Approximation of joint information gain for multi-sensor volumetric scene reconstruction

Mikko Lauri[1], Joni Pajarinen[2,3], Jan Peters[2], Simone Frintrop[1]

***Abstract*— Scene reconstruction builds a model of a real-world scene using images of the scene. In active scene reconstruction, usually a single robot moves a camera to different points of view to reconstruct the scene in high quality. In contrast to single robot scene reconstruction, we plan where each robot in a team should move to maximize scene reconstruction quality. We formulate this next-best-view planning problem as a centralized submodular optimization problem, leading to performance guarantees for the approximate greedy policy. We propose an approximation for the expected information gain from a set of sensor poses that takes into account coordination between the sensors, and apply the greedy policy to scene reconstruction tasks with two robot arms. Initial experimental results show that successful scene reconstruction can be achieved using the proposed approach. We identify several directions for future work that can be helpful to more precisely characterize the benefits and limitations of our method.**

I. Introduction

Scene reconstruction creates a digital model of a real-world scene from images of the scene. Using scene reconstruction, a robot can create a model of its workspace and the objects therein. Robotic applications requiring scene reconstruction range from household tasks involving object manipulation to industrial applications such as sorting waste.

To collect the images required for scene reconstruction, a robot views the scene from multiple viewpoints. Planning which views provide most information can increase scene reconstruction efficiency by avoiding exhaustively visiting all views. The problem of planning views to create high-resolution scene reconstructions is known as the next-best view (NBV) planning problem. To represent the scene to be reconstructed, NBV planning approaches often employ a volumetric [1] or surface based representation [2]. The next possible views are scored based on the distance from the current view, reachability, and overlap with previous views [3], or various quantifications of volumetric information [4], [1]. The next best view selected is the one with the greatest score.

In this paper, we address centralized multi-robot NBV planning, where a team of robots equipped with depth cameras reconstruct a scene. As our solution is centralized, it is applicable to problems where communication between the team members is available, for example when the robots work within a fixed work cell in close proximity. Fig. 1 shows an illustration of the task. We formulate a next-best-view planning problem as a centralized submodular optimization problem, and obtain performance guarantees for the approximate greedy policy. We propose an approximation for the expected information gain from a set of sensor poses that takes into account coordination between the sensors. We then apply the greedy policy of maximizing the expected information gain to scene reconstruction using two robot arms, and show that the method successfully explores the scene. We identify directions for future work and discuss potential extensions.

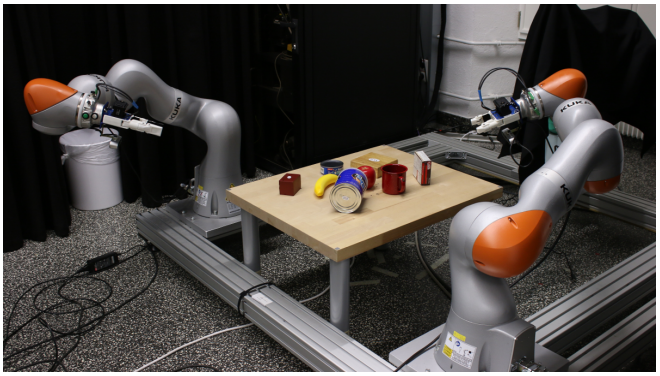

Fig. 1. Intel Realsense D435 depth cameras are attached to two KUKA LBR iiwa robot arms on the left and right. The robots move to poses around the table in the middle and record images of the items on the table. The objective is to plan a sequence of poses from which to view the items such that the quality of the resulting volumetric scene reconstruction is maximized.

The remainder of the paper is organized as follows. In Section II, we review related work in NBV planning, and multi-camera and multi-robot information gathering. In Section III, we formulate centralized multi-sensor scene reconstruction as a sensor selection problem and show that the problem is submodular. In Section IV, we introduce our proposed approximation for the information gain that encourages coordination between sensors. We present initial experimental results in Section V. Section VI concludes the paper by discussing directions for future work.

II. Related work

The NBV planning problem is a type of active vision problem [5], [6] where a camera is controlled to gather information to solve a task. For a general survey of active vision in robotic systems, see [7].

Single camera scene reconstruction. Delmerico et al. [1] compare various quantifications of information gain in a volumetric scene reconstruction task. They propose an occlusion-aware quantification of information gain that considers the visibility likelihood of voxels when planning next

[1]Department of Informatics, University of Hamburg, Hamburg, Germany {lauri,frintrop}@informatik.uni-hamburg.de
[2]IAS lab, TU Darmstadt, Darmstadt, Germany {pajarinen,peters}@ias.tu-darmstadt.de
[3]Tampere University, Tampere, Finland

views. The average entropy approach proposed in [4] selects views that have the greatest sum of entropies in the voxels potentially visible from that view. The area factor method proposed in [8] selects views that contain both occupied voxels and voxels on the frontier between known and unknown space. Robot control uncertainty is additionally considered in [3]. In [9], the related problem of planning how a robot should manipulate an object within the view of a stationary camera to create a 3D model of the object is investigated. The next pose to be selected is determined by maximizing expected information gain. All of the aforementioned works apply a volumetric scene representation. Recently, [2] proposed an NBV planning method based on a surface density representation that scales to large scenes of up to 40 meter scale. Views are selected to maximize the expected observed frontier volume. Other NBV planning applications using a surface representation include, e.g., [10].

Most of the works referenced above solve the next best view problem by a greedy strategy. The greedy strategy selects the view with the greatest immediate expected utility, without regard for utility over a longer horizon of time. Advantages of planning over multiple decisions non-greedily have been demonstrated, e.g., in active object detection [11], robotic exploration [12], [13], object manipulation [14] and search [15], and object classification in outdoor environments [16]. Information-theoretic exploration techniques can also be useful over long time horizons in changing environments [17]. However, in information gain maximization tasks with a submodular objective function, the greedy policy provides a good approximation to an optimal solution [18]. We use submodularity of our problem to justify the use of an approximate greedy policy.

Multi-camera scene reconstruction. Planning and control in multi-camera systems mainly concentrates on the two tasks of camera network design and control. The design task involves planning the placement of stationary cameras, for example for maximum coverage [19] or best 3D reconstruction performance [20]. The control task is mainly concerned with control for best possible area coverage in surveillance and security applications, see [21] for a survey. A generic method for online control of a camera network for a variety of computer vision tasks via policy search combined with heuristics is proposed in [22], but no quantitative evaluation of scene reconstruction is provided.

The problem of multi-robot information gathering is closely related to multi-camera control. Various decision-theoretic approaches to find decentralized solutions to multi-robot active information gathering problems have been proposed [23], [24], [25], [26]. However, finding such decentralized policies is computationally intensive.

Unlike the scene reconstruction methods reviewed above, we address multi-sensor scene reconstruction, where a team of robots or sensors co-operates to find the reconstruction. We target a team of robots that work in close proximity and connected via a low-latency network to each other, e.g., robots sorting waste along a processing line. Our approach is similar to [27] who consider the multi-robot setting, but differs as we do not consider specific regions of interest, but aim to obtain an overall coverage of the entire scene. Further, planning in [27] is decentralized, whereas we consider the centralized case where the robots in the team share their observations.

III. Multi-sensor active volumetric reconstruction

In this section, we formulate the problem of active volumetric reconstruction as a sensor selection problem. We show that our optimization problem is submodular, which leads to performance guarantees on the approximate solution by a greedy policy. Finally, we present an approximation of information gain that takes into account coordination between the sensors.

Notation. Random variables and sets are denoted by uppercase letters such as X. Lowercase letters, e.g., x, denote variables and realizations of random variables. A sequence of t objects is written $x_{1:t}$, and tuples of n objects as $(x^i)_{i=1}^n$.

A. Problem formulation

Consider a team of $n \geq 1$ sensors or robots as shown in Fig. 1. The robots construct a volumetric reconstruction of their workspace, in Fig. 1 the tabletop, by sequentially selecting sensor poses from which to capture depth images. The depth images captured by the sensors are combined into a single joint reconstruction that describes the scene geometry. Each pose of sensor $i = 1, 2, \ldots, n$ can be chosen from the set S_i. We denote the set of possible sensor poses for all sensors obtained as a Cartesian product $S = \times_{i=1}^n S_i$. As we target applications with robotic manipulators with high pose accuracy, we assume the sensor poses to be known.

We partition the workspace into a finite three-dimensional voxel grid V. For each voxel $v \in V$ a binary random variable X_v describes whether the voxel is free or occupied, that is, $X_v = 0$ or $X_v = 1$, respectively. We assume these random variables to be independent, and model information about the scene reconstruction as a collection of probabilities $p_v := P(X_v = 1)$ for each voxel $v \in V$. The uncertainty in a reconstruction $X = \{X_v \mid v \in V\}$ may be quantified by calculating its Shannon entropy

$$H(X) = \sum_{v \in V} H(X_v), \tag{1}$$

where $H(X_v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$ is the entropy of the binary random variable X_v.

A random variable Y_{s^i} depicts the depth image recorded by sensor i at pose $s^i \in S_i$. We write $Y_s = \{Y_{s^i} \mid s^i, 1 \leq i \leq n\}$ to denote the collection of random variables depicting depth images recorded by all sensors. We assume the depth images are mutually independent given the workspace occupancy[1], i.e., $P(Y_s \mid X) = \prod_{i=1}^n P(Y_{s^i} \mid X)$.

Suppose sensor poses $s = (s^i)_{i=1}^n$ are selected, and depth images $y = (y^i)_{i=1}^n$ are observed. Bayes' rule is applied to

[1]One requirement for this assumption is that no sensor can be at a pose where it occludes the view of another sensor.

find the posterior probability $P(X \mid Y_s = y)$. In this work we apply probabilistic sensor fusion on occupancy grids [28]. The following two steps are repeated for each depth image y^i. First, the 3D points corresponding to y^i are projected onto the voxel grid through ray casting. Second, for each voxel v intersected by the rays cast we compute the posterior probability $P(X_v \mid Y_{s^i} = y^i)$ applying the sensor model $P(Y_{s^i} \mid X)$ in [29]. The results produced by this sensor model are independent of the processing order of the images.

In the multi-sensor active volumetric reconstruction problem, we are given a budget of $T \geq 1$ sensor poses. Our task is to design a policy to select T sensor poses that maximizes a performance measure. Mutual information (MI) [30] is often used as a performance metric in such information gathering problems, see, e.g., [31], [32]. We are interested in maximizing the MI $I(X; Y_{s_1}, Y_{s_2}, \ldots, Y_{s_T}) := I(X; Y_{s_{1:T}})$ between the reconstruction X and the depth images recorded in the T sensor poses. A closed-loop policy uses all information from the preceding $(t-1)$ sensor poses and the corresponding depth images to select the tth sensor poses. We denote by $h_t = (s_1, y_1, s_2, y_2, \ldots, s_{t-1}, y_{t-1})$ the history information that can be applied to select the tth sensor pose[2].

Problem 1 (Multi-sensor active volumetric reconstruction)**.** *Given $n \geq 1$ sensors and the sets S_i of available poses, a budget of $T \geq 1$ poses, prior information $P(X)$, and a sensor model $P(Y_s \mid X)$, design a sequence of optimal closed-loop policies $\mu^*_{1:T}$ such that*

$$\begin{aligned} \mu^*_{1:T} &= \operatorname*{argmax}_{\mu_{1:T}} I(X; Y_{s_{1:T}}) \\ \text{s.t. } s_t &= \mu_t(h_t),\ t = 1, \ldots, T \\ Y_{s_t} &\sim P(Y_{s_t} \mid X),\ t = 1, \ldots, T \\ h_t &= (s_1, y_1, \ldots, s_{t-1}, y_{t-1}),\ t = 1, \ldots, T. \end{aligned} \tag{2}$$

We remark that the solution of Problem 1 is centralized, as the selection policy depends on the shared history of observations of all sensors.

B. Analysis of closed-loop greedy sensor selection policy

Problem 1 is equivalent to a finite-horizon partially observable Markov decision process problem, which are known to be computationally intractable [33]. We propose an approximate closed loop greedy policy that maximizes expected immediate information gain. We show Problem 1 is submodular, which gives performance bounds on the greedy policy.

The closed loop greedy policy selects the next sensor poses that maximize the expected immediate information gain. Formally, the tth sensor pose is selected according to

$$\mu^g_t(h_t) = \operatorname*{argmax}_{s_t \in S} I(X; Y_{s_t} \mid Y_{s_1} = y_1, \ldots, Y_{s_{t-1}} = y_{t-1}). \tag{3}$$

Since the MI of any two random variables A and B is equal to $I(A; B) = H(A) - H(A \mid B)$ where $H(A \mid B)$ is the conditional entropy, the closed loop greedy policy is equivalent to maximizing $H(X' \mid Y_{s_t})$, where $X' = X \mid Y_{s_1} = y_1, \ldots, Y_{s_{t-1}} = y_{t-1}$.

[2]With $h_1 = \emptyset$.

Submodularity [18] is a diminishing returns property of set functions that can provide performance bounds on various greedy selection policies.

Definition 1 (Submodularity)**.** *A function $f : 2^U \to \mathbb{R}$ is submodular if for every $A \subseteq B \subseteq U$ and $m \in U \setminus B$, $f(A \cup \{m\}) - f(A) \geq f(B \cup \{m\}) - f(B)$. Additionally, f is monotone on U if $f(A) \leq f(B)$.*

We view Problem 1 as a maximization of a set function. The following proposition shows that information gain in our problem is submodular and monotone. The proof is by Corollary 4 of [34].

Proposition 1 (Submodularity of information gain [34])**.** *Let $Y = \{Y_s \mid s \in S\}$, such that the variables in Y are independent given X. Then the information gain $I(X; A)$, where $A \subseteq Y$, is submodular and monotone on Y.*

Next consider the open loop variant of Problem 1, where we select a sensor pose sequence *without* observing the depth images or adapting our subsequent choices based on them. In other words, we want to find an optimal sequence $s^*_{1:T} = \operatorname*{argmax}_{s_{1:T}} I(X; Y_{s_{1:T}})$ of sensor poses such that the expected information gain is maximized. The expected information gain of the closed loop greedy policy (Eq. (3)) is related to an optimal open loop policy by the following theorem, whose proof is found in Thm. 5 of [32].

Theorem 1 ([32])**.** *Let $I(X; Y_{\mu^g})$ denote the expected value of the information gain $I(X; Y_{s_{1:T}})$ when the sensor poses $s_{1:T}$ are selected according to the closed loop greedy policy μ^g_t. Under conditions of Proposition 1, $I(X; Y_{\mu^g}) \geq \frac{1}{2} I(X; Y_{s^*_{1:T}})$, where $I(X; Y_{s^*_{1:T}})$ is the expected information gain of an optimal open-loop sequence $s^*_{1:T}$.*

Our analysis above shows that our suggested closed loop greedy policy from Eq. (3) yields at least one half of the expected information gain of an optimal open loop policy.

IV. Single- and multi-sensor planning

We consider two techniques of planning sensor poses to solve Problem 1. Both techniques approximate the expected information gain by calculating the volumetric information (VI) [1] contained in the visible part of the workspace given a set of sensor poses. In Subsection IV-A we review how to estimate volumetric information. In Subsection IV-B, we define an approximate policy that seeks to maximize the expected information gain of sensor poses by planning them individually. In Subsection IV-C we introduce a modification of the expected information calculation that takes into account coordination between multiple sensors.

A. Volumetric information of a voxel

We evaluate the VI of a voxel X_v given a sensor pose s^i via ray tracing. We denote by $V_{s^i} \subset V$ the subset of voxels in the visible part of the workspace, i.e., those that are potentially traversed by the rays from sensor i. To evaluate the total VI, we cast a set of rays from s^i and accumulate VI of each voxel along each ray.

We consider two ways of estimating volumetric information. The first method is the visible entropy (VE) criterion [4]

$$\mathrm{VE}(X_v, s^i) = \begin{cases} H(X_v) & \text{if } v \in V_{s^i} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

that directly estimates VI of a voxel by its entropy. Alternatively, we consider the occlusion-aware (OA) criterion [1]

$$\mathrm{OA}(X_v, s^i) = \begin{cases} p_{\mathrm{vis}}(v) H(X_v) & \text{if } v \in V_{s^i} \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $p_{\mathrm{vis}}(v)$ is the probability that voxel v is visible from the current pose, computed as the product of occupancy probabilities of all voxels along the ray passing through v.

B. Single-sensor planning

Given a sensor pose s^i, we approximate its information gain by

$$F_{\mathrm{VE}}(X, s^i) = \sum_{v \in V} \mathrm{VE}(X_v, s^i), \tag{6}$$

or alternatively as F_{OA} which is as above but the sum terms are replaced with Eq. (5). We then select the sensor pose that maximizes the information gain, e.g., $\operatorname{argmax}_{s^i \in S_i} F_{\mathrm{VE}}(X, s^i)$, by a separate maximization for each sensor i.

This approach considers individual sensors in isolation, without attempting to coordinate their activities. This may lead to suboptimal behaviour where many sensors are attempting to view the same parts of the workspace.

C. Joint multi-sensor planning

Ideally, the sensor poses s should be selected such that the visible parts of the workspace from each sensors' pose do not overlap. To account for this, we propose to define the information gain of sensor poses $s = (s_i)_{i=1}^n$ as

$$G_{\mathrm{VE}}(X, s) = \sum_{v \in V} \max_{1 \le i \le n} \mathrm{VE}(X_v, s^i) \tag{7}$$

where $F_k(X_v, s^i)$ is the volumetric information for voxel v for sensor pose s^i. A greedy policy is obtained by $\operatorname{argmax}_{s \in S} G_{\mathrm{VE}}(X, s)$. We obtain G_{OA} by replacing the maximization terms with Eq. (5) with $k \in \{\mathrm{VE}, \mathrm{OA}\}$. The joint information gain score of a set of poses s defined in Eq. (7) coordinates the activites of the sensors, as each voxel v only contributes by the maximum volumetric information from it available to any sensor.

V. Offline Experiments

We record sequences of images while moving the two robot arms shown in Fig. 1. With the recorded images, we compare NBV planning strategies for scene reconstruction.

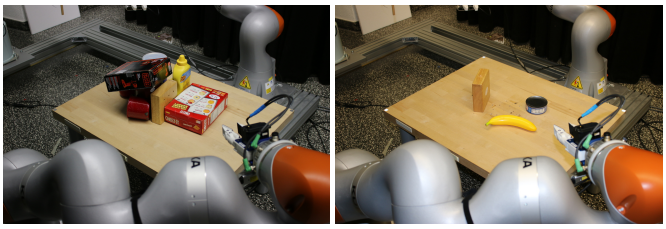

Fig. 2. Left: scene 1, IoU 0.25. Right: scene 8, IoU 0.53.

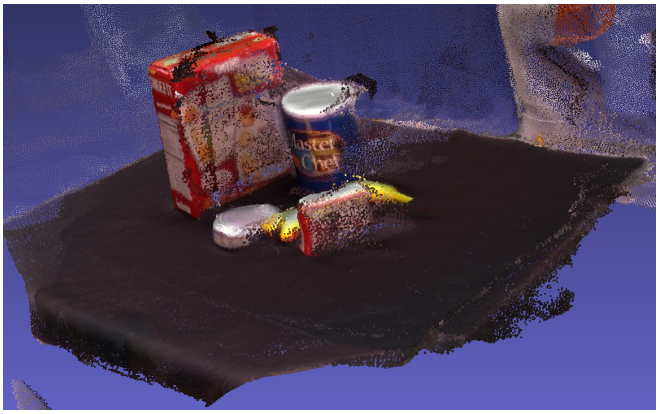

Fig. 3. Partial scene reconstruction of scene 4 using images only from the right camera.

A. Experimental setup

We set up 8 scenes similar to those shown in Fig. 2. Each scene contains a subset of YCB objects [35] arranged on a tabletop, with varying amounts of occlusion. On the opposite sides of the table, there are two robot arms equipped with RGBD cameras, see Fig. 1. The robots must plan poses to view the scene from to reconstruct the tabletop workspace.

To create a reference model to compare the created reconstructions to, for each scene we first move both robot arms in a predefined motion sequence while recording RGBD images. The motion sequence covers the full range of motion of each robot arm. We apply ElasticFusion [36] to create a 3D mesh of the scene from both robots' RGBD images. We align the meshes by the iterative closest point algorithm and segment the tabletop workspace to obtain a reference model.

For each scene, we quantify the amount of overlap between the visible voxels for each camera. We calculate the subsets P^i of points in the reference model that are observable by either of the cameras $i = 1, 2$. The more points there are that both of the cameras can observe, the greater the intersection of P^1 and P^2. The intersection over union (IoU) value $|P^1 \cap P^2| / |P^1 \cup P^2|$ quantifies the amount of overlap. Fig. 2 shows two scenes with low and high IoU values. Although the scene shown on the left is more cluttered, its IoU value is low – due to the occlusion, the potential overlap volume between the two cameras is reduced. In the scene shown on the right, the IoU value is high as the cameras can both potentially view many same areas of the scene. We number our scenes in increasing order of IoU: scene 1 has the lowest IoU, and scene 8 the greatest IoU.

We sample a set S_i of 10 possible sensor poses for each sensor. The poses cover the range of motion of each robot arm, with approximately uniform spatial distances. All the poses are such that the camera optical axis points towards the center of the tabletop workspace.

We represent the reconstruction V by an OctoMap [29] with a resolution of 0.01 meters per voxel. We obtain the initial information on X by first assigning to each voxel $v \in V$ an occupancy probability of $p_v = 0.5$, then we randomly sample a starting pose s_1 and record the corresponding depth images y_1 and fuse them into the map. We plan the remaining sequence $s_{2:T}$ of sensor poses to visit with a total of $T = 6$ views. On each scene, we run 20 trials.

We compare joint planning (Subsection IV-C) to two types of baselines. In the first baseline, we select the sensor poses by individually maximizing the expected information gain of each sensor (Subsection IV-B). In the second baseline, we select sensor poses randomly. For both planning based methods, we apply either visible entropy (Eq. (4)) or the occlusion-aware visible entropy (Eq. (5)) to quantify volumetric information. In all cases, sensor poses are selected based on the complete volumetric reconstruction including all depth images recorded by both sensors up to that time.

B. Evaluation metrics

We evaluate the success of the volumetric reconstruction process by the voxel grid entropy and the surface coverage.

Given sensor poses $s_{1:t}$ and the recorded depth images $y_{1:t}$, the voxel grid entropy at time t is evaluated as $H(X \mid Y_1 = y_1, \ldots, Y_t = y_t)$. A lower entropy indicates a lower average uncertainty of the reconstruction.

We calculate the surface coverage c_t at time t by comparing the current reconstruction obtained by comparing the reconstruction created from the depth images $y_{1:t}$ to the reference model. For each point in the reference model, we find the closest point in the current reconstruction. If the closest point is at a distance of less than 0.01 meters, we consider the corresponding reference model point observed. When $N_{o,t}$ is the number of observed points at time t, and N is the total number of points in the reference model, surface coverage is calculated as $c_t = N_{o,t}/N$.

C. Results

Fig. 4 shows the average entropy and its 95% confidence interval (CI) over all trials in each scene as a function of the number of views t. The scene numbers and IoU values are indicated in the respective axis titles. Since the first view is the same for each method, we show results for $t \geq 2$. The occlusion-aware (OA) and visible entropy (VE) criteria using joint planning are shown by the solid green and cyan lines, respectively. The baselines for single-sensor planning with OA and VE criteria are shown by the dashed red and blue lines, respectively. The solid black line indicates random view selection. The CIs are indicated by vertical bars.

There are few significant differences between the methods. In scene 1, joint planning with VE reaches significantly lower entropy than the other methods. In scene 4, joint planning with OA performs significantly better than other methods.

In some cases (Scenes 4 and 8), the entropy for some methods increases. This is due to error in the relative poses of the cameras. Due to the error, when depth images are integrated into the reconstruction, voxels previously assumed to be free with high likelihood are now observed to be occupied, leading to an increase in entropy. Voxels likely to be occupied are observed free, likewise increasing entropy.

As each scene is different, the range of entropy values also differs. As the scenes contain many voxels that can not be observed, e.g., voxels inside objects, entropy remains high even after many views.

Fig. 5 shows the average surface coverage and its 95% CI over all trials in each scene as a function of the number of views t. The colors for the different methods are as in Fig. 4. In four of eight scenes, joint planning has the highest final surface coverage: VE in scenes 1,6, and 8, and OA in scene 4. Sometimes, joint planning also reaches significantly greater surface coverage than other methods (scene 1, VE, steps 4 and 5; scene 4, OA, step 3; scene 5, OA, step 2; scene 8, VE, steps 4 and 5).

Reasons for the performance differences between OA and VE can be attributed to the characteristics of the scenes. A partial reconstruction of scene 4 using images from the right camera only is depicted in Fig. 3. The area behind the red cracker box and blue coffee can is only visible to the left camera. As the OA criterion considers the occlusion, it creates a plan to observe this area using the left camera, imaging other parts of the scene with the right camera. This leads to significantly better performance after three views, as seen from Fig. 5. The difference shrinks at subsequent steps as other methods also decide to observe the occluded area with the left camera. Scene 8, shown in Fig. 2 right, has very little occlusion. Thus, the VE strategy that maximizes visible entropy performs well as seen in Fig. 5.

We hypothesized that higher IoU values indicating larger overlap between the cameras' visible voxel sets would lead to an advantage for joint planning. However, we do not find conclusive differences between joint and individual planning in the scenes we examined. Joint planning most of the time outperforms individual planning with the same criteria (VE or OA), with the exception of scenes 6 and 7. Of all the scene we examined, scene 8 has the largest IoU score of 0.53. We conjecture that in scenes where there is a higher overlap (as measured by IoU) and overall more views are required for reconstruction, joint planning is more advantageous.

Fig. 5 shows that 70% surface coverage is reached in most cases after just two views. This partly explains the good performance of the random baseline, as there are not many views remaining where a significant amount of additional information could be obtained. In more cluttered scenes, such as scene 1 (Fig. 2 left), NBV planning has an advantage over random view selection. Here, by NBV planning 85% surface coverage is reached after four views, where the random strategy needs five or more views on average.

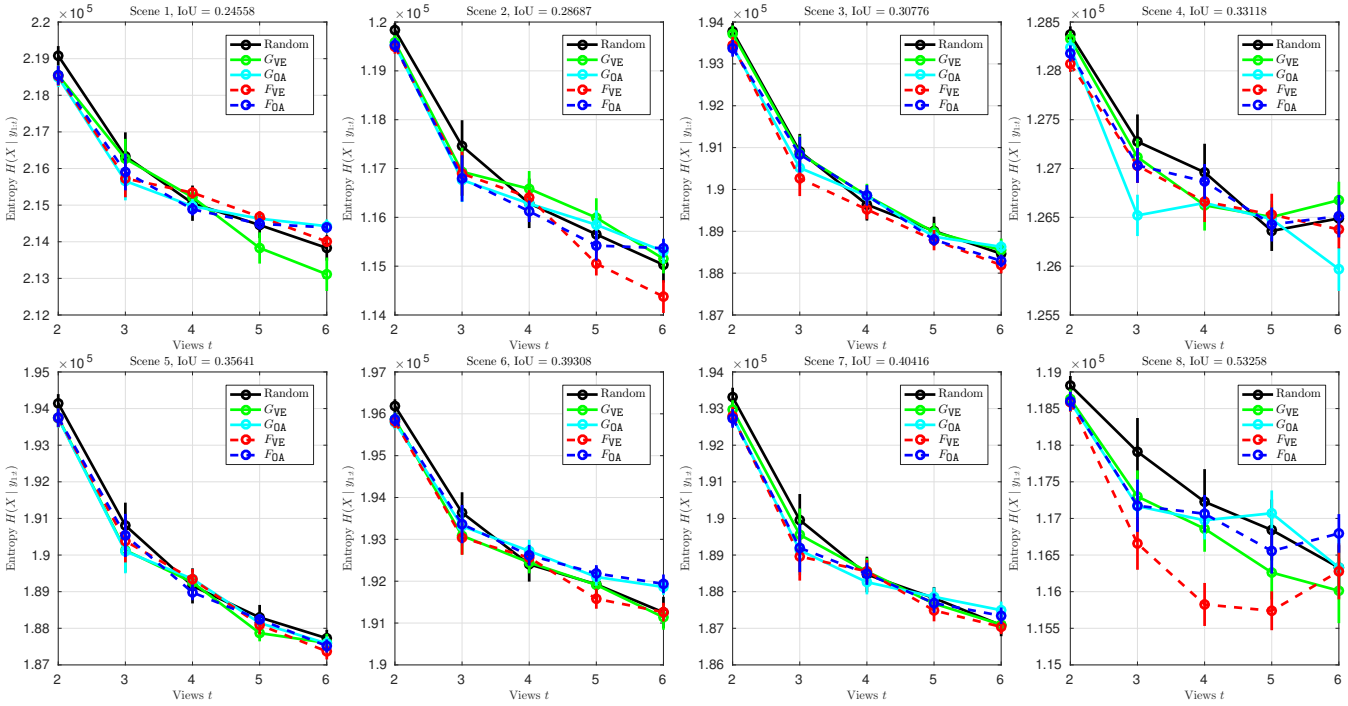

Fig. 4. Average reconstruction entropy as function of number of views t. The vertical bars show 95% confidence intervals. Our proposed joint planning variants are indicated by G_k, while F_k plan views independently for each sensor i. With k we indicate occlusion-aware (OA) [1] or visible entropy (VE) [4].

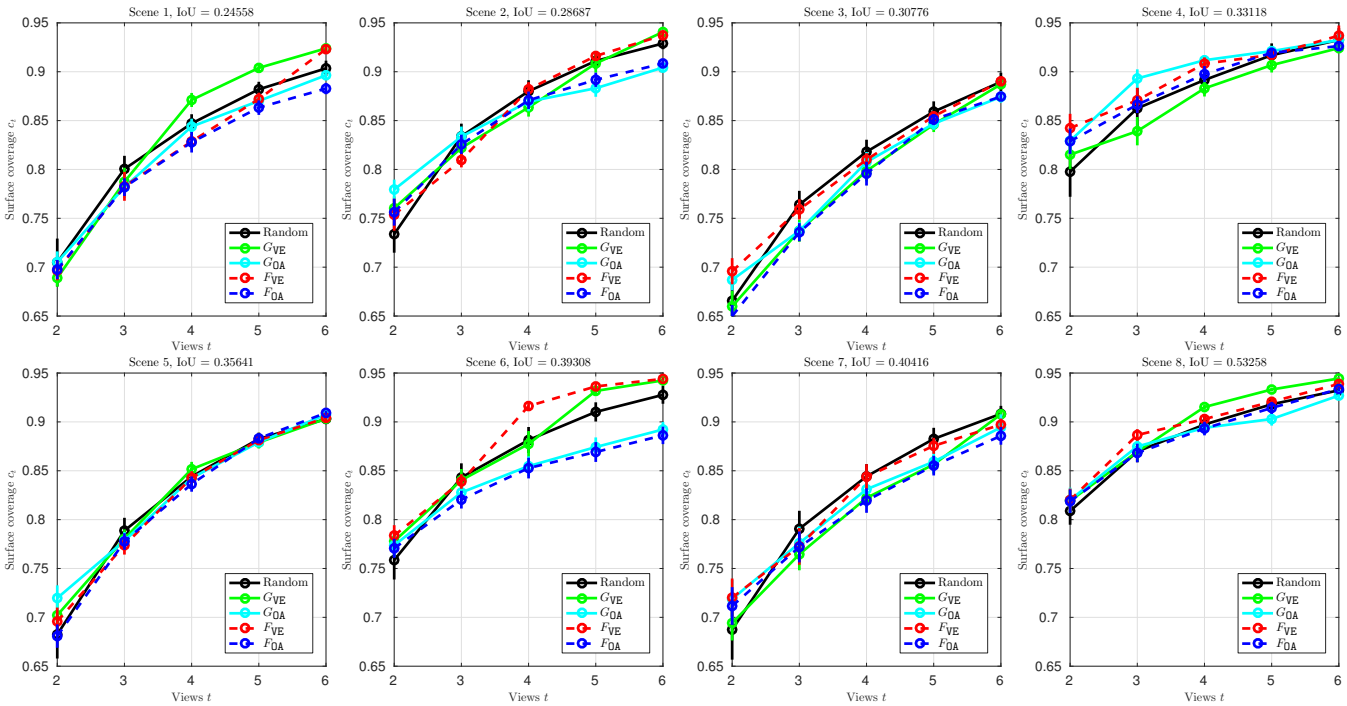

Fig. 5. Average surface coverage as function of number of views t. The vertical bars show 95% confidence intervals. Our proposed joint planning variants are indicated by G_k, while F_k plan views independently for each sensor i. With k we indicate occlusion-aware (OA) [1] or visible entropy (VE) [4].

VI. Discussion and future work

We propose to solve centralized multi-sensor scene reconstruction as a volumetric information gain maximization problem. The problem is submodular, which leads to approximation guarantees for the closed loop greedy policy. We applied a ray-tracing based method for approximation of the information gain, and proposed a modification for joint planning that encourages coordination between multiple sensors.

We provided some preliminary experimental results to illustrate application of our approach to multi-sensor scene reconstruction. However, we did not find conclusive differences between joint and individual planning in the scenes examined. Future work will expand the experimental evaluation in two major ways. Firstly, we will examine larger scenes where several views are required to obtain good surface coverage. We expect this to generally benefit planning-based approaches compared to randomly selecting views. Secondly, we will investigate more closely the hypothesis that a high overlap as measured by the IoU value leads to better performance for jointly planning the sensor views. This can be done by purposefully designing scenes with a high overlap between the sensors' fields of views.

In this work, we applied submodularity to justify our use of the approximate greedy policy. Another potential direction for future work is to consider how submodularity can be more effectively taken advantage of in solving the problem.

Finally, the relaxation of the centralization assumption can be considered. If observation histories cannot be shared between the robots, or can only be shared with a delay of one or multiple time steps, the centralized solution proposed here is not applicable. In such cases, planning a joint policy over multiple time steps that can be executed decentrally without implicit information sharing is required.

References

[1] J. Delmerico, S. Isler, R. Sabzevari, and D. Scaramuzza, "A comparison of volumetric information gain metrics for active 3D object reconstruction," *Autonomous Robots*, vol. 42, no. 2, pp. 197–208, 2018.

[2] R. Border, J. D. Gammell, and P. Newman, "Surface Edge Explorer (SEE): Planning Next Best Views Directly from 3D Observations," in *ICRA*, 2018.

[3] J. I. Vasquez-Gomez, L. E. Sucar, and R. Murrieta-Cid, "View/state planning for three-dimensional object reconstruction under uncertainty," *Autonomous Robots*, vol. 41, no. 1, pp. 89–109, 2017.

[4] S. Kriegel, C. Rink, T. Bodenmüller, and M. Suppa, "Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects," *Journal of Real-Time Image Processing*, vol. 10, no. 4, pp. 611–631, 2015.

[5] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, 1988.

[6] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[7] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments," *The International Journal of Robotics Research*, vol. 30, no. 11, pp. 1343–1377, 2011.

[8] J. I. Vasquez-Gomez, L. E. Sucar, R. Murrieta-Cid, and E. Lopez-Damian, "Volumetric next-best-view planning for 3D object reconstruction with positioning error," *International Journal of Advanced Robotic Systems*, vol. 11, no. 10, p. 159, 2014.

[9] M. Krainin, B. Curless, and D. Fox, "Autonomous generation of complete 3D object models using next best view manipulation planning," in *ICRA 2011*, 2011, pp. 5031–5037.

[10] G. A. Hollinger, B. Englot, F. S. Hover, U. Mitra, and G. S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity," *The International Journal of Robotics Research*, vol. 32, no. 1, pp. 3–18, 2013.

[11] N. Atanasov, B. Sankaran, J. L. Ny, G. J. Pappas, and K. Daniilidis, "Nonmyopic view planning for active object classification and pose estimation," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1078–1090, 2014.

[12] M. Lauri and R. Ritala, "Planning for robotic exploration based on forward simulation," *Robotics and Autonomous Systems*, vol. 83, pp. 15 – 31, 2016.

[13] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding Horizon "Next-Best-View" Planner for 3D Exploration," in *ICRA*, 2016, pp. 1462–1468.

[14] J. Pajarinen and V. Kyrki, "Robotic manipulation of multiple objects as a POMDP," *Artificial Intelligence*, vol. 247, pp. 213–228, 2017.

[15] Y. Xiao, S. Katt, A. ten Pas, S. Chen, and C. Amato, "Online Planning for Target Object Search in Clutter under Partial Observability," in *ICRA*, 2019.

[16] T. Patten, W. Martens, and R. Fitch, "Monte Carlo planning for active object classification," *Autonomous Robots*, vol. 42, no. 2, pp. 391–421, 2018.

[17] J. M. Santos, T. Krajník, J. P. Fentanes, and T. Duckett, "Lifelong Information-Driven Exploration to Complete and Refine 4-D Spatio-Temporal Maps," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 684–691, 2016.

[18] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.

[19] U. M. Erdem and S. Sclaroff, "Automated camera layout to satisfy task-specific and floor plan-specific coverage requirements," *Computer Vision and Image Understanding*, vol. 103, no. 3, pp. 156 – 169, 2006.

[20] G. Olague and R. Mohr, "Optimal camera placement for accurate reconstruction," *Pattern Recognition*, vol. 35, no. 4, pp. 927 – 944, 2002.

[21] P. Natarajan, P. K. Atrey, and M. Kankanhalli, "Multi-camera coordination and control in surveillance systems: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 11, no. 4, pp. 57:1–57:30, 2015.

[22] A. Ilie and G. Welch, "Online control of active camera networks for computer vision tasks," *ACM Transactions on Sensor Networks*, vol. 10, no. 2, pp. 25:1–25:40, 2014.

[23] N. Atanasov, J. Le Ny, K. Daniilidis, and G. J. Pappas, "Decentralized active information acquisition: Theory and application to multi-robot SLAM," in *ICRA*, 2015, pp. 4775–4782.

[24] M. Lauri, E. Heinänen, and S. Frintrop, "Multi-robot active information gathering with periodic communication," in *ICRA*, 2017, pp. 851–856.

[25] G. Best, O. M. Cliff, T. Patten, R. R. Mettu, and R. Fitch, "Dec-MCTS: Decentralized planning for multi-robot active perception," *The International Journal of Robotics Research*, 2018.

[26] M. Lauri, J. Pajarinen, and J. Peters, "Information Gathering in Decentralized POMDPs by Policy Graph Improvement," in *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Montreal, Canada, 2019, pp. 1143–1151.

[27] F. Sukkar, G. Best, C. Yoo, and R. Fitch, "Multi-Robot Region-of-Interest Reconstruction with Dec-MCTS," in *ICRA*, 2019.

[28] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *ICRA*, vol. 2, 1985, pp. 116–121.

[29] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: an efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[31] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[32] J. L. Williams, J. W. Fisher III, and A. S. Willsky, "Performance guarantees for information theoretic active inference," in *Artificial Intelligence and Statistics*, 2007, pp. 620–627.

[33] C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity of Markov Decision Processes," *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987.

[34] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Uncertainty in Artificial Intelligence*, 2005, pp. 324–331.

[35] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-CMU-Berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[36] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "ElasticFusion: Real-time dense SLAM and light source estimation," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1697–1716, 2016.