Auton Robot
DOI 10.1007/s10514-017-9634-0

A comparison of volumetric information gain metrics for active 3D object reconstruction

Jeffrey Delmerico[1] · Stefan Isler[1] · Reza Sabzevari[1] · Davide Scaramuzza[1]

Received: 1 March 2016 / Accepted: 3 April 2017

Abstract In this paper, we investigate the following question: when performing next best view selection for volumetric 3D reconstruction of an object by a mobile robot equipped with a dense (camera-based) depth sensor, what formulation of information gain is best? To address this question, we propose several new ways to quantify the *volumetric information* (VI) contained in the voxels of a probabilistic volumetric map, and compare them to the state of the art with extensive simulated experiments. Our proposed formulations incorporate factors such as visibility likelihood and the likelihood of seeing new parts of the object. The results of our experiments allow us to draw some clear conclusions about the VI formulations that are most effective in different mobile-robot reconstruction scenarios. To the best of our knowledge, this is the first comparative survey of VI formulation performance for active 3D object reconstruction. Additionally, our modular software framework is adaptable to other robotic platforms and general reconstruction problems, and we release it open source for autonomous reconstruction tasks.

Keywords Active vision · Information gain · 3D reconstruction

This is one of several papers published in *Autonomous Robots* comprising the Special Issue on Active Perception.

This research was funded by the Swiss National Science Foundation through the National Center of Competence in Research Robotics (NCCR).

✉ Jeffrey Delmerico
jeffdelmerico@ifi.uzh.ch
http://rpg.ifi.uzh.ch

[1] Robotics and Perception Group, University of Zurich, Zurich, Switzerland

1 Introduction

Object reconstruction in three dimensions is an important step in robust perception and manipulation tasks. In order to reconstruct an object, a mobile robot must position its sensors at different viewpoints in order to fully observe the object. Exhaustive observation is time consuming, so choosing the views that provide the most information is critical in performing this task efficiently.

This problem has been well studied in the robotics and computer vision literature Aloimonos et al. (1988), Bajcsy (1988), Blake and Yuille (1988), Chen et al. (2011), Scott et al. (2003), but often an a priori model of the object is assumed, the implementation is robot-dependent, or the sensor pose options are constrained. Based on the current state of the art, it is not clear that there is an optimal way to quantify the volumetric information for the object reconstruction task, with respect to choosing views based on maximizing information gain. Therefore, this paper's primary contribution is an analysis of many different formulations for this volumetric information, including the current state of the art Kriegel et al. (2015), Vasquez-Gomez et al. (2014), and several new metrics proposed here. We additionally release a robot-agnostic software framework for performing autonomous reconstruction with these formulations.

This paper specifically considers the problem of 3D reconstruction of an object or scene that is unknown a priori, but that is spatially bounded. We assume that we obtain dense 3D input data from a camera-based sensor, but do not restrict to a particular modality. We utilize a probabilistic volumetric map to represent the reconstruction, and we define the *information gain* (IG) in terms of the information contained in its voxels, which we denote as *volumetric information* (VI). We propose several metrics for quantifying this volumetric information based on different ways of measuring model quality

Published online: 22 April 2017

(e.g. completeness, entropy), which focus the reconstruction on observing unexplored regions or refining already observed ones. The reconstruction approach and software framework described in this paper was originally proposed in Isler et al. (2016), but the experimental evaluation here is significantly expanded.

1.1 Related work

Research on the *Next-Best-View problem* and conceptually similar problems in *Active Vision* dates back several decades Aloimonos et al. (1988), Bajcsy (1988) but remains an active area of research Forster et al. (2014). The most frequently referenced surveys of the field include an overview of early approaches by Scott et al. (2003) and an overview of more recent work by Chen et al. (2011). We will follow the categorization introduced by Scott et al. (2003) in distinguishing between model-based and non-model-based reconstruction methods.

Model-based methods assume at least an approximate a priori model of the scene, e.g. from aerial imagery Schmid et al. (2012). They rely on knowledge of the geometry and appearance of the object, which may not be available in many real world scenarios. Non-model based approaches use relaxed assumptions about the structure of the object, but the required information for planning the next best view must be estimated online based on the gathered data Banta et al. (1995), Forster et al. (2014). We utilize a non-model based approach since we do not assume anything about the object aside from its spatial bounds.

The method used to reason about possible next actions depends on the environment representation in which the sensor data is registered. Scott et al. (2003) distinguished between surface-based and volumetric approaches, and more recently methods have been proposed that employ both Kriegel et al. (2015). In a surface-based approach, new view positions are evaluated by examining the boundaries of the estimated surface, represented by e.g. a triangular mesh Pito (1999), Chen and Li (2005). The approach from Krainin et al. (2011) assumes a Gaussian distribution for the uncertainty of reconstruction along the ray from each pixel in a depth camera. Information gain is then the sum of the entropy reduction along all of these rays, weighted by the surface area represented by the pixels. A surface-based approach can be advantageous if the surface representation is also the output of the algorithm because it permits examination of the quality of the model during its construction. However, it is computationally expensive due to the more complex visibility operations that come with a surface representation. A volumetric representation, on the other hand, facilitates simple visibility operations and also allows probabilistic occupancy estimation Hornung et al. (2013). View positions are evaluated by casting rays into the model from the candidate sensor pose and examining the traversed voxels, therefore simulating the image sampling process of a camera. We choose a volumetric representation for its compactness and efficiency with respect to visibility, which forms the basis of several of our VI formulations.

Existing volumetric information metrics fall into two categories: counting metrics and probabilistic metrics. Connolly (1985) and Banta et al. (2000) count the number of unknown voxels. Yamauchi (1997) introduced the concept of *frontier voxels*, defined as voxels bordering free and unknown space, and counted those. This approach has found heavy use in the exploration community, where the exploration of an unknown environment is the goal, rather than reconstruction of a single object Wettach and Berns (2010). The research of Vasquez-Gomez et al. (2014) is a recent example where a set of frontier voxels is used for reconstruction. They count what they call *occplane voxels* (short for occlusion plane), defined as voxels bordering free and occluded space.

Among probabilistic approaches, one method is to use information theoretic entropy to estimate expected information Kriegel et al. (2015). This necessitates the use of occupancy probabilities but has the advantage that the sensor uncertainty is considered. Potthast and Sukhatme (2014) argue that the likelihood that unknown voxels will be observed decreases as more unknown voxels are traversed and that this should be considered in the information gain calculation. They model the observability using a Hidden Markov Model and introduce empirically determined state transition laws to calculate posterior probabilities in a Bayesian way.

We propose several VI formulations of both the counting and probabilistic types, and specifically compare to the state of the art metrics in Kriegel et al. (2015) and Vasquez-Gomez et al. (2014).

1.2 Contributions and outline

In this paper, we propose a set of volumetric information formulations and evaluate them along with recent formulations in the literature:

- *Occlusion Aware VI* Quantifies the expected visible uncertainty by weighting the entropy within each voxel by its visibility likelihood.
- *Unobserved Voxel VI* Restricts the set of voxels that contribute their VI to voxels that have not been observed yet.
- *Rear Side Voxel VI* Counts the number of voxels expected to be visible on the back side of already observed surfaces.
- *Rear Side Entropy VI* Quantifies the expected amount of VI as defined for the Occlusion Aware VI, but restricted to areas on the rear side of already observed surfaces.

- *Proximity Count VI* Weighted higher for unobserved voxels that are close to already observed surfaces.

We evaluate all of these VIs in synthetic experiments designed to isolate their performance from any environmental factors. We consider the following criteria: the amount of discovered object surface (surface coverage), the reduction of uncertainty within the map, and the computational cost of next best view selection. This is the first such comparative survey of information gain metrics.

We release our modular software framework for active dense reconstruction to the public. The ROS-based, generic system architecture enables any position controlled robot equipped with a depth sensor to carry out autonomous reconstructions.

The paper is organized as follows: we introduce our proposed volumetric information formulations in Sect. 2, then give an overview of our software framework in Sect. 3. Experiments comparing the performance of the VI formulations in a simulated environment are shown in Sect. 4. In Sect. 5, we discuss the results of our experiments, and finally, in Sect. 6 we summarize our findings.

2 Volumetric information

To find the next best view within a set of candidates, we estimate the obtainable *Information gain* (IG) for each view by evaluating the amount of *Volumetric information* (VI) contained in the visible area of the map. We define VI as the amount of information a single voxel is expected to provide when seen from a particular view Isler et al. (2016). The next best view (NBV) is the view that maximizes this metric, minus any estimated costs.

For every view v within a set of candidate sensor positions $\mathcal{V}$, the 3D points from the camera-based range sensor are projected into the map. The projection is carried out through ray casting, yielding a set $\mathcal{R}_v$ of rays cast for every view. As each ray traverses the map we accumulate the volumetric information within the set of visited voxels $\mathcal{X}$. During ray casting, a ray ends when it is incident on a physical surface or when it reaches the limit of the map. The predicted IG for a view v, denoted as $\mathcal{G}_v$, is then the cumulative volumetric information $\mathcal{I}$ collected along all rays r cast from v, such that:

$$\mathcal{G}_v = \sum_{\forall r \in \mathcal{R}_v} \sum_{\forall x \in \mathcal{X}} \mathcal{I}. \tag{1}$$

The formulation of VI in Eq. 1 and the set of views for which it is evaluated define the behavior of the system. By choosing a VI formulation that is directly proportional to voxel uncertainty, we can favor views that observe unknown areas in our map. If we choose a VI formulation that assigns high values at the frontiers of unknown areas and previously observed objects, the system will favor views that explore these boundaries and gather more information about partly observed objects within the map. We discuss a set of formulations for VI considering different aspects like uncertainty in the map or the proximity of known surfaces in the following sections.

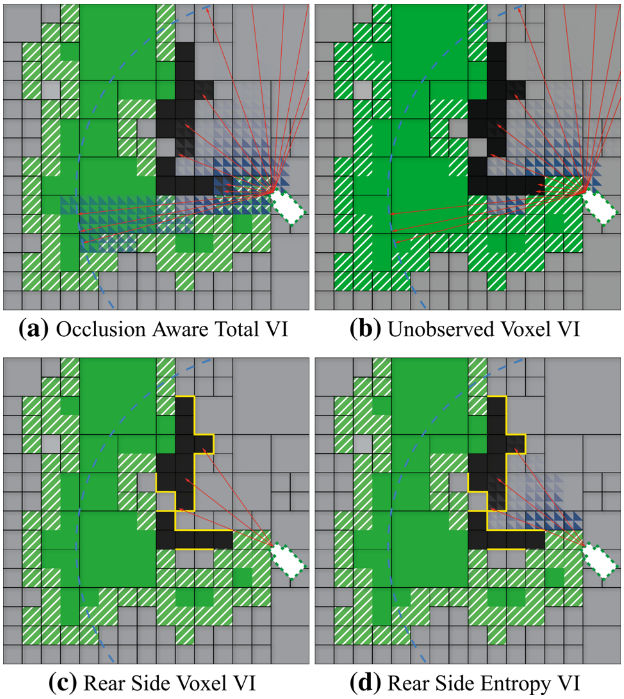

Fig. 1 Visualization of the IG function with different VI formulations in 2D on an exemplary state of the map: The map shows occupied (*black*), unknown (*grey*) and unoccupied (*green*) regions and a view candidate (*white camera*). Additionally frontier voxels (*striped white*), unknown object sides (*yellow*), considered ray sets (*red*), maximal ray length (*dashed blue circle*) and VI weights (opacity of *blue triangles*) are shown. Note that the proposed *Proximity Count VI* behaves like the *Rear Side Voxel VI* (*bottom left*), but with a weight that is dependent on distance from previously observed surface voxels, which would be difficult to visualize clearly in this diagram (Color figure online)

We illustrate several of our proposed VI formulations in Fig. 1. These diagrams show the state of the map at one point during the reconstruction, but simplified to 2D for clarity. A single candidate view is being evaluated under the different VI metrics, and this view has not yet been taken by the camera, so the camera position is in an unexplored part of the map. The voxels that are considered by each VI formulation are illustrated with colors and shading.

2.1 Considering map uncertainty

Uncertainty within a volumetric map that encodes occupancy probabilities for each voxel can be defined using the voxel's entropy:

$$\mathcal{H}(x) = -P_o(x)\ \ln P_o(x) - \overline{P}_o(x)\ \ln \overline{P}_o(x), \tag{2}$$

where $P_o(x)$ denotes the probability of voxel x being occupied, while $\overline{P}_o(x)$ is the complement probability of P_o, i.e. $\overline{P}_0 = 1 - P_o$. A voxel for which we have no information about its occupancy ($P_o = 0.5$) has the highest uncertainty (and hence entropy) with $P_o(x) = 0.5$ and $\mathcal{H}(x) = 1.0$ *Shannon*. If we consider that a view observing a map area where we have high uncertainty is likely to yield more information, entropy can be used as a metric to maximize the total amount of new information gathered in each iteration of the reconstruction. We can therefore define a VI formulation based purely on entropy:

$$\mathcal{I}_e(x) = \mathcal{H}(x). \tag{3}$$

The corresponding IG that estimates the entropy for a view $v \in \mathcal{V}$ is given by substituting Eqs. 3 into 1.

Kriegel et al. (2015) used the entropy as defined in Eq. 2 to propose an entropy-based IG:

$$\mathcal{G}_{v,Kriegel}(v) = \frac{1}{n} \sum_{\forall r \in \mathcal{R}_v} \sum_{\forall x \in \mathcal{X}} \mathcal{H}(x), \tag{4}$$

where n is the total number of traversed voxels. We therefore refer to this as *Average Entropy VI*. While Eq. 3 favors views for which the cast rays traverse deep into the map and visit many voxels, Eq. 4 may also yield a high IG for views close to known surfaces where the rays traverse fewer voxels, but where their entropy is high.

2.2 Visibility and occlusions

Since our map is a probabilistic voxel grid, we can consider the likelihood of a voxel being visible from a particular view instead of simply integrating entropy over all traversed voxels. We call this formulation *Occlusion Aware VI*. The visibility likelihood P_v of a voxel x_n is given by:

$$P_v(x_n) = \prod_{i=1}^{n-1} \overline{P}_o(x_i), \tag{5}$$

where $x_i, i = 0 \ldots n-1$ are all voxels traversed along a ray before it reaches voxel x_n. Using Eq. 5 we define:

$$\mathcal{I}_v(x) = P_v(x)\ \mathcal{H}(x). \tag{6}$$

By substituting Eq. 6 in the IG formulation from Eq. 1, this IG formulation estimates the *visible entropy* for a particular view $v \in \mathcal{V}$, thus favoring views with a high visibility uncertainty. This is a very natural way of dealing with occlusions within the map: a voxel with a large unobserved volume between its position and the view candidate is less likely to be visible due to occlusions, and is therefore less likely to contribute information than voxels that are closer to the sensor position, or that are behind more certain free space. This VI formulation is illustrated in Fig. 1a.

2.3 Focusing on areas of interest

In goal directed tasks, not all voxels provide the same amount of information, and intuition about the task can be exploited to drive the choice of NBV. For example, when reconstructing an object, views that favor parts of the object that have not yet been observed can reveal more of the object's surface. To favor view candidates that observe task-specific areas of interest, we can define a VI formulation that assigns high information content to voxels that have a high likelihood of belonging to the interest area or a VI that removes areas of no interest from consideration.

An example of focus by exclusion is to only sum up VI as defined in Eq. 6 over voxels that are thus far unobserved, and therefore remove areas with high confidence from consideration. We set up an indicator function based on the observation state of the voxel:

$$\mathcal{I}_u(x) = \begin{cases} 1 & x \text{ is unobserved} \\ 0 & x \text{ is already observed} \end{cases} \tag{7}$$

In combination with the visible entropy formulation from Eq. 6 we get:

$$\mathcal{I}_k(x) = \mathcal{I}_u(x)\ \mathcal{I}_v(x) \tag{8}$$

We denote this VI the *Unobserved Voxel VI*. It estimates the hidden information in unobserved voxels. This VI formulation is illustrated in Fig. 1b.

An example of a class of interest areas is the set of voxels that have not been observed but are adjacent to an occupied voxel on a ray, what we refer to as *rear side voxels*, because they represent voxels on the back side of the occupied voxels that have already been observed. The intuition here is that rays that are incident on the rear side of an already observed surface are very likely to be incident on previously unobserved parts of the object. The simplest formulation for this is an indicator function, which determines whether a voxel is part of an interest area in a binary fashion:

$$\mathcal{I}_b(x) = \begin{cases} 1 & x \in \mathcal{S}_o \\ 0 & x \notin \mathcal{S}_o \end{cases}, \tag{9}$$

where $\mathcal{S}_o$ is the set of *rear side voxels*, defined as unobserved voxels such that the next voxel on their ray is estimated to be occupied. Substituting Eqs. 9 into 1 we obtain an IG that counts how many of the rays cast for a particular view are incident on an unknown side of a previously observed object surface. Such a ray is necessarily incident on an unknown surface of the object of interest. We denote this *Rear Side Voxel VI*, and visualize it in Fig. 1c.

As an alternative to this count, we reformulate $\mathcal{S}_o$ in Eq. 9 as the set of all unknown voxels between the sensor and an occupied voxel, combined with the entropy and visibility based formulation from Eq. 6, such that:

$$\mathcal{I}_n(x) = \mathcal{I}_b(x)\, \mathcal{I}_v(x). \tag{10}$$

This type of VI estimates the visible uncertainty within the unknown volume behind known surfaces. We denote it as *Rear Side Entropy VI*, and visualize it in Fig. 1d.

A problem with the intuition behind the *Rear Side Voxel VI* (Eq. 9) and *Rear Side Entropy VI* (Eq. 10) is that we consider all of the voxels behind an observed surface to have the same weight. For continuous objects, the voxels that are closer to the rear side of the surface are more likely to be occupied than those that are farther away. We propose another VI formulation that introduces a weighting factor on the information in a voxel based on its distance behind these previously observed surfaces.

When estimating IG from a candidate sensor pose, it would be computationally expensive to compute the distance behind the nearest surface for each traversed voxel. Instead, we augment our map and implement this computation during the data registration step when the surface is first observed. Given a point cloud of observations from the most recent NBV that was chosen, we continue the rays for the observed points behind the occupied voxel up to a maximum distance d_{max}. For each voxel that is traversed beyond the point, we mark it with the distance $d(x)$ to the surface voxel. If a voxel is already marked, we keep the smaller distance. We then define the *Proximity Count VI* as:

$$\mathcal{I}_p(x) = \begin{cases} d_{max} - d(x) & x \quad \text{is unobserved} \\ 0 & x \quad \text{is already observed} \end{cases}. \tag{11}$$

This VI functions as a weighted version of the *Rear Side Voxel VI* where the weight is higher the closer it is to an already observed surface voxel.

The *Occlusion Aware VI*, *Unobserved Voxel VI*, *Rear Side Voxel VI* and the *Rear Side Entropy VI* are visualized in an exemplary 2D scenario in Fig. 1. The images show a snapshot of a possible state in the map during reconstruction and how IG is estimated: each voxel has a state that is estimated based on registered point measurements. Based on this state, we compute the voxel's volumetric information. This VI is then integrated for the voxels along the rays to obtain the information gain estimate.

3 System overview

We approach the autonomous reconstruction task as an iterative process consisting of the three largely independent parts (i) 3D model building, (ii) view planning, and (iii) the camera positioning mechanism, as observed in Torabi and Gupta (2012). The orthogonality of the involved tasks has inspired us to design our autonomous reconstruction system in a modular manner to allow for fast reconfiguration of the software for different applications and robotic setups. We utilize the Robot operating system (ROS) Quigley et al. (2009) software framework, which allows a hardware-agnostic design through the use of its interprocess communication interfaces. Within this framework, we use off-the-shelf components for the 3D model building and camera positioning sub-tasks, and focus only on view planning based on our proposed IG formulations.

Conceptually, our systems build upon the framework presented in Isler et al. (2016): it consists of several independent modules that interact through well-defined interfaces, yielding a very flexible system architecture, as shown in Fig. 2. Single components can be exchanged without affecting other parts of the system, and only the sensor and robot interfaces need to be implemented for use with a new robot platform.

The components that are part of the *Perception System* are responsible for data acquisition and processing. The output of this module is a point cloud of observed 3D points in world coordinates. Our software framework allows additional information such as color or measurement uncertainty to be included and incorporated into the reconstruction, but we do not consider any data other than the geometry in this work.

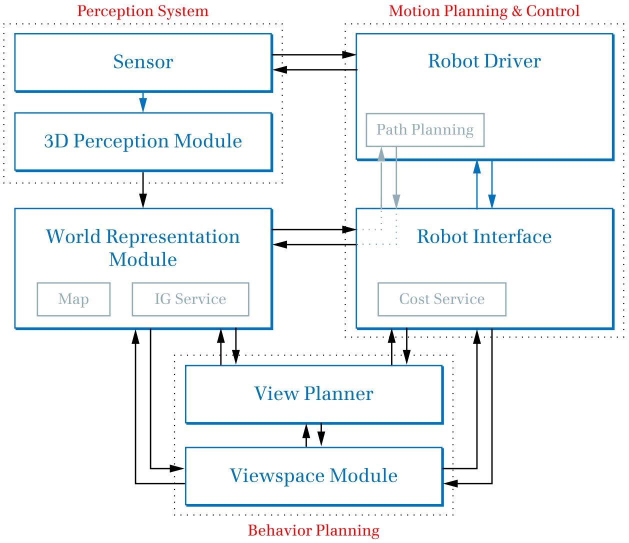

Fig. 2 Conceptual system overview: Main modules, important components, and their communication interfaces (*arrows*) are visualized

The *Motion Planning and Control* components control the movement of the robot, where the *Robot Interface* (RI) defines the interaction between robot specific code and the other modules. Part of this interface is a service that calculates the cost to reach candidate views, e.g. based on their distance from the current position or the estimated energy necessary to carry out the movement. This component is used to keep the robot movement bounded. The robot receives commands to move the sensor to a given new viewpoint but carries out path planning to reach this position itself using the *Robot Interface*.

All perceived data is registered within the map, which is part of the *World Representation Module* (WM). This module gives access to the current map and additionally provides a service for information gain (IG) evaluation of given views, as discussed in Sect. 2.

The high level behavior of the robot is controlled from the two *Behavior Planning* components. We define the *viewspace* as the space of candidate sensor positions, often a discrete set of 6 DoF poses (position and orientation of the sensor). The *Viewspace Module* (VM) provides the current set of candidates to the *View Planner*. This may be a static set that is fixed a priori, or a dynamic set that is recreated and evaluated at each iteration. Static viewspaces are usually sampled from simple geometries circumscribed around the object, such as a cylinder Pito (1999), a sphere Trummer et al. (2010), or a combination of cylinder and hemisphere Isler et al. (2016). Dynamic viewspaces are generated based on the current map and the possible poses of the robot Kriegel et al. (2015), Wettach and Berns (2010), or may also be sampled from a fixed geometry, but randomly resampled with every iteration Vasquez-Gomez et al. (2014).

The *View Planner* evaluates the set of candidate views it receives from the *Viewspace Module* and determines the next best view (NBV), which it commands to the robot as the next target position. For every view in the viewspace, it requests the IG $\mathcal{G}_v$ from the *World Representation Module* and the cost $\mathcal{C}_v$ from the *Robot Interface*, calculating the utility $\mathcal{U}_v$ of the view:

$$\mathcal{U}_v = (1-\gamma)\,\frac{\mathcal{G}_v}{\sum_{\mathcal{V}}\mathcal{G}} - \gamma\,\frac{\mathcal{C}_v}{\sum_{\mathcal{V}}\mathcal{C}}, \tag{12}$$

where $\sum_{\mathcal{V}}\mathcal{G}$ and $\sum_{\mathcal{V}}\mathcal{C}$ are the total IG and cost, respectively, predicted for the current iteration over all view candidates and $\gamma \in [0, 1]$ is the user defined cost weight. The NBV v^* in the current viewspace is found by maximizing Eq. 12:

$$v^* = \arg\max_v \mathcal{U}_v. \tag{13}$$

The robot stops when a predefined termination criterion is fulfilled. For example, this criterion could be that the highest expected information gain for any candidate view falls below a user defined threshold,

$$\mathcal{G}_v < g_{thresh} \quad \forall v \in \mathcal{V} \tag{14}$$

or that a sufficient amount of the map has been observed.

Algorithm 1 Active Volumetric Reconstruction

```
repeat
    Command RI Layer to move the sensor to the NBV.
    Signal robot to collect data from the sensors.
    Request the view candidate set from VM.
    Request cost for each view candidate from RI Layer.
    Request IG for each view candidate from WM.
    Calculate the utility function combining IGs and costs.
    Determine NBV.
until Termination Criteria met
return Volumetric map and point cloud of object
```

4 Experiments

We will first discuss how we use the presented approach from Sects. 2 and 3 to have a simulated mobile robot generate a complete volumetric model of an object that is unknown a priori, but spatially bounded. The robot positions the sensor at different viewpoints, pointing at the volume that contains the object, with the goal of carrying out the reconstruction as quickly as possible. Reconstruction proceeds according to Algorithm 1 within the *View Planner module* until reaching the user's termination criterion.

Information gain based on VI is a metric used as an indicator to estimate which next view will be most informative to the reconstruction. An informative view maximizes (i) the amount of new object surface discovered and (ii) the uncertainty reduction in the map. Additionally, (iii), we are interested in minimizing the computational cost of evaluating candidate views. We therefore evaluate our VI formulations on these three criteria.

Our map representation is a probabilistic volumetric voxel grid based on OctoMap Hornung et al. (2013). The *Viewspace Module* feeds the *View Planner* a static viewspace, which is a reasonable choice for these experiments since we know a priori the volume in which the unknown object is contained. We report the performance of the reconstruction planning using the VI formulations presented in Sect. 2. We consider the amount of object surface discovered over time and the uncertainty reduction in the map, as a way of quantifying the progess of the reconstruction. We also measure the computation time for each VI formulation, in order to asses its efficiency.

4.1 Experimental setup

Our simulated experiments are designed to isolate the performance of the VI formulation in the reconstruction from any environmental factors. We utilize an uncluttered scene and an idealized depth sensor in order to provide optimal performance from each view. We also use a free-moving sensor as an idealized "robot" without movement constraints or sensor pose uncertainty. Consequently, the only independent variable in evaluating the performance is the next best view chosen by each VI formulation.

The reconstruction scene for the simulation consists of an object placed in an empty environment with no ground plane. Each model was adjusted in size to be approximately 0.5 m in length along its largest dimension, so that its extent is bounded within a 1.0 m cube. Around the object we generate a set of 48 candidate views, distributed uniformly across a cylinder with a half-sphere on top, such that they face the model from different poses. We present results for a total of 11 models: All of them have been generated from data available online. Our dataset features the Stanford bunny and dragon,[1] the Armadillo model from TU Munich,[2] and eight models generated from color-tagged 3D range data.[3] The robot is a free-flying, idealized RGB-D sensor that captures the structure of the scene exactly, and with which we can carry out unconstrained movements in 6 DoF.

For a simulation environment, we use *Gazebo*[4] in conjunction with ROS. The generated, Gazebo-ready models are available online.[5]

All simulated reconstructions begin by placing the RGB-D sensor at a randomly chosen view from the view space. The sensor output is a point cloud that is integrated into the World Module's OctoMap Hornung et al. (2013). We use a resolution of 1 cm for the map, with 0.97 as the likelihood threshold to consider a voxel to be occupied, and 0.12 as the threshold to be considered free. Views are removed from the candidate set once visited in order to only visit novel views during the reconstruction. We do not use the termination criterion but instead run 20 iterations for each trial.

4.2 Evaluation

We present the results of our simulated experiments in Table 1, and more extensive visualizations for a few select models in Fig. 3. Reconstruction for each model and VI combination was performed 20 times, each with a random starting view, and the results were averaged over all of the trials.

To quantify the reconstruction progress in terms of surface coverage, we compare the pointcloud models obtained during reconstruction with a pointcloud generated from the ground truth model. For each point in the ground truth model, the closest point in the reconstruction point cloud is queried. If this point is closer than a registration threshold[6] the surface point of the original model is considered to have been observed. The surface coverage c_s is then the percentage of observed surface points compared to the total number of surface points of the model:

$$\text{Surface coverage } c_s = \frac{\text{Observed surface points}}{\text{Surface points in original model}}. \tag{15}$$

To calculate the total entropy we consider a bounding cube with 1.28 m side length around the object and define the total entropy to be

$$\text{Entropy in map} = \sum \text{Entropy of voxels within cube.} \tag{16}$$

All the models evaluated in simulation have an extent less than 1 m^3. For our OctoMap resolution of 0.01 m, a cube with $(1.28 = 0.01 \times 2^7)$ m side length is the smallest level in the octree completely containing our models. The resulting maximal entropy within this bounding box, where each voxel is initialized to an occupancy likelihood of 50%, is $-(2^7)^3 \log_2 0.5 = 2.097 \times 10^6$ *Shannon*.

For both surface coverage and entropy, we have computed the area under the curve (AUC) as a way to summarize the performance of each VI over the course of the full reconstruction procedure. The view iteration is normalized by the maximum number of views (20 in our experiments) and the maximum value of the metric, and then the area is computed for each VI. Therefore, the AUC varies between 0.0 and 1.0, with a higher value being better for surface coverage, and a lower value being better for entropy. Note that since there are many unobserved voxels in the scene, the entropy AUC remains high, even for the best performing VIs.

We also measured the time required to evaluate the next best view for each VI formulation, averaged over each iteration, over all of the 20 trials and all of the 11 models. All trials were performed on a PC with an 8-core Intel i7-4770K CPU, operating at 3.50GHz, and using 8 parallel threads for the ray-casting step in evaluating each view. The timing results are presented in Table 2.

We compare our formulations to the information gain methods proposed by Kriegel et al. (2015) (see Eq. 4) and Vasquez-Gomez et al. (2014). Vasquez-Gomez et al. (2014) define desired percentages $\alpha_{des,oc} = 0.2$ and $\alpha_{des,op} = 0.8$ of occupied and occplane voxels in the view, respectively, and

[1] Available from the Stanford University Computer Graphics Lab.

[2] Presented in Rodolà et al. (2013); available from TUM Computer Vision Group.

[3] Generated with MeshLab from the *multiple view stereo* (MVS) dataset Jensen et al. (2014) from the Image Analysis and Computer Graphics section at DTU Denmark.

[4] http://www.gazebosim.org.

[5] http://rpg.ifi.uzh.ch.

[6] We chose $d_{reg} = 0.5\,\text{cm}$.

Table 1 Simulation results for all tested models and VI formulations, averaged over 20 trials each

Model	VI	Surface coverage [%] Iteration nr.					Entropy in map	Model	VI	Surface coverage [%] Iteration nr.					Entropy in map
		2	4	6	10	AUC	AUC			2	4	6	10	AUC	AUC
 Angel	OA	0.732	0.904	0.960	0.963	0.882	0.9473	 Armadillo	OA	0.645	0.882	**0.980**	0.993	0.893	0.9466
	UV	0.741	0.910	0.959	0.963	0.883	0.9473		UV	0.594	0.857	**0.980**	**0.994**	0.887	0.9466
	RSV	0.686	0.941	**0.961**	**0.964**	0.881	0.9471		RSV	0.647	0.833	0.908	0.986	0.876	0.9464
	RSE	0.714	0.949	**0.961**	**0.964**	0.884	0.9472		RSE	0.636	0.811	0.933	0.988	0.878	0.9464
	PC	0.798	0.930	0.954	**0.964**	0.888	0.9472		PC	0.766	**0.950**	0.979	0.990	**0.908**	0.9465
	VG	**0.844**	**0.950**	**0.961**	0.963	**0.895**	0.9471		VG	**0.776**	0.944	0.973	0.981	0.906	**0.9463**
	Kr	0.742	0.919	0.955	**0.961**	0.883	**0.9470**		Kr	0.652	0.903	0.974	0.989	0.896	0.9466
	Rand	0.719	0.883	0.922	0.958	0.872	**0.9470**		Rand	0.643	0.876	0.957	0.986	0.885	0.9464
 Buddha	OA	0.665	0.894	**0.955**	0.957	0.869	0.9401	 Bunny	OA	0.607	0.810	**0.915**	**0.924**	0.829	0.9462
	UV	0.657	0.889	**0.955**	0.956	0.867	0.9401		UV	0.568	0.800	**0.915**	**0.924**	0.824	0.9462
	RSV	0.669	**0.951**	**0.955**	**0.958**	0.882	0.9402		RSV	0.605	0.721	0.868	0.916	0.812	0.9461
	RSE	0.692	**0.951**	0.954	**0.958**	0.884	0.9402		RSE	0.622	0.794	0.869	0.920	0.825	0.9461
	PC	**0.883**	0.949	0.954	**0.958**	**0.893**	0.9406		PC	0.707	0.861	0.898	0.917	0.839	0.9464
	VG	0.815	0.934	0.946	0.952	0.882	0.9403		VG	**0.716**	**0.880**	0.908	0.916	**0.842**	0.9463
	Kr	0.718	0.894	0.928	0.950	0.867	**0.9400**		Kr	0.575	0.870	0.899	0.910	0.827	**0.9459**
	Rand	0.714	0.904	0.935	0.952	0.870	0.9401		Rand	0.583	0.781	0.865	0.903	0.813	0.9460
 Doves	OA	0.676	0.932	**0.963**	**0.969**	0.883	0.9418	 Dragon	OA	0.494	0.717	0.883	**0.936**	0.820	0.9443
	UV	0.718	0.932	**0.963**	**0.969**	**0.886**	0.9418		UV	0.521	0.741	**0.907**	0.934	0.825	0.9442
	RSV	0.683	**0.935**	0.955	0.967	0.884	0.9419		RSV	0.417	0.615	0.859	0.927	0.798	0.9440
	RSE	0.645	**0.935**	0.961	0.966	0.883	0.9419		RSE	0.461	0.599	0.883	0.930	0.807	0.9437
	PC	**0.731**	0.925	0.952	0.966	0.885	0.9423		PC	0.604	0.789	0.873	0.932	**0.836**	0.9447
	VG	0.724	0.883	0.915	0.953	0.870	0.9417		VG	**0.622**	**0.798**	0.858	0.910	0.826	**0.9436**
	Kr	0.670	0.893	0.940	0.963	0.872	**0.9415**		Kr	0.467	0.795	0.869	0.911	0.812	0.9443
	Rand	0.656	0.849	0.919	0.959	0.864	0.9418		Rand	0.524	0.737	0.820	0.899	0.800	0.9439
 Head	OA	0.814	0.942	**0.957**	0.957	0.888	0.9411	 Owl	OA	0.723	0.871	0.992	**0.999**	0.901	0.9462
	UV	0.836	0.946	**0.957**	0.957	0.890	0.9410		UV	0.689	0.887	0.994	**0.999**	0.901	0.9462
	RSV	0.812	**0.954**	**0.957**	**0.959**	0.893	**0.9409**		RSV	0.685	0.717	0.965	**0.999**	0.886	0.9462
	RSE	0.800	**0.954**	**0.957**	0.958	0.892	**0.9409**		RSE	0.651	0.743	0.964	**0.999**	0.883	0.9462
	PC	**0.924**	0.951	0.956	0.957	**0.897**	0.9411		PC	0.881	0.993	**0.998**	**0.999**	0.929	0.9464
	VG	0.874	0.940	0.944	0.949	0.887	**0.9409**		VG	**0.896**	**0.994**	**0.998**	**0.999**	**0.931**	0.9463
	Kr	0.798	0.945	0.952	0.955	0.888	0.9410		Kr	0.695	0.991	0.996	0.998	0.917	**0.9458**
	Rand	0.760	0.925	0.945	0.955	0.882	**0.9409**		Rand	0.739	0.935	0.988	0.998	0.912	0.9459
 Rabbit	OA	0.548	0.717	0.879	**0.889**	0.791	0.9445	 Skull	OA	0.532	**0.766**	**0.869**	**0.901**	**0.807**	0.9450
	UV	0.617	0.750	0.878	**0.889**	0.799	0.9446		UV	0.466	**0.766**	0.868	0.898	0.800	0.9452
	RSV	0.549	0.606	0.714	0.876	0.754	0.9445		RSV	0.481	0.623	0.693	0.843	0.750	0.9451
	RSE	0.577	0.702	0.801	0.886	0.780	0.9445		RSE	0.527	0.673	0.745	0.861	0.776	0.9448
	PC	0.633	0.836	0.881	0.887	0.810	0.9450		PC	0.494	0.722	0.821	0.887	0.788	0.9451
	VG	**0.734**	0.840	0.879	0.885	**0.815**	0.9448		VG	**0.616**	0.744	0.827	0.874	0.804	0.9446
	Kr	0.573	**0.851**	**0.882**	0.888	0.806	**0.9433**		Kr	0.563	0.751	0.816	0.893	0.804	0.9446
	Rand	0.566	0.791	0.867	0.887	0.795	0.9437		Rand	0.473	0.692	0.784	0.878	0.778	**0.9445**
 Snowmen	OA	0.601	**0.853**	0.888	**0.919**	**0.828**	0.9439	VI Legend	OA	Occlusion Aware (Sec. 2.2)					
	UV	0.569	**0.853**	**0.890**	**0.919**	0.827	0.9439		UV	Unobserved Voxel (Sec. 2.3)					
	RSV	0.565	0.665	0.850	0.907	0.797	0.9440		RSV	Rear Side Voxel (Sec. 2.3)					
	RSE	0.638	0.816	0.868	0.904	0.815	0.9439		RSE	Rear Side Entropy (Sec. 2.3)					
	PC	0.636	0.833	0.873	0.894	0.816	0.9446		PC	Proximity Count (Sec. 2.3)					
	VG	**0.657**	0.830	0.871	0.892	0.819	0.9437		VG	Area Factor (Vasquez-Gomez et al. 2014)					
	Kr	0.570	0.787	0.881	0.907	0.810	**0.9433**		Kr	Average Entropy (Kriegel et al. 2015)					
	Rand	0.593	0.786	0.848	0.900	0.808	0.9437		Rand	Random View					

Surface coverage is shown for several iterations during the reconstruction procedure. Additionally, the *Area Under the Curve* (AUC) is shown for the surface coverage and entropy. Surface coverage AUC is normalized by total reconstruction steps (20); higher is better, with a maximum of 1.0. Entropy AUC is normalized by the maximum possible entropy in the map, and total reconstruction steps; lower is better. We compare our proposed formulations to the state of the art (Kriegel et al. 2015; Vasquez-Gomez et al. 2014) as well as randomized view selection without consideration of IG. The best performing VI is shown in **bold** for each model and column

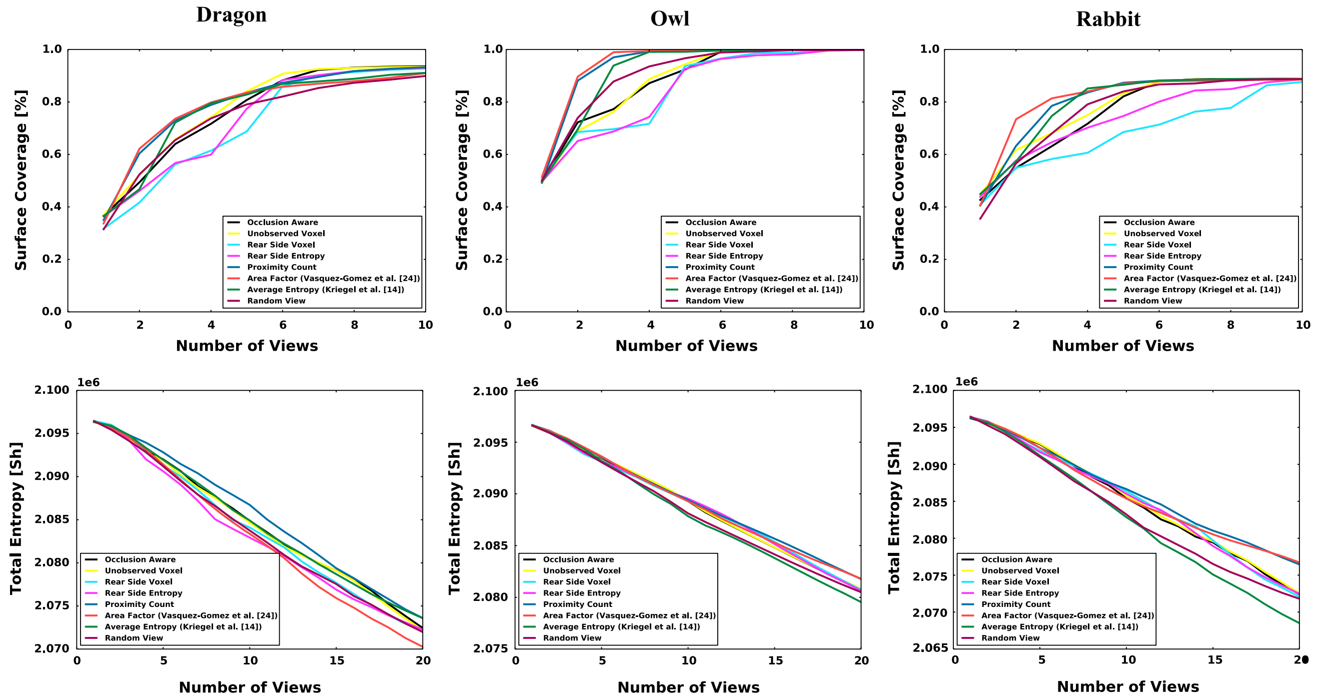

Fig. 3 Surface completion (*top*) and total map entropy (*bottom*) for selected models. These models illustrate how the *Area Factor VI* Vasquez-Gomez et al. (2014) and our proposed *Proximity Count VI* formulation perform significantly better in surface completion over the first five views of an object than the other VIs. The *Average Entropy VI* Kriegel et al. (2015) also performs well, but not as consistently in the first views. The total map entropy plots also illustrate how that metric is not well correlated with surface completion

Table 2 Timing results for each VI formulation, averaged over all trials, models, and views in the simulated experiments. Differences in computational cost are negligible across all of the proposed VIs and the state of the art

Volumetric information	Avg. time per view [s]
Occlusion Aware	3.7798
Unobserved Voxel	3.7786
Rear Side Voxel	3.7709
Rear Side Entropy	3.7889
Proximity Count	3.7756
Area Factor (Vasquez-Gomez et al. 2014)	3.7618
Average Entropy (Kriegel et al. 2015)	3.7714

base the IG formulation on the difference in the expected percentages α_{oc} and α_{op}:

$$\mathcal{G}_{v,Vasquez}(v) = f(\alpha_{oc}, \alpha_{des,oc}) + f(\alpha_{op}, \alpha_{des,op}) \tag{17}$$

with

$$f(\alpha, \alpha_{des}) = \begin{cases} h_1(\alpha, \alpha_{des}) & \text{if } \alpha \leq \alpha_{des} \\ h_2(\alpha, \alpha_{des}) & \text{if } \alpha > \alpha_{des} \end{cases} \tag{18}$$

where

$$h_1(\alpha, \alpha_{des}) = -\frac{2}{\alpha_{des}^3}\alpha^3 + \frac{3}{\alpha_{des}^2}\alpha^2 \tag{19}$$

and

$$\begin{aligned} h_2(\alpha, \alpha_{des}) = &-\frac{2}{(\alpha_{des}-1)^3}\alpha^3 + \frac{3(\alpha_{des}+1)}{(\alpha_{des}-1)^3}\alpha^2 \\ &-\frac{6\alpha_{des}}{(\alpha_{des}-1)^3}\alpha + \frac{3\alpha_{des}-1}{(\alpha_{des}-1)^3} \end{aligned} \tag{20}$$

$f(\cdot)$ is equal to one if the estimated percentage matches the desired percentage. This formulation is referred to as *Area Factor VI*.

We also compare all of the VI formulations to a next best view planner that chooses randomly from the available views in the view space at each iteration.

5 Discussion

Considering our target scenario—3D reconstruction of a bounded object by a mobile robot with a camera-based depth sensor— we evaluated the proposed and state of the art volumetric information formulations based on their ability to choose views that efficiently lead to a complete object model. Based on our simulated experiments, the *Area Factor VI* proposed by Vasquez-Gomez et al. in Vasquez-Gomez et al. (2014) and our proposed *Proximity Count VI* exhibit superior performance during the first few views of the object. Using the Area Under the Curve as a metric for the efficiency of the reconstruction over the full trial, both of these VIs also outperform the other formulations, performing best on 4 of the 11 models each. However, other VIs typically achieve comparable or superior surface coverage after 4–6 views, and beyond 6–10 views, depending on the model, the surface completion is asymptotic.

Choosing random views performs reasonably well when averaged over many trials, but the variance is much larger than any of the proposed or state of the art VIs. While this approach would be computationally less expensive than evaluating all of the candidate views for information gain, the computation time per view in our tests was small enough ($<< 10$ s) that taking additional random views would not be more efficient based on the time for most mobile robots to move between views.

The *Average Entropy* formulation from Kriegel et al. (2015) is most effective at reducing the entropy in the map, achieving the best AUC for most of the models. However, measuring the total entropy reduction in the map does not discriminate between the VIs very well, based on our trials. The ability of a VI formulation to effectively reduce the entropy in the map is not well correlated with its effectiveness in surface reconstruction. Indeed, this conclusion is supported by the update procedure for a volumetric representation like ours. Since we must observe the same voxel multiple times to increase our certainty about its occupancy, and therefore decrease its entropy, we would therefore fail to optimize our observation of new regions of the object. Additionally, for many robotic applications such as the estimation of grasping affordances, map entropy would be much less informative than surface completion in completing the task. However, for scenarios in which the certainty about the occupancy of the volumes is more important than completeness, *Average Entropy* is more effective than any of the other VIs, but the results are not very conclusive.

Within our software framework, computational cost is negligibly different between all of the proposed and state of the art VI formulations. Consequently, the efficiency of the reconstruction is dependent primarily on the number of views required to reconstruct the object.

Based on the results of our experiments, in which we isolated the choice of formulation for volumetric information as the primary independent variable, the performance of the Vasquez-Gomez et al. *Area Factor VI* and our proposed *Proximity Count VI* make them the best choices for efficient reconstruction.

6 Conclusion

In this work, we have considered the problem of next-best view selection for 3D reconstruction by a mobile robot equipped with a camera, where the robot builds a probabilistic map in real time, and quantifies the expected information gain from a set of discrete candidate views. We proposed several formulations to quantify this information gain for the volumetric reconstruction task, including visibility likelihood and the likelihood of seeing new parts of an object when performing volumetric reconstruction. The next best view is selected by optimizing the expected information gain over the candidate views of the object.

We evaluated these formulations with extensive simulated experiments in order to asses the contribution that each VI formulation makes in the performance at the reconstruction task. Due to the use of an uncluttered scene containing only the object, and an idealized sensor with no uncertainty in its measurements or position, our experiments isolated the performance of each VI formulation, without effects from environmental factors. The results of the experiments indicate that our proposed *Proximity Count VI* and the *Area Factor VI* from Vasquez-Gomez et al. (2014) both provide comparably high levels of reconstruction completeness during the first few views, as well as over the course of the whole procedure, on a set of models with a variety of shapes and degrees of complexity and convexity. However, the experimental results also showed that in most cases, the reconstruction is able to achieve most of its model completion within a small number (<10) of well-chosen views, regardless of the choice of VI formulation.

Our active reconstruction framework is adaptable to other robotic platforms and reconstruction problems, and has been released open source. The software and videos demonstrating its performance are available at: http://rpg.ifi.uzh.ch.

References

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, *1*(4), 333–356.

Bajcsy, R. (1988). Active Perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

Banta, J., Wong, L., Dumont, C., & Abidi, M. (2000). A next-best-view system for autonomous 3-D object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(5), 589–598.

Banta, J. E., Zhien, Y., Wang, X. Z., Zhang, G., Smith, M. T., & Abidi, M. A. (1995). Best-next-view algorithm for three-dimensional scene reconstruction using range images. In *Proceedings of the SPIE 2588, intelligent robots and computer vision XIV: Algorithms, techniques, active vision, and materials handling*, 418 (October 3, 1995). doi:10.1117/12.222691.

Blake, A., & Yuille, A. (1988). *Active vision*. Cambridge: The MIT Press.

Chen, S., & Li, Y. (2005). Vision sensor planning for 3-D model acquisition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B Cybernetics*, *35*(5), 894–904.

Chen, S., Li, Y., & Kwok, N. M. (2011). Active vision in robotic systems: A survey of recent developments. *International Journal of Robotics Research*, *30*(11), 1343–1377.

Connolly, C., et al. (1985). The determination of next best views. In *IEEE international conference on robotics and automation (ICRA)* (Vol. 2, pp. 432–435). IEEE.

Forster, C., Pizzoli, M., & Scaramuzza, D. (2014). Appearance-based active, monocular, dense depth estimation for micro aerial vehicles. In *Robotics: Science and Systems (RSS)*.

Hornung, A., Wurm, K. M., Bennewitz, M., Stachniss, C., & Burgard, W. (2013). OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*,. doi:10.1007/s10514-012-9321-0. http://octomap.github.com. Software http://octomap.github.com.

Isler, S., Sabzevari, R., Delmerico, J., & Scaramuzza, D. (2016). An information gain formulation for active volumetric 3d reconstruction. In *IEEE international conference on robotics and automation (ICRA)*.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., & s H., A. (2014). Large scale multi-view stereopsis evaluation. In *Proceedings of IEEE international conference on computer vision and pattern recognition*.

Krainin, M., Curless, B., & Fox, D. (2011). Autonomous generation of complete 3d object models using next best view manipulation planning. In *IEEE international conference on robotics and automation (ICRA)* (pp. 5031–5037). IEEE.

Kriegel, S., Rink, C., Bodenmüller, T., & Suppa, M. (2015). Efficient next-best-scan planning for autonomous 3d surface reconstruction of unknown objects. *Journal of Real-Time Image Processing*, *10*, 611. doi:10.1007/s11554-013-0386-6.

Pito, R. (1999). A solution to the next best view problem for automated surface acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(10), 1016–1030.

Potthast, C., & Sukhatme, G. S. (2014). A probabilistic framework for next best view estimation in a cluttered environment. *Journal of Visual Communication and Image Representation*, *25*(1), 148–164.

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., & Ng, A. Y. (2009). ROS: An open-source robot operating system. In *ICRA workshop on open source software* (Vol. 3, p. 5).

Rodolà, E., Albarelli, A., Bergamasco, F., & Torsello, A. (2013). A scale independent selection process for 3d object recognition in cluttered scenes. *International Journal of Computer Vision*, *102*(1–3), 129–145.

Schmid, K., Hirschmüller, H., Dömel, A., Grixa, I., Suppa, M., & Hirzinger, G. (2012). View planning for multi-view stereo 3D reconstruction using an autonomous multicopter. *Journal of Intelligent and Robotic Systems*, *65*(1–4), 309–323.

Scott, W., Roth, G., & Rivest, J. F. (2003). View planning for automated 3d object reconstruction and inspection. *ACM Computing Surveys*, *35*(1), 64–96.

Torabi, L., & Gupta, K. (2012). An autonomous six-DOF eye-in-hand system for in situ 3D object modeling. *International Journal of Robotics Research*, *31*(1), 82–100.

Trummer, M., Munkelt, C., & Denzler, J. (2010). Online next-best-view planning for accuracy optimization using an extended e-criterion. In *International conference on pattern recognition (ICPR)* (pp. 1642–1645). IEEE.

Vasquez-Gomez, J. I., Sucar, L. E., & Murrieta-Cid, R. (2014). View planning for 3d object reconstruction with a mobile manipulator robot. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*.

Vasquez-Gomez, J. I., Sucar, L. E., Murrieta-Cid, R., & Lopez-Damian, E. (2014). Volumetric next best view planning for 3d object reconstruction with positioning error. *International Journal of Advanced Robotic Systems*, *11*, 159.

Wettach, J., & Berns, K. (2010). Dynamic frontier based exploration with a mobile indoor robot. In *international symposium on Robotics (ISR)* (pp. 1–8). VDE.

Yamauchi, B. (1997). A frontier-based approach for autonomous exploration. In *IEEE international conference on robotics and automation (ICRA)* (pp. 146–151). IEEE.

Jeffrey Delmerico is a postdoctoral researcher at the Robotics and Perception Group at the University of Zurich. He received his Ph.D. in 2013 from the State University of New York at Buffalo (with Jason Corso), and was a postdoc at the University of Hawaii at Manoa before joining the Robotics and Perception Group.

Stefan Isler received his M.Sc. degree in Mechanical Engineering from ETH Zurich in 2016. He is now a computer vision engineer for Insightness IG.

Reza Sabzevari was a postdoctoral researcher at the Robotics and Perception Group at the University of Zurich and is now with Robert Bosch GmbH Corporate Research. He received his Ph.D. from the Italian Institute of Technology in Genoa, in 2013.

Davide Scaramuzza is Assistant Professor of Robotics at the University of Zurich and head of the Robotics and Perception Group. He received his Ph.D. (2008) in Robotics and Computer Vision at ETH Zurich (with Roland Siegwart). He was Postdoc at both ETH Zurich and the University of Pennsylvania (with Vijay Kumar and Kostas Daniilidis).