A next-best-view algorithm for 3D scene recovery with 5 degrees of freedom

J.M. Sanchiz †, R.B. Fisher §
†§Division of Informatics, Institute of Perception, Action and Behaviour
University of Edinburgh, 5 Forrest Hill, Edinburgh EH1 2QL, UK
†Dept d'Informàtica, Universitat Jaume I, 12071 Castelló, Spain
sanchiz@inf.uji.es, rbf@dai.ed.ac.uk

Abstract

We present an algorithm for determining the next best position of a range sensor in 3D space for incrementally recovering an indoor scene. The method works in five dimensions: the sensor navigates inside the scene, and can be placed at any 3D position and oriented by a pan-tilt head. The method is based on a mixed exhaustive search and hill climbing optimisation, and outputs the next position in reasonable time. Results are shown on a simulated mobile robot with a simulated range sensor navigating in a CAD model of a scene.

1 Introduction

This paper describes an improved method of computing the next best position for a complete and accurate three-dimensional recovery of an unknown indoor scene. We consider a mobile robot equipped with a range sensor, and we compute the next position and pose to place the sensor for taking the next view. Different views are registered, building up an incremental scene model. The problem addressed in this paper is sensor placement: finding the next view that would best improve the current recovered scene model. In general 3D motion this is a problem with six degrees of freedom, since the sensor can be placed anywhere in space, and can be oriented by three rotations: pan (rotated around the vertical axis), tilt (rotated around the horizontal axis), and roll (rotated around the optical axis).

Sensor rolling can be useful because the horizontal and vertical apertures of the sensor may be different, so different scene areas can be recovered by rolling the sensor. But in this work we assume that rolling is not allowed because conventional pan-tilt heads do not perform this motion. So we can think of a mobile base which moves over a floor, with a vertical bar to lift or lower the sensor, which is mounted on a pan-tilt head at the extreme of the bar. This makes the problem of finding the next best view a five-dimensional one.

Other research has addressed the next best view problem for object reconstruction where the outsides of objects are seen, but not for 3D scene recovery, where the inside of a scene is explored and the sensor can navigate into the scene. Massios

and Fisher [1] compute the next best position of a range sensor for object acquisition with orthogonal projection. The sensor position space was two-dimensional: a sphere at a fixed radius enclosing the object with the sensor always pointing to the centre. As in our approach they defined a quality criterion and used a voxel map for view reasoning. García et al. [2] also addressed a similar 2D sensor space, although discretised in a different way. They used a voting scheme to compute the next view, maximising the observation of occluded areas, and used a triangular mesh to model the object.

Reed at al. [3] presented a method to recover object models including the computation of the next view to maximise occluded areas. They computed visibility volumes from where occluded areas are fully visible, following the method presented by Tarabanis et al. [4]. But that research did not solve for the best position of the sensor inside the visibility volumes. In that work Reed et al. assumed a 2D sensor position space, an enclosing sphere, and computed its intersection with the viewing volumes. Pito [5] proposed a method for view planning in object modelling with 4 degrees of freedom. The sensor position space was a cylinder, and it could be oriented within a range of pan-tilt angles. He used a voting scheme to maximise the observation of occluded areas.

In summary, the question that this paper addresses is: **Is it possible to define an effective and efficient algorithm for scanning an environment such that all surfaces are observed with high quality measurements ?** The results shown have answer the question positively, and we believe this is the first implemented algorithm to do so.

2 Scene representation

Our objective is not to recover a surface model of the scene, which has been done elsewhere by triangulation of the sensed 3D points [6] (including texture information to make it more realistic). Instead, we aim at finding the set of sensor poses that best acquire the 3D points according to some criteria (explained later). Our recovered scene model has to be accurate enough to compute these criteria.

We use a voxel map representation. Voxels, volume elements, are small cubes of a fixed size. The voxel map is a 3D rectangle whose size depends on the available memory, size of the scene to be modelled, and resolution at which we work. The voxel map is implemented as a 3D circular buffer, and it can be placed anywhere in space, so that if new 3D points are sensed that do not fit in the voxel map, new space can be allocated for the new area without moving data in memory.

A voxel map representation allows ray tracing by a 3D Bresenham algorithm [7] using only integer operations. It also allows a straightforward registration of the new sensed points with the recovered scene [1] (voxel map update) just by assuming that the voxel size is bigger than the errors that may arise in the sensed point positions due to inaccurate sensor placement (navigation errors).

In our scene model a voxel consists of a *label* indicating its type, a *surface normal*, and a *quality*, indicating how accurately this voxel has been sensed so far. The voxel labels include:

- *Unmarked* voxel. A voxel that has never been observed by the sensor.

- *Empty* voxel. A voxel that has been observed and found to be empty.
- *Occupied* voxel. A voxel in which 3D sensed points have fallen.
- *Occluded* voxel. A voxel so far occluded by an occupied voxel.
- *Occlusion Plane* voxel. A special kind of *Occluded* voxel adjacent with an *Empty* voxel through any of its six faces.

A voxel's *surface normal* and *quality* are only defined for *Occupied* voxels. Normals are estimated at every point in a range image. The voxel *surface normal* is the average of the normals of all the range points that have updated this voxel. The *sensed quality* of an *Occupied* voxel is the cosine of the angle formed by the *surface normal* and the viewing ray. The voxel *quality* is the best *sensed quality* of the voxel so far.

3 Voxel map update

Every time a new range image is taken the voxel map is updated, incrementing our knowledge of the scene.

A range image is a matrix $[d_{uv}]$ ($u \in [0..N-1]$, $v \in [0..M-1]$) of distances sensed in the direction $\vec{n}_{uv\phi\theta}$ from the optical centre c, (ϕ, θ) being the pan-tilt angles the sensor is oriented. Representing a 3D rotation by $\mathbf{R}^{axis}_{angle}$, $\vec{n}_{uv\phi\theta}$ are computed as:

$$\begin{cases} \phi_u = \beta(\frac{1}{2} - \frac{u}{M-1}) \\ \theta_v = \alpha(\frac{1}{2} - \frac{v}{N-1}) \\ \vec{a} = \mathbf{R}^{\vec{x}}_{\phi_u}(0,0,1)' \\ \vec{n}_{uv00} = \mathbf{R}^{\vec{a}}_{\phi_u}\mathbf{R}^{\vec{x}}_{\theta_v}(0,1,0)' \\ \vec{n}_{uv\phi\theta} = \mathbf{R}^{\vec{y}}_{\phi_u}\mathbf{R}^{\vec{x}}_{\theta_v}\vec{n}_{uv00} \end{cases} \tag{1}$$

where α and β are the horizontal and vertical angular apertures. The 3D co-ordinates of a sensed point p_{uv} are:

$$p_{uv} = c + d_{uv}\vec{n}_{uv\phi\theta} \quad . \tag{2}$$

A voxel is identified by its three indices in the voxel map $(i,j,k)'$. The centre of voxel $(0,0,0)'$ is placed in space at world co-ordinates $(x_w, y_w, z_w)'$ and the voxel array is aligned with the world co-ordinate axes. A voxel is a cube of side *scale*, so a point in 3D space $(x,y,z)'$ falls inside voxel $(RoundScale(x-x_w), RoundScale(y-y_w), RoundScale(z-z_w))'$, where $RoundScale(x) = Round(\frac{x+sgn(x)\frac{scale}{2}}{scale})$.

The voxel map is updated with the following method:

- Compute the camera position in the voxel map $c^{voxel} = RoundScale(c)$.
- For every range image point p_{uv} compute its voxel co-ordinates $p^{voxel}_{uv} = RoundScale(p_{uv})$ using (1–2).
  - Compute the intersection l^{voxel}_{uv} of the ray $c + \lambda\vec{n}_{uv\phi\theta}$ with the limits of the voxel map in voxel co-ordinates.

- Do ray tracing from c^{voxel} to p_{uv}^{voxel} marking as *Empty* the voxels that are not *Occupied*.
- Do ray tracing from p_{uv}^{voxel} to l_{uv}^{voxel} marking as *Occluded* the voxels that are still *Unmarked*.
- Mark voxel p_{uv}^{voxel} as *Occupied*.

- Traverse the voxel map marking as *Occlusion Plane* the voxels that are of type *Occluded* and have a face touching an *Empty* voxel.

4 Fitness function

In order to compute the best next position we have set some criteria for the goodness of a sensor pose, which are formulated as a mathematical function to maximise. The criteria are defined on the scene area that a test position covers, and are: 1) Providing overlap with previously acquired data for fine registration of the data (as wheel slip on the vehicle is likely to introduce dead-reckoning registration errors). 2) Eliminating occlusion plane areas. 3) Observing new unseen areas.

Let a_{ov} be the proportion of overlapping area of the image taken from a certain camera pose, a_{op} be the proportion of occlusion plane area, and let a_{us} be the proportion of unseen area, $a_{ov}, a_{op}, a_{us} \in [0..1]$ and $a_{ov} + a_{op} + a_{us} = 1$. The data for computing these values comes from projecting the current scene model, when viewed from a viewpoint and direction, onto an internal image plane. The function to be maximised has been designed with these characteristics:

- A unique maximum at a certain value of a_{ov} (we have fixed 40% for this value) and for $a_{op} = a_{us}$ (thus favouring at the same time the sensing of occlusion plane and unseen areas).
- Zero at $a_{ov} = 0$, forcing some overlap.
- A fixed value greater than zero at $a_{ov} = 1$, to make possible views with no occlusion planes or unseen areas. This can occur at late stages of the scene recovery, when all parts have been observed and new views aim at increasing the quality of the sensed data.

A simple (polynomial) function that satisfies the above criteria is

$$f_{area} = \left(5a_{ov}^3 - 10.5a_{ov}^2 + 6a_{ov}\right)\left(1 - \frac{1}{2}\left|a_{op} - a_{us}\right|\right) \quad . \tag{3}$$

This function has a maximum of 1.04 at $a_{ov} = 0.4$ and $a_{op} = a_{us}$, a local minimum of 0.5 at $a_{ov} = 1$, value 0 at $a_{ov} = 0$, and decreases as a_{op} differs from a_{us}. In the space defined by axes (a_{ov}, a_{op}, a_{us}), the domain of f_{area} is the triangle defined by the plane $a_{ov} + a_{op} + a_{us} = 1$ and the conditions $a_{ov}, a_{op}, a_{us} \in [0..1]$. Fig. 1 shows the domain triangle and the shape of f_{area}.

The area-based evaluation has the advantage that small occluded areas will tend to be examined more closely, since the overlapping area attempts to be about 40%. This forces the vehicle to approach unobserved areas until they occupy about

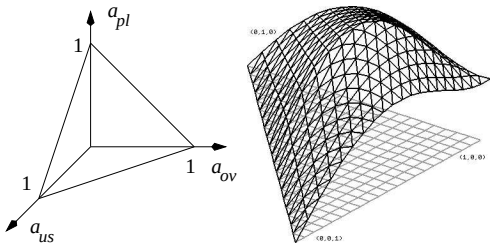

Figure 1: Area-based fitness function, f_{area} on (a_{ov}, a_{op}, a_{us})

60% of the image. On the other hand, and for the same reason, big occluded areas, like those hidden by a salient corner, will tend to be imaged from a further distance, and if new detail appears (in the form of occlusion planes), it will be examined closer in further views.

We have qualified as *basic* the above area-based criteria. Other criteria can be represented by factors $f_i \in [0..1]$ which multiply by f_{area}, thus increasing the total evaluation, $f = f_{area} \prod_i (1 + f_i)$. These secondary criteria may include:

- Quality improvement, $f_{quality}$.
- Structure of the overlapping area, $f_{structure}$. The purpose of this factor is to favour the sensing of areas where the surface has non-degenerate shape, thus easing the registration.
- Navigation cost, $f_{navigation}$, modelling the cost of reaching a new position and orientation from the current position of the sensor.

From these proposed criteria we have only tested $f_{quality}$, leaving the other two to further work. $f_{structure}$ could be computed from the variance of the surface normals at the new sensed points. $f_{navigation}$ should rely on robot path planning, reasoning about the known obstacles in the scene, the distance to travel, and trajectory of the robot. We have implicitly introduced a simple navigation factor in the definition of the feasible space when optimising the fitness function, but a reliable navigation factor should be computed by a navigation module.

For $f_{quality}$ we use the ratio of *Occupied* voxels that would improve quality from this view to the total number of *Occupied* voxels updated, multiplied by the mean quality improvement. Clearly $f_{quality} \in [0..1]$. The total fitness function becomes then

$$f = f_{area}(1 + f_{quality}) \ . \tag{4}$$

5 Optimisation

The feasible space is related to the physical characteristics of the sensor. If the sensor is mounted on a mobile base that moves on the floor, the mobile base cannot even move safely unless areas of the floor have been scanned and an obstacle-free path is found. Nevertheless, a navigation reasoning module is not taken into account in the present work, so we define the feasible space by these simple

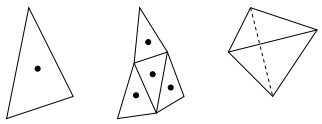

Figure 2: *Left*: face of an icosahedron. *Middle*: refinement. *Right*: simplex (tetrahedron) used in search algorithm

constraints: 1) The sensor must lie inside an *Empty* voxel. 2) The nearest *Occupied* voxel is not closer that $K \times scale$ (K times the voxel size) so the sensor does not collide with obstacles. $K = 4$ has been used.

The search strategy to optimise the fitness function can be one of the following: 1) Exhaustive search. This will find a global maximum but it would be extremely costly due to the five degrees of freedom of the problem. 2) Hill climbing methods. They will end up in a local maximum close to the starting position of the search. 3) Statistical methods: simulated annealing. These would need lots of fitness function computations, and are not guaranteed to end up in a global maximum. 4) Evolutionary methods: genetic algorithms. It would require an extremely large number of function computations to maintain a population of test positions.

To choose the search method one has to consider that our goal is to provide an answer, the next best position, in a reasonable amount of time. Several dozens of views will be necessary to recover a normal-sized room, so response times of the order of a minute, at most, are desirable. The fitness function is based on ray tracing on a subsampled range image used for view prediction, while the full range image is used for modelling. Sensor sizes may be of about 50-250 thousand points, and after subsampling the number of points may be still of about two thousand (64×32 for example).

We use a mixed method: exhaustive search in the 2D space formed by the pan-tilt angles, and a hill climbing method in the 3D space of sensor positions.

To perform an exhaustive (coarse) search in the pan-tilt space, we choose the centres of the 20 faces of an icosahedron as the values to test. These orientations are evenly distributed around a sphere, and in case there is spare time, a face can be subdivided as shown in Fig. 2 providing four new faces, which can be further subdivided to the desired resolution. So, provided the next best position and orientation of the sensor is worked out at this resolution, the orientation can be refined as desired.

For the hill climbing optimisation we use the N-dimensional *simplex* method [8]. The method starts from the current position of the sensor, and finds a nearby local maximum. A simplex in 3D space is a tetrahedron (Fig. 2). The vertices of the initial simplex are set as follows: the first vertex is set as the current sensor position, the other three are set randomly choosing an icosahedron face [1..20] and a ray within this face. These directions are projected a random distance (within a range).

A simplex evolves in 3D space changing its shape, size and position, aiming at high values of the fitness function. This is done by performing reflections of the worst point through the opposite face, expansions of a point along the direction of the opposite face, 1D contractions of the worst point toward the opposite face, and

3D contractions of all but the best point toward this point. A deeper explanation of the simplex optimisation algorithm can be found in [8].

The simplex optimisation is stopped when the range of change of the fitness function among the four vertices of the simplex is below a threshold (0.001 for example). This will always be reached eventually since the simplex gets smaller as it contracts toward vertices where the fitness function is better, and at a certain iteration the whole simplex will be contained inside a unique voxel, so the fitness of its four vertices will be the same, and the range will be zero.

The combination of the exhaustive search in the pan-tilt space and the simplex method in position space is done by computing the fitness of all 20 directions (faces of the icosahedron) at every position tested (a simplex vertex), and keeping the best evaluation as the fitness for that position.

The termination criterion is aimed at ensuring that the whole scene is recovered, and it can be: 1) No more unseen area is covered. 2) No more unseen area is covered and the quality of every pixel is above a threshold. 3) No more unseen area is covered and no more quality improvements are achieved. We used criterion 2.

The covered area and the occlusion plane area can be roughly computed as the number of voxel faces that touch an empty voxel, times the area of a voxel face.

6 Results

Although the proposed approach does not guarantee the selected best view is globally the best, which could be computed by exhaustive search in five dimensions, it provides a feasible solution which: 1) is locally a maximum of the fitness function, 2) is near to the previous sensor position, 3) improves quality of the covered area, and covers occlusion planes and new unseen areas.

To show the goodness of the method, experiments have been carried out using a simulated range sensor and mobile base. The base is able to move forward/backwards, left/right, and lift the sensor up/down. It can rotate, thus panning the sensor, besides the sensor can tilt from 0 to 180 degrees. The simulated sensor navigates in a scene model built with a CAD tool, accepting commands through a UNIX socket to perform the motion and to take range images. The range sensor observes 64×30 points, with horizontal angular aperture of 60 degrees, providing a $2\frac{1}{2}$D range image. The scene used consisted of a closed room of $5 \times 3 \times 3$ metres, with three boxes inside, Fig. 3. The voxel size was $scale = 10cm$ and the voxel map was of $64 \times 48 \times 32$ voxels.

Two experiments were carried out, one with the quality factor switched off $f_{quality} = 0$, and another taking into account this factor. The experiments were run till the 5D space to optimise the fitness function was almost flat (50 views with the quality factor off and 200 views with the factor on). The main difference was that in the second experiment the sensor reexamined the walls that had been scanned by almost-vertical views just at the beginning.

Fig. 4 shows several plots giving information of how the method was working. These include: *Fitness function*, which shows that the optimisation method finds good poses for the sensor, but declines as the scene is recovered. *Number of iterations* of the optimisation method: the number of function evaluations is this

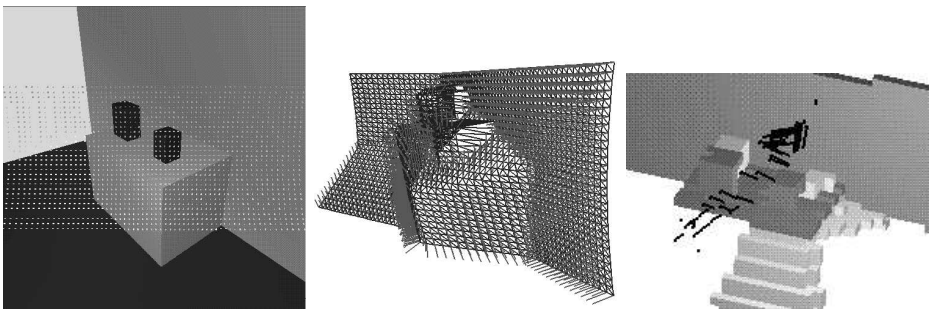

Figure 3: *Left*: original scene and range points. *Centre*: range image and normals. *Right*: positions tested by the simplex during one optimisation cycle. Lines indicate the best direction with length proportional to their fitness

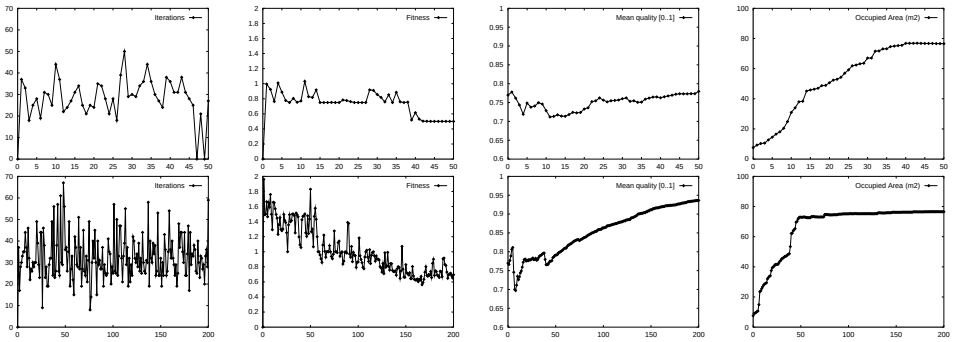

Figure 4: Plots showing the scene recovery process versus view number. *Top line*: with the quality factor off. *Bottom line*: with the quality factor on. *From left to right*: number of iterations to locate the next best view. Optimised fitness function at the next best view. Mean quality of *Occupied* voxels. Area of *Occupied* voxels

number multiplied by twenty (number of faces of an icosahedron). *Mean quality* of *Occupied* voxels: in the second experiment quality improves monotonically after the scene is initially completely observed. *Occupied area* in m^2: this figure stabilises in both cases after about 50 views, when the whole scene has been observed.

Fig. 5 shows the recovered scene after a number of views. The *Occupied* and *Occlusion Plane* voxels can be identified as dark and clear cubes. Fig. 3 shows all the positions tested by the optimisation method in a sample view as the simplex evolved in the 3D space. The lines start at the position of the vertices (dots) and have the direction of the best evaluation, with length proportional to the fitness function.

The test results show that the scene scanning is virtually complete after a reasonable number of views (50) with or without the quality measure. Additional scans are needed to obtain the remaining isolated unobserved voxel faces, or to improve the quality. A drawback of the approach is that when almost the whole scene is recovered and there just remain few isolated viewpoints, the space is al-

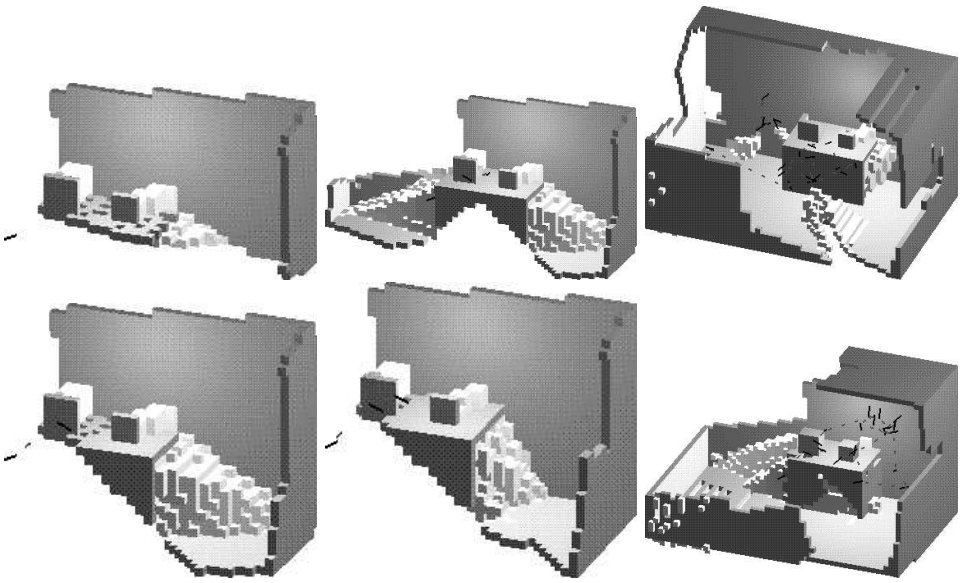

Figure 5: *Left*: recovered scene after first and second views. *Centre*: recovered scene after third view, quality factor off (up) and on (down). *Right*: recovered scene after view 20, quality factor off (up) and on (down). *Occupied* voxels are in dark and *Occlusion Plane* voxels are in clear. The path followed by the sensor is pointed out by dashed lines, sensor positions and orientations by solid lines

most flat regarding the fitness function, and the local hill-climbing method cannot find a good direction to "climb". In this case the simplex evolves randomly until a timeout is signalled. This "tiding-up" phase should be then directed by a deterministic approach (detecting holes in the scene by morphology analysis and computing viewpoints for them). But a voxel hole in an area of *Occupied* voxels does not mean that the 3D scene model represented by a triangular mesh could not be realistically built.

The scene used has an exact area of $80.64m^2$. The minimum number of views to cover this area, assuming a 40% overlap, and that the sensor stays at $2m$ from the surfaces, covering an area of $2(2\tan\frac{\alpha}{2})(2\tan\frac{\beta}{2}) = 2.5m^2$ ($\alpha = 60deg, \tan\beta = \frac{M}{N}\tan\alpha, M = 30, N = 64$), would be of 81 views. On the other hand, the maximum number of views is given by visiting all the voxels inside the room, that is $\frac{5\times3\times3}{scale\times scale\times scale} = 4500$. As we can see the number of views taken by the present approach is quite reasonable.

The time to deduce the next observation position depends on the current scene complexity, but the experiments reported here took approximately 832 seconds for the first experiment (50 views) and 1252 seconds for the second one (200 views) on a *Pentium* processor at 166 MHz. This is approximately 16 seconds per view on average. The average distance travelled by the sensor from one view to the next was less than one metre. The storage requirements for the voxel representation were about 3/4 of a M-byte, which is also low enough for practical use.

7 Conclusions

We have presented an approach to full 3D scene recovery by a range sensor mounted on a mobile robot. The recovered scene model is represented by a voxel map at just enough resolution for computing the criteria to find the next best view. This allows a straightforward registration of new views with the scene model (voxel map update). During scene model recovery, one could also acquire higher resolution surface and texture data. This data would not be needed for the best next view planning process, but could be used for scene modelling.

The next best view is recovered by mixed exhaustive search and hill climbing optimisation, and the fitness function is based on area proportions of the new image to be sensed. This means that detailed areas are examined closer. We have envisaged other criteria that can modify the basic one, aimed at improving the quality of the sensed data, or at favouring the recovery of high structured parts to ease the registration, if a realistic recovery of the scene is to be performed by another task.

We have presented results that show the feasibility of the method. The experiments have been carried out on a simulated sensor and mobile robot navigating in a CAD scene, with and without the aim of quality improvement of the data. Recovering our sample scene with a mean quality of 0.9, for example, takes 4 times the number of views required for recovering it with any data quality. We expect to perform real-scene experiments in due time.

References

[1] N.A. Massios and R.B. Fisher. A best next view selection algorithm incorporating a quality criterion. In *Proc 6th British Machine Vision Conference*, pages 780–789, 1998.

[2] M.A. García, S. Velázquez, and A.D. Sappa. A two-stage algorithm for planning the next view from range images. In *Proc 6th British Machine Vision Conference*, pages 720–729, 1998.

[3] M.K. Reed, P.K. Allen, and I. Stamos. Automated model acquisition from range images with view planning. In *Proc of Int Conf on Computer Vision and Pattern Recognition, CVPR'97*, pages 72–77, 1997.

[4] K.A. Tarabanis, R.Y. Tsai, and A. Kaul. Computing occlusion-free viewpoints. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(3):279–292, 1996.

[5] R. Pito. A sensor-based solution to the next best view problem. In *13th International Conference on Pattern Recognition, ICPR'96*, volume I, pages 941–945, 1996.

[6] E. Wolfart, V. Sequeira, K. Ng, S. Butterfield, and J.G.M. Gonçalves. *Hybrid approach to the construction of triangulated 3D models of building interiors*, volume LNCS 1542, pages 489–508. Springer-Verlag, 1999.

[7] J.E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.

[8] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P.Flannery. *Numerical Recipes in C, the art of scientific computing.* Cambridge University Press, 1995.