Active View Selection for Efficient 3D Scene Reconstruction

Vítor Sequeira, João G. M. Gonçalves, M. Isabel Ribeiro(*)
European Commission, Joint Research Centre, Italy
(*)Instituto Superior Técnico/Institute for Systems and Robotics, Portugal

Abstract

This paper presents an active sensing process for the efficient reconstruction of a real world environment using multiple views of a scene captured by a laser range finder onboard a mobile robot. To resolve the ambiguities caused by occlusions there is the need to acquire multiple range views at different viewpoints. A strategy for optimising the selection of the next sensing position and orientation based on previously detected occlusions is proposed. No a priori knowledge about the scene is used for planning the next views. The system evaluates the set of capture points from which all the necessary data can be acquired aiming at minimising its number. Among these points it then selects the one(s) maximising the occluded areas to be resolved.

Keywords: *laser range images, active sensing, perception planning, multiple views, CAD-based vision.*

1. Introduction

Single images often contain occluded regions, being insufficient to reconstruct a complete 3D model of a scene. This model can be obtained by fusing multiple range views, stressing the requirement for mobility in range acquisition systems. Efficient reconstruction can be obtained by controlling the relationship between the occluded surfaces and the acquisition system. This strongly points out for an active data gathering strategy.

Although several sensor planning systems have been developed for automatic camera and light source placement to meet the requirements of robot vision tasks [1-5], or for planning the focus of attention to perform visual feedback control tasks (e.g., grasp an object) [6], less work has been done for 3D scene reconstruction [2,10].

This paper proposes a strategy for optimising (i.e., minimising the number of capture points, maximising the occlusion areas to be resolved and selecting the best acquisition conditions) the planning of the next view based on occlusions. The system relies only on the acquired data. A first range image is acquired from an arbitrary or pre-defined capture point. The next capture points are estimated based on the model already extracted from previous images. Fig. 1 is an overview of the iterative procedure for 3D reconstruction of static environments.

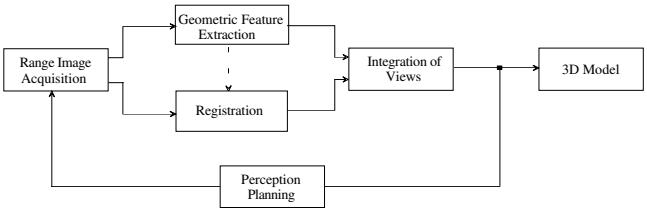

Fig. 1: Iterative procedure for 3D scene reconstruction.

Details on the 3D scene reconstruction (i.e., geometric feature extraction, registration and integration of multiple views) can be found in [7,8].

The paper is organised as follows. Section 2 describes the active strategy for planning the next view based on the extraction of the occlusions and Section 3 presents experimental results. Finally, in Section 4 conclusions are drawn and work in progress is outlined.

2. Perception Planning

The problem of determining the next capture point, defined as the 3D point (x,y,z) and corresponding 3D orientations from where data will be acquired, depends on how much *a priori* geometrical information about the scene is available.

Some works tackled the problem of sensor positioning when the fully environment is known [3,4], for 3D object recognition [9], for the reconstruction of unknown small objects [2,10,11,12].

Two issues are to consider when selecting the next capture point for the 3D reconstruction of an unknown environment (no *a priori* geometrical information about the scene is available): i) resolving occlusions in the already extracted model, ii) acquiring additional views of the non-modelled environment. These are required to have overlapping regions with the previously extracted model, aiming at being integrated with it. There is hence a need to detect the occlusions, and to evaluate the next most appropriate capture point(s) as well as the associated scanning parameters. After resolving all the

e-mail: vitor.sequeira@jrc.it; joao.goncalves@jrc.it; r@isr.ist.utl.pt

occlusions in the extracted model, the system plans additional views of the non-modelled environment by following the surfaces on the boundary of the reconstructed model.

2.1 Occlusion Detection

Occlusions are detected based on surface discontinuities (jump edges) [7]. A relationship between the jump edge and the far away surface from the sensor is established providing boundary limits for both adjacent surfaces along the jump edge. These boundary limits define the occlusion plane (see Fig. 2). In the case of curved jump edges, they are broken into line edges on its maximum curvature.

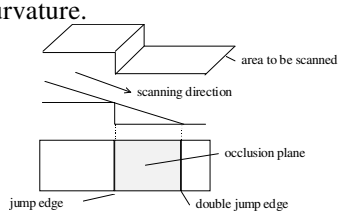

Fig. 2: Determining the occlusion plane.

On complex indoor environments a significant number of occlusions may occur, leading to a large number of occlusion planes and thus a large number of potential capture points.

2.2 Planning the Next View

To select the next optimal capture point(s), the perception planning system first evaluates the set of capture points from which each occlusion area is completely visible. These initial set of points are then combined aiming at minimising its number while keeping all the occlusions areas fully visible. The procedure then selects the one(s) maximising an objective function that accounts for the acquisition conditions.

2.2.1 Potential Capture Points

Initially, the system reconstructs all scene occlusion planes based on the jump edges and their projections (double jump edge) as seen in Fig. 2. For each of those planes, the set of possible capture points that will resolve the occlusion is evaluated. The whole procedure works on 3D space, but for visualisation and clarity purposes some of the next figures illustrate the method in a 2D representation. This is the case of Fig. 3 that shows the 2D set of initial capture points for resolving a single occlusion.

For each occluded plane a viewing direction passing in the centre of the occlusion area (occlusion point), and normal to it is selected. From this initial viewing direction (axis Y′) a set of possible viewing segments[1] constrained to the following conditions is evaluated:

1. The initial viewing direction, Y′, is reduced to its included segment (viewing segment) defined by the occlusion point and the closest previously extracted surface that it intersects.

2. The distance between any point on the viewing segment and any already extracted surface must be greater than a safety operational distance, d_s.

3. The viewing position must be within the sensor measurement interval ($d_{min} < d < d_{max}$).

4. The angle α between the initial viewing direction along *Y'* and the occlusion plane should belong to the interval]-90;90[. For each of the directions in the referred interval, the corresponding viewing segments are evaluated along the previous constraints.

5. All the occlusion area must be visible.

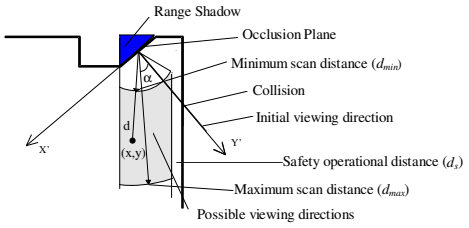

Fig. 3: Possible viewing position constraints.

These previous placement constraints are automatically combined to produce a set of candidate capture points, a 3D viewing volume where the scanning device can be placed, which is then approximated by a polyhedron in 3D denoted by *R*. This is represented by the light shaded area in Fig. 3.

Up till now the method has concentrated on the determination of a viewing volume (set of capture points) fulfilling the associated constraints for each occlusion. The next step in the perception planning scheme is to consider all the occlusions and to reduce the number of the volumes from which all the occlusions can be recovered. For doing this, the viewing polyhedrons from each occlusion are intersected and the global viewing volumes minimising the number of capture points determined. The global viewing volume(s) results from the global intersection, if non-empty (Fig. 4), or from several partial intersections otherwise.

Fig. 4 is an example of the procedure for planning the next set of views, when the current description of the environment has detected three occlusion regions A_1, A_2, and A_3. Fig. 4a), b) and c) show the possible viewing area (set of possible capture points) R_1, R_2, and R_3, associated with each occlusion region. Fig. 4d) displays the final viewing area, a single global viewing volume, resulting from the intersection of the three viewing areas, $R_1 \cap R_2 \cap R_3$. In this example the viewing area resulting

[1] It is considered that all the points in a viewing segment are potential capture points candidates.

from the global intersection is non-empty and thus a single capture point inside $R_1 \cap R_2 \cap R_3$ is enough to resolve all the three occlusions.

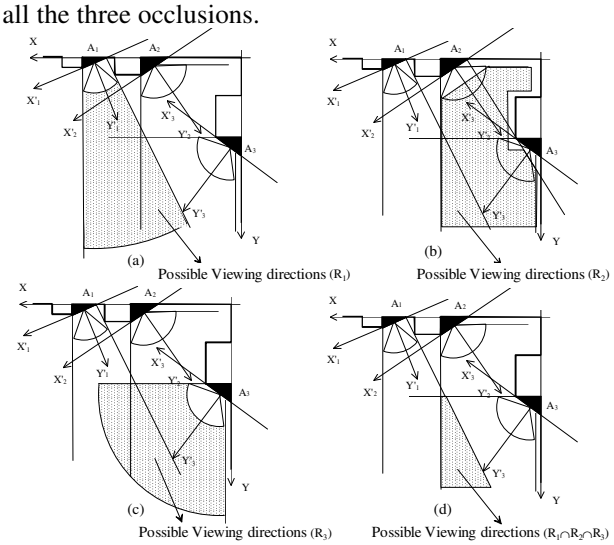

Fig. 4: Possible viewing directions to resolve 3 occlusions.

As the number of occlusion areas increases, along with the complexity of the objects' 3D shapes (e.g., curved objects, non-uniform shapes) and working environments (e.g., objects in a middle of a room), the complexity of determining the possible global volumes increases rapidly.

The problem of determining the final set of global viewing volumes aiming at minimising the number of capture points can be formulated as follows:

- Let R_i, $i = 1,\ldots,M$ be the 3D viewing volume from which occlusion i is resolved.
- Find global viewing volumes, V_k, such that for every viewing area R_i, there exists a region V_k such that

$$R_i \cap V_k \neq \varnothing, \qquad k = 1,\ldots,n_V \tag{1}$$

and their number, n_v, is minimal. The number of global viewing volumes, n_v, determines the final number of capture points from where it is possible to resolve all the occlusions. By minimising the number of capture points, one minimises the robot movements required to positioning the sensor devices. This will reduce the time needed for acquiring all the unknown data and thus the time needed to perform a full 3D reconstruction.

When there are multiple solutions of the intersection (1) for the same n_v, the choice among them is supported on the evaluation of an objective function that will be introduced in the next sub-section.

2.2.2 Objective Function

In the previous section the minimum number of capture points, n_v, that resolve all the occlusions was determined. At this stage, a criterion for selecting the "best" capture point within each global viewing area V_k, k=1,...,n_v must be defined. In the present case, "best" stands for enabling the reconstruction of the 3D model of environments as complete and accurate as possible.

The precision of laser range data depends strongly on data acquisition conditions. A main limitation is due to the fact that the laser beam is not an ideal line, but rather a cylinder. The distance information of each point on the 3D image is an integration of range values over a 2D area, resulting from the intersection of the cylinder beam with the target surface, the laser footprint. Moreover, the size of the footprint, and thus the measurement accuracy, also depends on the angle between the surface normal and the beam. In conclusion, the "best" (higher precision) position to view a surface is in its surface normal direction. Another issue is that measurement accuracy depends on the distance to the target [13].

An objective function was thus derived favouring measurements done near the surface normal direction, and penalising measurements done out of the best distance precision. This last factor depends very much on the type of Laser Range Finder (LRF) used. Some LRFs have a beam that is always divergent, some others have the beam focused to a distance different from the minimum, i.e., first the beam converges up to a given distance, and then it diverges. Finally, in the presence of several occluded surfaces, an higher priority is given to the larger surface.

Thus, to evaluate the position of the next n_V capture points, $(x_{c_k}, y_{c_k}, z_{c_k})$, the perception planning selects from each V_k the point that maximises the objective function:

$$F_k(x,y,z) = \sum_{i=1}^{M_k} F_i(x,y,z), \quad (x,y,z) \in V_k, \quad k = 1,\ldots,n_v \tag{2}$$

with $\sum_{k=1}^{n_v} M_k = M$, where M_k is the number of occlusions that are resolved with the global viewing volume V_k and M represents the total number of occlusions. In (2),

$$F_i(x,y,z) = A_i \cdot \eta_i(x,y,z) \cdot \varepsilon_i(x,y,z), \tag{3}$$

where A_i is the area of the i^{th} occlusion resolved with V_k at the current position (x,y,z) (benefit),

$$\eta_i(x,y,z) = \frac{1}{\sigma_\alpha \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\alpha_i(x',y',z')}{\sigma_\alpha}\right)^2}, \tag{4}$$

is the component of the cost function associated with $\alpha_i(x',y',z')$, the angle between the viewing direction (**V**) and the occlusion plane as represented in Fig. 5, where

$$\begin{cases} x' = l_1(x-x_i) + m_1(y-y_i) + n_1(z-z_i) \\ y' = l_2(x-x_i) + m_2(y-y_i) + n_2(z-z_i) \\ z' = l_3(x-x_i) + m_3(y-y_i) + n_3(z-z_i) \end{cases}. \tag{5}$$

In (5), $l_1,l_2,l_3;m_1,m_2,m_3;n_1,n_2,n_3$ are the direction cosines of the (X'_i,Y'_i,Z'_i) axes relative to the (X,Y,Z) axes respectively. Finally

$$\varepsilon_i(x,y,z)=\frac{1}{\sigma_{lrf}\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{d_i(x,y,z)-d_c}{\sigma_{lrf}}\right)^2} \quad (6)$$

is the cost function (Normal density) associated with $d_i(x,y,z)$, the distance between the acquisition point and the origin of the occlusion co-ordinate system at the occlusion point (see Fig. 5). In (6), d_c is the best (higher resolution) acquisition distance and σ_{lrf} is associated with the accuracy of the range measurements. Both values were experimentally evaluated in [13].

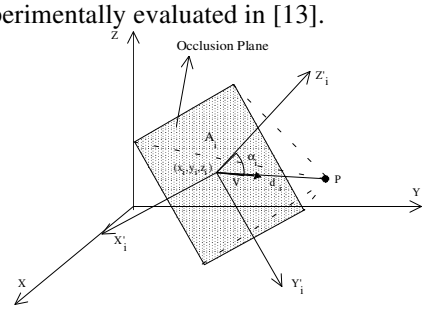

Fig. 5: Occlusion plane.

For each k, the result of the optimisation process in (2) is not necessarily unique, depending largely on the type of environment and on the constraints imposed by it. In case of more than one solution, (n_r solutions), the final solution is the one minimising the maximum angle $\alpha_i(x',y',z')$ in (4):

$$\min_r\left[\max_i\left(\alpha_i(x',y',z')\right)\right], \quad i=1,\ldots,M_k, \quad r=1,\ldots,n_r \quad (7)$$

If there is more than one solution of the intersection in (1) for the same number, n_v, of global volumes, there is the need to choose the final solution. Fig. 6 illustrates a case with three viewing volumes (R_1,R_2,R_3) whose global intersection is empty and a solution is found for n_v=2. In this case, two sets n_m=2 of solutions (V_{11},V_{21}) and (V_{12},V_{22}) were found.

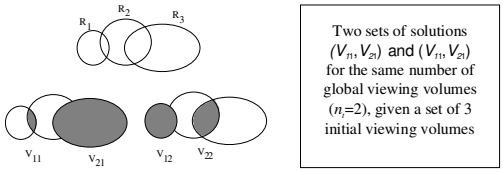

Fig. 6: Sets of solutions.

The final solution is the one that maximises the sum of the maximum values of the objective function for each solution[2],

$$\max_m \sum_{k=1}^{n_t} F_{km}(x_{c_{km}},y_{c_{km}},z_{c_{km}}), \quad m=1,\ldots,n_m, \quad (8)$$

[2] Note that F_{km}= F_m and V_{km}= V_m for n_m =1.

with $(x_{c_{km}},y_{c_{km}},z_{c_{mk}})\in V_{km}$, $k=1,\ldots,n_t$. In (8) n_t is the number of capture points that are being disambiguated, n_m is the number of multiple solutions for a particular n_t, and $(x_{c_{km}},y_{c_{km}},z_{c_{km}})$ is the capture point that gives a global maximum for F_{km}.

The objective function was used to determine the final solution for each of the n_m possible solutions and thus (8) automatically leads to the final corresponding capture points.

After computing the next capture point(s) the system automatically selects all the scanning acquisition parameters, namely the image field of view (scanning angle in both the horizontal and vertical directions), the area to be covered in one scan and the image resolution, i.e., the angular distance between two consecutive samples. These parameters are chosen aiming at having enough overlapping between views to allow registration. If the areas to be scanned are sufficiently apart, i.e., if the angle between the limits of two scanning areas are higher than a certain threshold, or if the field of view is higher than the maximum allowed by the scanning mechanism, several views are taking from the same capture point in order to cover all the area to be recovered.

3. Experimental Results

The algorithm was tested on a large number of different real-world scenes, with different levels of complexity. The range images were acquired using a time-of-flight LRF mounted on a pan and tilt unit (with two rotational degrees of freedom) onboard of a mobile robot (see Fig. 7). The first image is acquired from an arbitrary or pre-defined capture point. The next capture points are estimated based on the information extracted in the previous images, along the lines described in Sec. 2.

Fig. 8 shows the results for a complex scene taken from a laboratory environment with a large number of objects with different shape complexity and surface types (e.g., planar and convex or concave curved surfaces). This leads to a large number of occlusions (Fig. 8a), and thus to a large number of different initial viewing directions. In this case, the algorithm yields three different positions (P_1, P_2, P_3 in Fig. 8b) from where the detected occluded areas can be resolved. The region fusing procedure enables the elimination of occluded areas (or areas with noisy/ambiguous data in the raw range image) in the 3D model extracted from a single range image (Fig. 8a) by fusing another models extracted from the range images of the occluded areas (Fig. 8c), into a single model (Fig. 8d). The models in Fig. 8c were extracted from the range images acquired from the three capture points (P_1, P_2, P_3) displayed in Fig. 8b. The developed integration scheme [7] is powerful in that the

combination of views may be done at the level of the final representation.

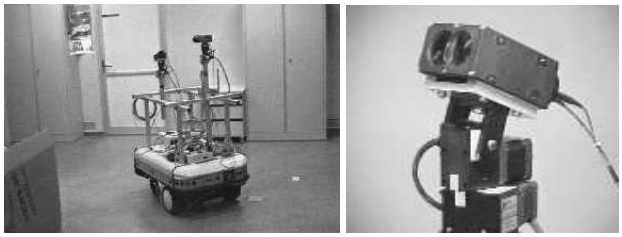

Fig. 7: LRF on board of the mobile robot.

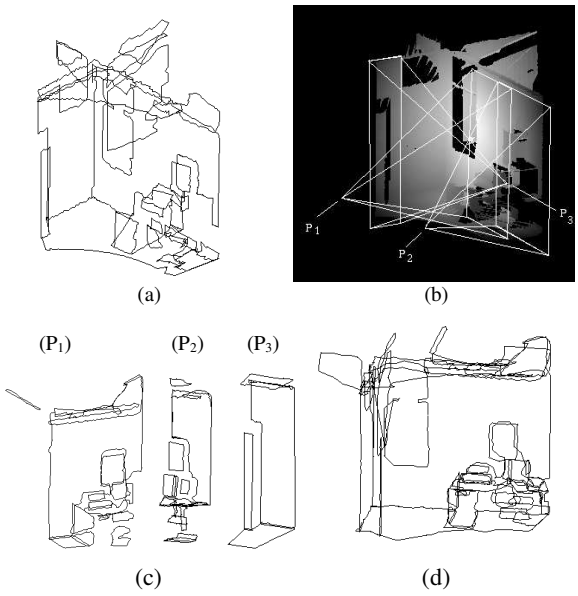

Fig. 8: Next capture points (3) for a complex scene.

4. Conclusions and Work in Progress

A strategy for the planning of the next view based on occlusions was developed using an objective function. The motivation for the objective function comes from the fact that there are some locations where data can be acquired with higher precision (e.g., the best acquisition position is perpendicular to the surface being scan) and also to minimise the number of capture points, maximising the occlusion areas to be resolved and to optimise the acquisition conditions.

The range sensor must enter the scene to collect data, and thus the algorithm for planning the next view takes into account the environment that is incrementally being built and the associated constraints: topological (imposed by the objects being scanned) and operational (imposed by the already reconstructed environment and by the acquisition system). The system operates autonomously, with no a priori knowledge about the scene to be reconstructed and with no human intervention.

The characteristics of the cost function for planning the next view depends strongly on the type of environment and on the constraints imposed by it. It is not possible to a priori determine its characteristics, but it can happen that the presence of a single global optimum is surrounded by multiple local minima. As a consequence, to ensure successful minimisation for many different types of complex environments the conventional gradient descent approach is not a fully robust search method. Current work is being done to use very fast simulated re-annealing (VFSR) to minimise the cost function.

5. Acknowledgements

V. Sequeira, PhD student of Instituto Superior Técnico, acknowledges the PRAXIS XXI program from JNICT, Portugal, for his PhD grant. This research was partially funded by ACTS program of the European Union (RESOLV project).

6. References

[1] C. K. Cowan, P. D. Kovesi, Automatic Sensor Placement from Vision Requirement, IEEE-PAMI-10 (2), 1988, pp. 407-416.

[2] J. Maver, R. Bajcsy, Occlusions as a Guide for Planning theNext View, IEEE-PAMI-15(5), May 1993, pp. 417-433.

[3] S. Sakane, T. Sato, Automatic Planning of Camera and Light Source Placement for an Active Photometric Stereo System, Proc. IEEE ICRA'91, pp.1080-1087.

[4] K. Tarabanis, R.Y. Tsai, Viewpoint Planning: the visibility constraint, Proc. Image Underst. Ws., 1989, pp. 893-903.

[5] S. Yi, R. M. Haralick, L. G. Shapiro, Automatic Sensor and Light Source Pos. for Machine Vision, Proc. ICPR'90, pp.54-59.

[6] S. Sakane, T. Koruma, T. Omata, Planning Focus of Attention for Multifingered Hand with Consideration of Time-Varying Aspects, Proc IEEE CAD-Based Vision Ws.,1994, pp.151-160.

[7] V. Sequeira, J.G.M. Gonçalves, M.I. Ribeiro, High-Level Surface Descriptions from Composite Range Images, in Proc. IEEE ISCV'95, pp. 163-168.

[8] V. Sequeira, J.G.M. Gonçalves, M.I. Ribeiro, 3D Environment Modelling Using Laser Range Sensing, in J. Robotics Auton. Systems - Elsevier, 16(1), Nov. 1995, pp. 114-127.

[9] S.A. Hutchinson, A.C. Kak, Planning Sensing Stategies in a Robot Work Cell with Multi-Sensor Capabilities, IEEE Trans. on Robotics and Aut., 5(6), pp. 765-783, 1989.

[10] C.I. Connolly, The Determination of Next Best Views, Proc. IEEE ICRA'85, pp. 432-435.

[11] P. Whaite, F.P. Ferrie, From Uncertainty to Visual Exploration, IEEE-PAMI-13 (10), 1991, pp. 1038-1049.

[12] P. Whaite, F.P. Ferrie, Active Exploration: Knowing When We're Wrong, IEEE-ICCV'93, May 1993, pp. 41-48.

[13] V. Sequeira, Active Range Sensing for Three-Dimensional Environment Reconstruction, PhD Thesis, to be submitted to Instituto Superior Técnico, Portugal.