam model with a Gaussian profile [17]. Given a measurement, the occupancy grid gets updated according to Eq. 2. An example for the result of this simulation is shown in Fig. 3(b).

We posed the validation problem of occupancy grid estimation as a binary classification into empty or occupied cells. For each estimated grid at any time in the exploration process, the number of false and true positives can be computed for different thresholds resulting in an ROC curve. For evaluating the development of this curve over time, we choose the area under the ROC curve (AUC) as a measure. It corresponds to the probability that the state of a cell is correctly classified.

2) Predicting Occupancy Grids with GPs:

Covariance Functions Compared: In [6], it is claimed that the neural network covariance function is more suitable for predicting the non-stationary behavior of a typical map data set. That data comes from indoor and outdoor environments either with hallways, rooms, walls or streets bounded by buildings. However, our experimental data showed superior performance with the squared exponential covariance function in which blob-like objects spread on a table were more correctly modeled.

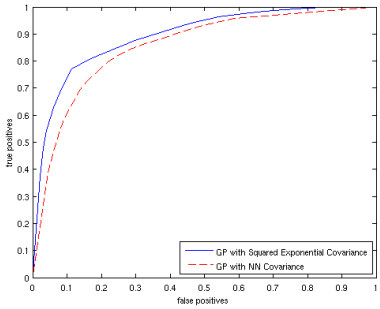


Fig. 6. ROC for Squared Exponential (SE) and Neural Network (NN) covariance function. SE outperforms NN for OG prediction.

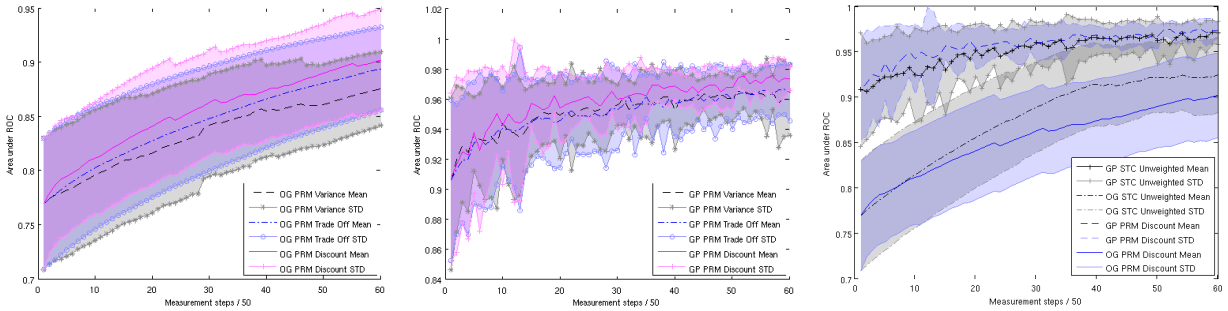
We predicted each of the 50 occupancy grids with a GP by sampling training points from the space observed by the camera and then querying $p(s_i = occ \mid \{z\}_t)$ for $C_i \in \mathcal{P}_u$. We compared a neural network with a squared exponential covariance function. The ROC curves for the whole data set are shown in Fig. 6 from which the improved performance of the squared exponential covariance function for our scenario is confirmed.

Prediction vs Occupancy Grid: An important question is whether the GP makes a valid prediction of the scene map and how this compares to the traditional occupancy grid. In an OG, only those estimates for $p(s_i = occ \mid \{z\}_t)$ are different from the initial value of 0.5 for which at least one measurement has been obtained. However, the GP predicts occupancy probability for all cells. To confirm that this inference is valid, we calculated the mean AUC for all occupancy grids after they have been observed by the camera and compared it to the mean AUC of the GP predicted maps. While for the unpredicted occupancy grid this value is 0.7697, it is 0.9058 for the predicted maps; a clear increase of 13%. Therefore we conclude in agreement with [6] that a GP prediction provides valid inference about regions of the scene in which no measurements are available.

3) Exploration Strategies Compared: We compare the different exploration strategies based on the mean and variance of the AUC measure over time and all synthetic scenes. We start from the scenes partially explored with the camera.

First, we consider the utility based exploration (Sec. VII-B). Three functions were proposed that incorporate uncertainty in the map prediction and/or distance to traverse. Fig. 5(a) and 5(b) show the results for the OGs and for the GP predicted maps. There is a clear difference between the utility functions. The discounted version (Eq. 10) performs best both in terms of mean and variance in the OGs and GP predicted maps. This is due to the explicit consideration of the whole exploration path instead of just a high predictive variance goal point that might never be reached.

Second, we consider STC based exploration (Sec. VII-A). Measurement paths are planned such that each $C_i \in \mathcal{P}_u$ is traversed only once. Fig. 5(c) shows the results for the occupancy grids and for the GP predicted maps. The estimation of the occupancy grid converges relatively fast towards ground truth.



(a) Utility based exploration of OG (b) Prediction with GP during utility based exploration (c) Best STC and utility based exploration compared

Fig. 5. Mean and variance of area under ROC curve (AUC) for occupancy grid (OG) as well as Gaussian Process (GP) prediction under different exploration strategies. (a), (b): Utility based exploration. Discounted predictive variance outperforms pure variance and trading off between distance and variance value. (c) Comparison between Spanning Tree Covering (STC) based method and best utility based method. The latter achieves a more accurate GP prediction early in the process, while STC explores the unknown space in the OG faster.

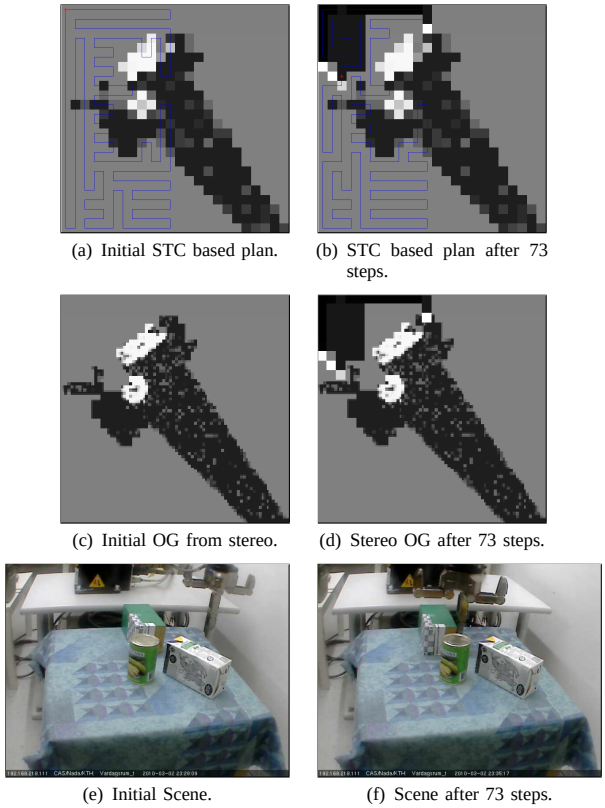
4) Summary: Fig. 5(c) also contains the results for the discounted utility function for direct comparison with the STC based exploration. GP predicted maps traversed based on the discounted utility are more accurate early in the exploration process. This is an expected result since here points of high variance are chosen to be measured first. They will therefore have a high positive influence on the quality of the prediction. However, the occupancy grid does not converge as fast towards ground truth as in the STC based exploration. We conclude that if a good map estimate is needed quickly, active learning based exploration is advantageous over systematically traversing the space. If there is time for an exhaustive exploration, STC based measurement paths are more beneficial.

B. Demonstration in the Real World

In this section, we demonstrate our approach for the real-world scenario described in Sec. I: exploration of a table top populated with several unknown objects using both, point clouds coming from a stereo camera and haptic data from a robotic hand. As exploration strategy we consider STC (see Sec. VII-A) and a PRM based scheme with the discounted utility function defined in Eq. 10.

The real world scene contains three objects. For visual exploration, we manually select the initial fixation point for two of them. This could be replaced by using an attention system as described in [8]. The objects are segmented and stereo reconstructed (Sec. IV-A). The resulting point cloud is projected onto an occupancy grid aligned with the table (Sec. III). By excluding the third object from visual observation, we can demonstrate the map update upon finger contact. Examples for the occupancy grid are given in Fig. 7(c) and 8(b).

In this stereo based grid, we can now detect locations that were not observed with the vision system. The reachable unexplored spaces are explored with the haptic sensors on the hand. For doing so, we are using one of the three fingers pointing downwards as shown in Fig. 7(e). The hand is moved at a constant height over the table. The haptic sensor arrays on the finger are always pointing in the direction of movement. Planning is done in a slice of the robot task space that is aligned with the table top. This is a reasonable simplification since all the points on the table within a radius



(a) Initial STC based plan. (b) STC based plan after 73 steps.

(c) Initial OG from stereo. (d) Stereo OG after 73 steps.

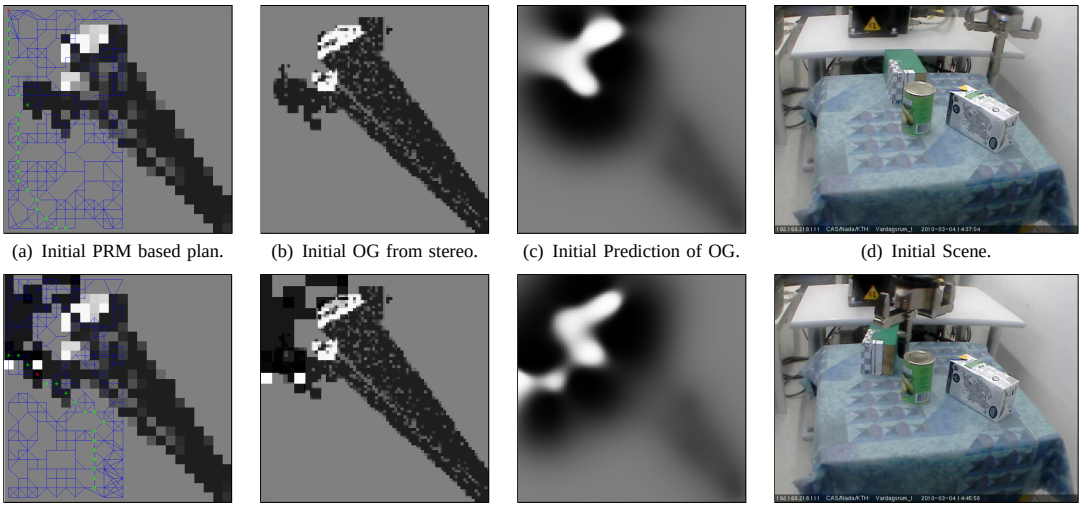
(e) Initial Scene. (f) Scene after 73 steps.

Fig. 7. Snapshots from an exploration using an STC based plan (covers only reachable workspace).

of 780mm are reachable with a valid joint configuration. For more complex environments planning has to be done in the six-dimensional C-space of the arm which is considered as future work.

Since the fingers of the Schunk hand are very thick, we are keeping two grids in parallel: the coarse planning grid and the finer grid from the stereo data. For every measurement, both grids are updated in parallel: one cell in the planning grid, a set of cells in the stereo grid.

In Fig. 7, the STC based measurement path planned on the initial occupancy grid is shown as well as the updated grid after 73 measurement steps. As expected from the results on the synthetic data, the area close to the starting position of the hand at the top left corner is explored systematically without leaving holes. In Fig. 8, the first PRM based mea-



(a) Initial PRM based plan. (b) Initial OG from stereo. (c) Initial Prediction of OG. (d) Initial Scene.

(e) PRM based plan after 68 steps. (f) Stereo OG after 68 steps. (g) Prediction after 68 steps.. (h) Scene after 68 steps.

Fig. 8. Snapshots from an exploration using an PRM based plan (covers only reachable workspace).

surement path is shown as well as the updated grids after 68 measurements. The area close to the start position is not yet fully explored, but the next measurement path leads towards the lower left of the grid that has a high uncertainty. A movie of the demonstration is available at [22].

Opposed to the synthetic experiments, in the real world, objects can move upon contact with the hand. This can be observed when comparing Fig. 7(e) and 7(f) or 8(d) and 8(h). To avoid the map becoming inconsistent, visual tracking is needed.

IX. Conclusion

We proposed a method for multi-modal scene exploration. Initial object hypotheses formed by active visual segmentation were confirmed and augmented through haptic exploration. The current belief about the state of the map is updated with measurements and yet unknown parts of the map are predicted with a Gaussian Process. Through the integration of different sensor modalities, we achieved a more complete scene model. We showed that the prediction of the scene structure leads to a valid scene representation even if the map is not fully traversed. Furthermore, different exploration strategies were proposed and evaluated quantitatively on synthetic data. Finally, we showed the feasibility of our scene representation and exploration strategies in a real world scenario. The demonstration on the robot also exposed further challenges. Constant visual tracking of the scene during the hand interaction is necessary to keep the scene estimate up to date. This is considered as future work. Furthermore, we aim to generalise the approach to 3D.

References

[1] L. Petersson, D. Austin, and D. Kragic, "High-level control of a mobile manipulator for door opening," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, vol. 3, Oct. 2000, pp. 2333–2338.

[2] S. Ekvall, P. Jensfelt, and D. Kragic, "Integrating active mobile robot object recognition and slam in natural environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2006, pp. 5792–5797.

[3] D. Kragic, M. Björkman, H. I. Christensen, and J.-O. Eklundh, "Vision for robotic object manipulation is domestic settings," *Robotics and Autonomous Systems*, pp. 85–100, Jun 2005.

[4] M. Björkman and D. Kragic, "Active 3d scene segmentation and detection of unknown objects," in *IEEE Int. Conf. on Robotics and Automation*, 2010.

[5] Y. Gabriely and E. Rimon, "Spanning-tree based coverage of continuous areas by a mobile robot," *Annals of Mathematics and Artificial Intelligence*, vol. 31, pp. 77–98, May 2001.

[6] S. T. O'Callaghan, F. Ramos, and H. Durrant-Whyte, "Contextual occupancy maps using gaussian processes," in *IEEE Int. Conf. on Robotics and Automation*, Kobe, Japan, May 2009.

[7] S. Harnad, "The symbol grounding problem," in *PhysicaD: Nonlinear phenomena*, vol. 42, 1990, pp. 335–346.

[8] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic, "An Active Vision System for Detecting, Fixating and Manipulating Objects in Real World," *Int. J. of Robotics Research*, vol. 29, no. 2-3, 2010.

[9] D. Katz and O. Brock, "Manipulating articulated objects with interactive perception," in *IEEE Int. Conf. on Robotics and Automation*, May 2008, pp. 272–277.

[10] J. Kenney, T. Buckley, and O. Brock, "Interactive Segmentation for Manipulation in Unstructured Environments," in *IEEE Int. Conf. on Robotics and Automation*, 2009, pp. 1377–1382.

[11] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with gaussian processes for object categorization." in *Int. Conf. of Computer Vision*. IEEE, 2007, pp. 1–8.

[12] Y. Yu and K. K. Gupta, "C-space entropy: A measure for view planning and exploration for general robot-sensor systems in unknown environments," *I. J. Robotic Res.*, vol. 23, no. 12, pp. 1197–1223, 2004.

[13] M. P. Deisenroth, C. E. Rasmussen, and J. Peters, "Gaussian process dynamic programming," *Neurocomput.*, vol. 72, no. 7-9, pp. 1508–1524, 2009.

[14] F. Bourgault, A. A. Makarenko, S. B. Williams, B. Grocholsky, and H. F. Durrant-Whyte, "Information based adaptive robotic exploration," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2002, pp. 540–545.

[15] G. Lidoris, D. Wollherr, and M. Buss, "Bayesian state estimation and behaviour selection for autonomous robotic exploration in dynamic environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, San Diego, USA, October 2007.

[16] C. Stachniss, G. Grisetti, and W. Burgard, "Information gain-based exploration using rao-blackwellized particle filters," in *Proc. of Robotics: Science and Systems (RSS)*, Cambridge, MA, USA, 2005.

[17] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, pp. 46–57, 1989.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, December 2005.

[19] R. Prim, "Shortest connection networks and some generalisation," *Bell System Technical Journal*, no. 36, pp. 1389–1401, 1957.

[20] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical Science*, vol. 10, pp. 273–304, 1995.

[21] L. E. Kavraki, P. Svestka, J. C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 566–580, 1996.

[22] J. Bohg, M. Johnson-Roberson, M. Björkman, and D. Kragic, "Strategies for multi-model scene exploration," Movie, http://www.csc.kth.se/ bohg/IROS2010Grasp.mp4.