Using NEWUOA to drive the autonomous visual modeling of an object by a Humanoid Robot

Torea Foissotte [1,2], Olivier Stasse [2], Pierre-Brice Wieber [3], Abderrahmane Kheddar [1,2]

[1]*CNRS-LIRMM, France*
[2]*CNRS/AIST JRL, Japan*
[3]*INRIA, France*

***Abstract*— A significant amelioration of our previous work is presented which aims at building autonomously visual models of unknown objects, using a humanoid robot. The use of embedded sensors inside the head of the robot in order to construct this model necessitates an original solution compared to previous approaches in the litterature.**

Previously we introduced a Next-Best-View solution using two steps: (i) an optimization algorithm without derivatives, NEWUOA, is used to find a camera pose which maximize the amount of unknown data visible, and (ii) a whole robot posture is generated by using a different optimization method where the computed camera pose is set as a constraint on the robot head. This paper presents modifications of the original algorithm in order to improve the robustness while broadening the cases that can be handeld.

I. Introduction

A. Context of the work

The main question addressed in this paper is how to generate successive postures of a humanoid robot in order to build a 3D model of an unknown object while taking into account different constraints on the robot body as well as the visual characteristics of the object perceived by the robot. The perception of the object is done using the stereo cameras embedded in the humanoid head by constructing a disparity map and also by detecting landmarks represented by SIFT features [1]. The disparity map is used to perform space carving on an occupancy grid which corresponds to the 3D model to construct. We take as an hypothesis that the environment is known so that its distinction from the object to model is simplified.

The work presented in this paper details our current solution to generate humanoid whole-body postures given an occupancy grid, a list of SIFT landmarks and the constraints on the humanoid body: self-collisions, collisions with obstacles in the environment, stability and joint limitations. To achieve this, we rely on two complementary optimization methods: a derivative-free optimization method and a gradient-based method. This work continues the one presented in [2] by improving the robustness of the criterion used in the derivative-free optimization method and by widening the range where the algorithm can be applied.

B. Overview of related work

The planning of sensor positions in order to create a 3D model of an unknown object is known as the Next-Best-View (NBV) problem and has been addressed for several years, the first notorious work in the field being [3]. Amongst following works, we can cite [4], [5], [6] or [7]. Hypotheses and limits of such works are detailed with more details in these two surveys: [8] and [9]. Most works in the NBV field make the assumptions that the depth range image is dense and accurate by using laser scanners or structured lighting, and that the camera position and orientation is correctly set and measured relatively to the object position and orientation. The object to analyze is also considered to be inside a sphere or on a turntable, i.e the sensor positioning space complexity to evaluate is reduced since its distance from the object center is fixed and its orientation is set toward the object center. The main aim is to get an accurate 3D reconstruction of an object, using voxels or polygons, while reducing the number of viewpoints required. Such methods can be efficiently used in controlled environments with relatively small objects, and using a sensor with accurate positionning. Though the problem of sensor positionning has

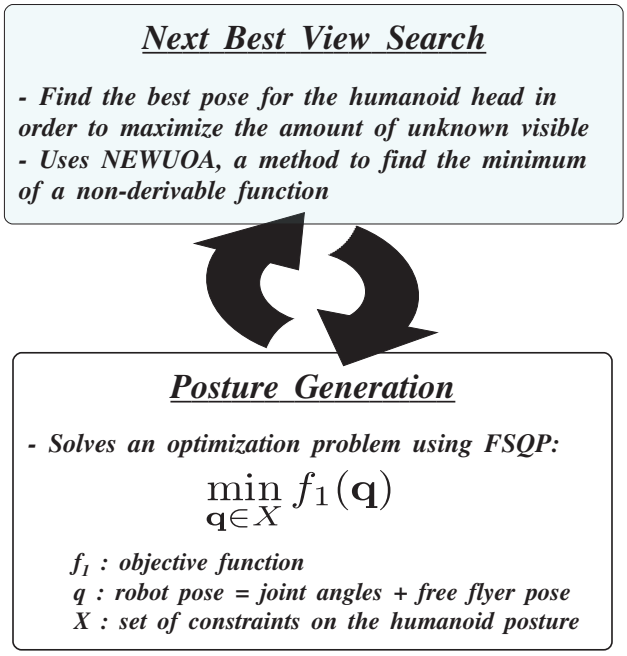

Fig. 1. Two-steps approach for the generation of posture.

been adressed in some works such as [10], they aim at the exploration of configuration space with a simplified robot. The aim is different in our work as the sensor is embedded in a humanoid which can move in a known, but not necessarily obstacle-free, environment. Furthermore our goal is to have a method useful for the modeling of objects with no specific particularities: objects can be small or big, and have complex shapes.

C. Contribution

Though our modeling process also requires a NBV solution, it appears that working hypotheses are quite specific for a humanoid robot and thus our work differs in few important issues:

1) the sensor pose is constrained due to it being embedded in a humanoid robot. Moreover the presence of obstacles in the environment must be dealt with.
2) the sensor's result positions need not being further constrained to some precomputed discrete positions on a sphere surface, and its viewing direction is not forced toward a sphere center. Thus the algorithm can be easily used to model objects of different sizes,
3) to correct possible positionning errors, a constraint that keeps some landmarks visible to the camera is implemented,
4) an accurate 3D model of the object is not required. Our goal is to get a set of visual features around the object to allow its effective detection and recognition.

In [11], the object modeling was performed by generating postures with the robot head pose set as a constraint given by a human supervisor. In [12], a first attempt to complete this work by using visual cues to guide the modeling process automatically was proposed by using a formulation which can be directly integrated into our posture generator. However this formulation results in convergence problems when generating a pose, thus the two-steps approach, outlined in Fig. 1, was designed and presented in [2]. Section II summarizes this approach and presents the latest improvements added. Analysis of these modifications are presented in section III and finally section IV concludes this paper.

II. Two steps NBV approach

The solution presented in [2] decomposes the problem in two: first, find a camera position and orientation that maximizes the amount of new visual information while solving specific constraints related to the robot head, then generate a whole-body posture for the robot using the desired camera pose as a constraint on the robot head pose. If a whole-body posture cannot be generated, the first step is run again with some modifications in its input data.

For the first step, we use NEWUOA [13], a method that can find the minimum of a function by refining a quadratic approximation of it through a deterministic iterative sampling, and which can thus be used for non-differentiable functions. NEWUOA has the advantages of being fast and robust to noise while allowing us to keep the 6 degrees of freedom of the robot camera.

The second step uses the posture generator (PG) proposed as part of the work in [14] and [11]. This PG relies on FSQP to give a posture that minimizes an objective function while solving given constraints expressed as derivable functions.

A. Evaluation of visual data

To define the best view, we need an estimation of new information that can be discovered depending on the view posture. Following the works of [5] and [3], our approach uses an occupancy grid and the space carving algorithm for this purpose. The object model can be composed of perceived (known) voxels and occluded (unknown) voxels, and is updated each time a disparity map is constructed by stereo vision. The NBV algorithm is based on the evaluation of the unknown surface visible from a specific robot pose. This evaluation can be done in a relatively fast way by rendering known and unknown voxels as 3D cubes, using OpenGL.

In [2], by coloring known voxels in blue and unknown voxels in green, the amount of unknown visible is defined as the number of green pixels on the screen. This gives a useful estimation when the voxels have a relatively small size compared to their distance to the robot and the camera field-of-vision angle. As a constraint which sets a minimum distance between the robot head and the object is needed in order to create the disparity map using stereo vision, we can define a maximum threshold for the size of the object to be detected depending on the characteristics of the cameras embedded in the robot head. But a problem arises when the size of voxels gets big relatively to the minimum distance allowed: the maximum surface of a voxel that can be projected increases, resulting in computed views where the amount of unknown voxels visible gets significantly reduced. In such cases, the number of poses necessary to build the model increases as the robot gets close to few voxels instead of trying to perceive the maximum number of unknown voxels possible.

A simple way to deal with this problem is to increase the resolution of the occupancy grid depending on the object size but this increases drastically both the memory space required and the computation time. Instead we are now using a fast estimation of the number of unknown voxels visible, relying on OpenGL, by associating a unique color to each unknown voxel. The number of unknown voxels visible in one frame is thus equal to the number of different pixel colors.

B. Constraints on the camera pose

Though NEWUOA is supposed to be used for unconstrained optimization, some constraints on the camera pose need to be solved in order to be able to generate a posture with the PG from the resulting desired camera pose. The constraints on the camera position $\mathbf{C}$ and orientation $\mathbf{\Theta_c}$ included in the evaluation function of the first step given to

NEWUOA are:

$$\begin{cases} C_{zmin} < \mathbf{C}_z < C_{zmax} & (1) \\ \forall i, d_{min} < d(\mathbf{C}, \mathbf{Vox_i}) & (2) \\ \Theta c_{xmin} < \mathbf{\Theta_c}_x < \Theta c_{xmax} & (3) \\ \Theta c_{ymin} < \mathbf{\Theta_c}_y < \Theta c_{ymax} & (4) \\ N_l > N_{lmin} & (5) \\ \forall i, \mathbf{C} \neq F_i \vee \mathbf{\Theta_c} \neq Fr_i & (6) \end{cases}$$

The range for the camera height is limited by (1) to what is accessible by the humanoid size and joints limits. A minimum distance d_{min} is imposed by (2) between the robot camera and all the non-empty voxels of the object. This corresponds to a requirement in order to generate the disparity map with the two cameras embedded in the robot head. No maximum distance constraint is used. The rotations on X and Y axes are restricted by (3) and (4) to ranges manually set according to the robot particularities. The constraint (5) keeps a minimum number of landmarks, i.e. features that were detected in previous views, visible from the resulting viewpoint. By matching them with features detected within the new viewpoint, it is possible to correct the odometry errors due to the movement of the humanoid and thus the position and orientation of the features detected all around the object, relatively to each other, can also be corrected. Finally, we added a new constraint (6) which ensures that the resulting pose will not be near previously found poses, with position F_i and orientation Fr_i, which could not be reached by the PG in the second step. Though we didn't run into such cases when doing our experiments in [2], this is necessary to ensure the algorithm can find a pose even when constraints not expressed in the first step can block the convergence in the second step, e.g. the presence of obstacles in the environment.

C. New constraints formulation in the evaluation function

Two modifications have been made concerning the formulation of constraints in the evaluation function given to NEWUOA: a reformulation of the constraint on landmark visibility (5) and the addition of a new constraint to avoid camera poses which resulted in problems of convergence for the PG.

In [2], the landmark constraint relies on the visibility of corresponding voxels. Our new formulation relies both on the visible surface of landmarks and their normal vector as the SIFT landmarks cannot always be detected whenever they are visible. The surface visibility for each landmark *i* is computed relatively to its amount of pixels visible from the current viewpoint pv_i using a sigmoid function:

$$ls_i = \frac{1}{1 + exp\left(pmin_i - pv_i\right)} \tag{7}$$

The parameter $pmin_i$ is the minimum amount of pixels required to consider the landmark *i* visible, and its value depends on the original landmark size.

The visibility of each landmark relatively to their normal vector $\mathbf{Nl}_i$ and the current camera view direction vector $\mathbf{C}_{view}$ is expressed using another sigmoid:

$$ln_i = \frac{1}{1 + exp\left(\beta\left((\mathbf{C}_{view}.\mathbf{Nl}_i) + \phi\right)\right)} \tag{8}$$

where ϕ is related to the angle range allowed, and β determine the slope of the sigmoid function. The final visibility coefficient for each landmark is computed by multiplying ls_i with ln_i. We set an arbitrary defined minimum number of visible landmark Nlm_{min} which is compared to the obtained coefficients for all landmarks N using:

$$lv = \left(\sum_{i=0}^{N} ls_i.ln_i\right) - Nlm_{min} \tag{9}$$

The constraint for the evaluation function is defined in one of two ways depending on the sign of lv. Configurations maximizing lv are slightly encouraged when it is positive:

$$K_l = -\eta\, lv \tag{10}$$

The η parameter can be small so that the minimization of other constraints and the maximization of unknown visible both have a greater priority than the increase of number of visible landmarks beyond the defined threshold. In the other case, where $lv \leq 0$, the configurations are greatly penalized:

$$K_l = \left(\frac{lv}{Nlm_{min}}\right)^2 \tag{11}$$

The new constraint to avoid unreacheable postures is formulated as:

$$K_f = \sum_p exp\left(-\delta.D_{fp}\right) \tag{12}$$

where D_i represents the sum of absolute differences between the values of the actual camera pose and the unreachable pose fp. δ corresponds to the sensitivity of the constraint.

The evaluation function to minimize which is used as input to the NEWUOA algorithm becomes:

$$\begin{aligned} f_e = &\lambda_z K_{\mathbf{C}_z} + \lambda_x K_{\mathbf{\Theta_c}_x} + \lambda_y K_{\mathbf{\Theta_c}_y} \\ &+ \lambda_d K_d + \lambda_l K_l + \lambda_f K_f - \lambda_n N_{up} \end{aligned} \tag{13}$$

The λ parameters are fixed manually to modify the balance between the constraints. N_{up} is the number of unknown voxels visible from the camera pose. $K_{\mathbf{C}_z}$, $K_{\mathbf{\Theta_c}_x}$, $K_{\mathbf{\Theta_c}_y}$ and K_d parameters corresponds respectively to the height limit, X and Y rotation limit, and the minimum distance constraints which are formulated in [2].

D. Evaluation function behavior

Due to the constraints used and the specificities of objects to model, many different cases can result in local minima in our evaluation function that are quite disjoint as can be seen in the example shown in Fig. 2. This figure illustrates the behavior of some constraints in the evaluation function for different camera position around the object carved once. The best orientation found by a NEWUOA search is chosen for each position. The object simulated is 3 meters high, the camera is placed at a height of 1.3 meters and its X and

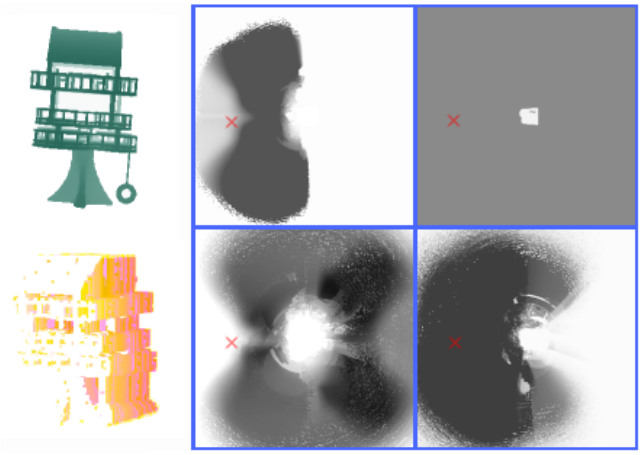

Fig. 2. Example of the evaluation function components obtained with a fixed camera height and depending on the camera 2D position on the plane XY around an object (left-top) which has been perceived once. The red cross indicates the position of the camera used for the carving. Left-bottom: view of the occupancy grid which is used to predict the amount of unknown visible, center-top: f_e, center-bottom: N_{up}, right-top: K_d and right-bottom: K_l.

Y positions are in the interval [-15,15] meters. There is a distance of 0.1 meter between each position tested.

The number of unknown voxels visible (center-bottom) as well as the constraint on the landmarks (right-bottom) are composed of many abrupt variations due to the occlusions between known and unknown voxels occurring with relatively complex shapes. We can note in the final evaluation (center-top) that the landmark constraint decourages the algorithm to select a camera pose where the visibility of unknown is the best: near the opposite side of the perceived object face.

A particularity of our algorithm is highlighted by the constraint on the minimum distance (right-top) where the camera is allowed to be placed inside the object. As the object is composed of many holes, when space carving is applied, there appears some parts inside the occupancy grid which are empty. This means that the algorithm has the possibility to set a robot pose inside the object if it is big enough, for example a house, and thus, though it is not its main goal, the algorithm can still be used for exploration tasks too.

E. NEWUOA configuration

NEWUOA seeks the minimum of f_e by approximating it with a quadratic model, inside a trust region. Thus an initial configuration is provided to the software which limits the initial sampling to a subspace according to a range given by the user. Nevertheless NEWUOA's complete search is not limited to the trust region and can test vectors outside depending on the quadratic approximation obtained. In fact, result poses are most of the time clearly outside of the bounds of the region.

After extensive tests, two techniques were introduced in our previous paper to cope with local minima seen in Fig. 2 and improve the search of an optimum solution: first, we run NEWUOA in an iterative way, i.e run it once with a manually chosen starting pose then run it again by setting the starting pose with the previously found one, etc. The computation stops when two successive iterations give the same result or a maximum number of iteration is reached. Second, we set manually different starting poses around a reference position, launch the iterative process from each chosen position and select the best one.

A new significant improvement introduced here is the decoupling of the camera position and orientation searches. In [2], both the position and orientation are given as input to NEWUOA but, in fact, the quadratic approximations of the evaluation function depending on the orientation parameters themselves depend on the current position. Thus NEWUOA is now used in cascade: one occurrence searches for the best position and each time a position is tested, a new occurence is launched which searches for the best orientation corresponding to this position.

F. Second step: Posture Generator

Once an optimal camera pose has been found, the result is used as a constraint on the humanoid robot head in order to generate a whole-body posture that takes into account all other constraints such as stability, collisions, etc.

The starting robot pose is set using a pre-computed posture at the position and orientation of the desired camera pose. In cases where the PG cannot converge, the goal camera pose is put inside the list of forbidden poses which is used in the constraint 6 and NEWUOA is launched again to find another pose.

III. Simulations

A. NEWUOA tests for camera pose evaluation

First we compare the improvement of using NEWUOA in cascade in order to look for the position and orientation separately. Examples of results are illustrated in Fig. 3 where the starting pose of the camera is translated on the Y axis near a carved object with a height of 0.5 meters (top), and 3 meters (bottom). The results of our previous approach (left) are compared with those of the new one (right). The main difference is the ability of the cascade method to reach a better pose in one iteration. Thus the number of iterations needed in order to reach an optimum is significantly reduced. However the computation time is still much higher than for the previous method as several NEWUOA searches are performed per iteration.

Though there appears to have greater variability in the results for our latest method when considering big objects, in overall, the resulting poses still have better evaluations than those obtained with our first approach.

B. Constraint on forbidden poses

The Fig. 4 illustrates the influence of our new constraint to avoid unreachable poses on our evaluation function. The graphics are obtained in the same conditions than for Fig. 2 though the object evaluated differs. The pose to avoid

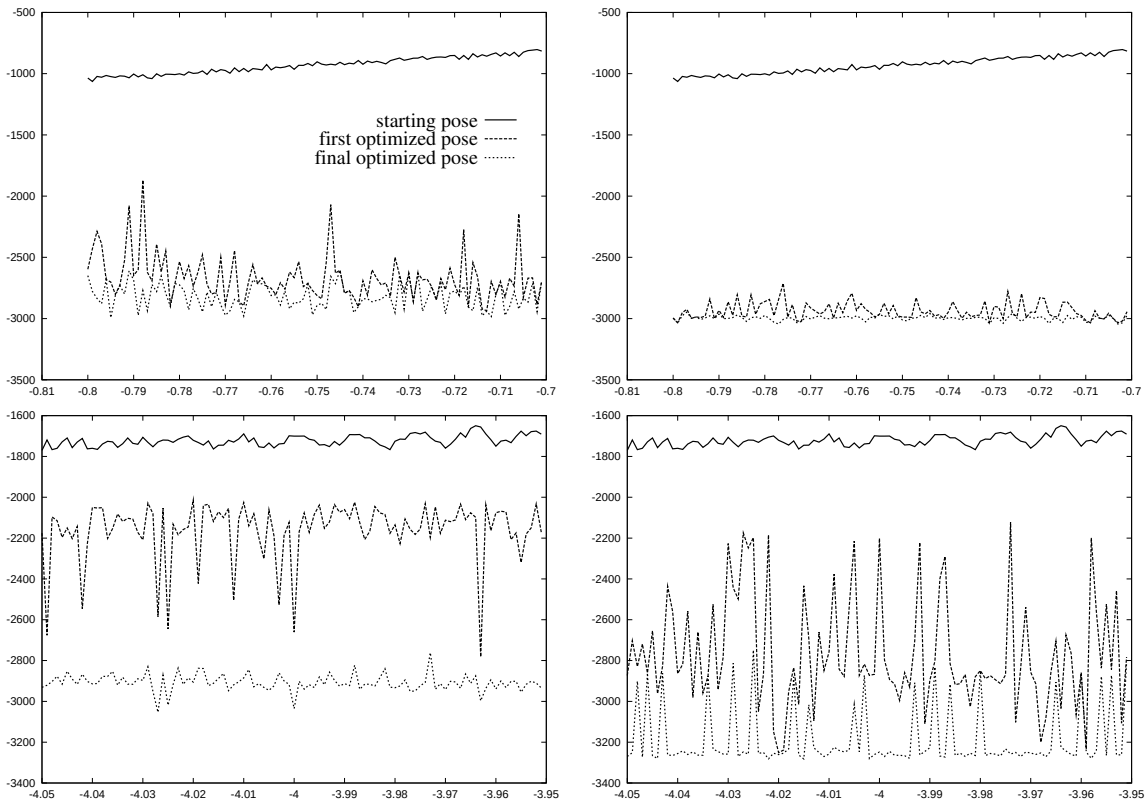

Fig. 3. Evaluation (Y axis) of the poses obtained before, at the beginning and at the end of the iterative NEWUOA search depending on the starting position (X axis). Left column: evaluation results when NEWUOA is searching for the best position and orientation simultaneously. Right column: results with NEWUOA used in cascade. Top line: results obtained with a small object. Bottom line: results obtained with a big object.

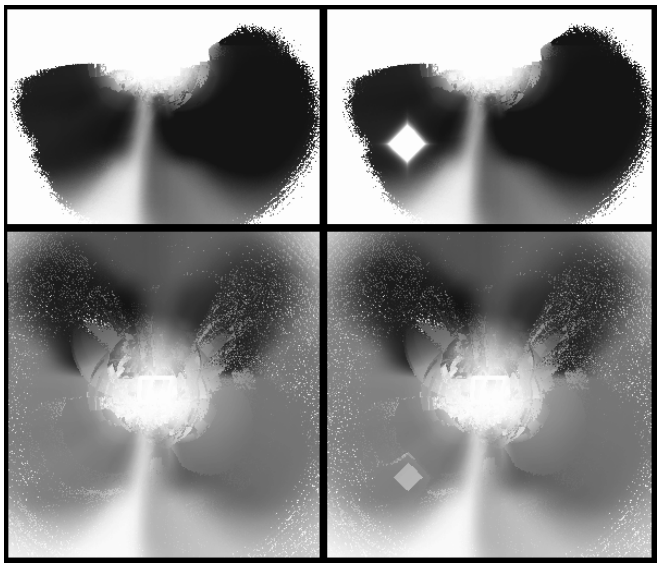

Fig. 4. Influence of the pose avoidance constraint on the evaluation function. Top line: evaluation results depending on fixed X and Y positions. Bottom line: Estimation of unknown data visible. Left column: initial evaluation. Right column: evaluation done after the manual input of a pose to avoid. The XY position of the forbidden pose is located in the middle of the white square on the dark area in the top-left image, and in the middle of the grey square in the bottom-left image.

was chosen as one of the pose obtained initially at a predetermined XY position. The parameter δ from equation 12 is set to a large value so that the constraint appears clearly in the evaluation. Using a δ value small enough results in an evaluation similar to the initial one as a change of orientation relatively to the reference pose is enough to set the constraint inactive.

C. Modeling process simulation

The experimental setting is simulated by having a virtual 3D object perceived by a virtual camera. The modeling process loops through the following steps:

1) The disparity map is constructed using the object 3D informations and is used to perform a space carving operation on the occupancy grid. Some known voxels are randomly selected to be considered as landmarks.
2) The NEWUOA routine is then called in order to find an optimal camera pose by minimizing our evaluation function. In our previous work, the starting reference point was set by rotating the actual robot position around the object then a simple fixed sampling of the 3d space was computed from this point to be used as starting poses for the iterative search. Our new implementation does not require this initial rotation due to the improvements of the search and thus the sampling is done using the current robot position.

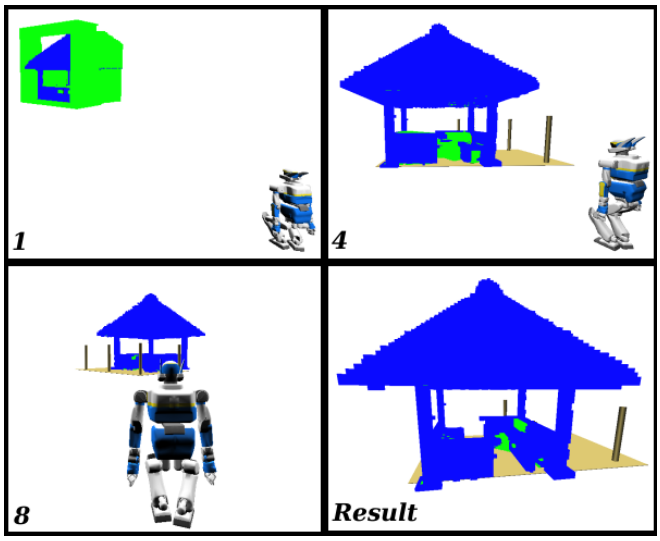

Fig. 5. Selection of 3 postures among the 8 generated during the modeling of a house and the resulting model. Updated occupancy grids are displayed with known voxels in blue and unknown voxels in green.

3) When an optimal camera pose is found, it is sent to the PG in order to generate a whole-body posture. If the PG does not converge, we add this camera pose in the list of poses to avoid, using the new constraint (6), and run the first step again.

An example of postures generated during the successful modeling process of a six meters high house is illustrated in Fig. 5. The modeling was completed after 8 poses. The typical posture of the robot is due to the starting posture used by the PG. As the object is quite big and the robot cameras have a field of view of 25 degrees, the postures generated are relatively far away from the object. At the end of the process, we can note that some voxels inside the house could not be perceived: the last poses generated were too far away from a pertinent pose to fully complete the model thus the robot is stuck inside a local minima. A possible solution is to increase the number and ranges of sampled starting poses for the NEWUOA routine at the cost of increasing the computation time.

During our tests, we could observe a significant reduction, which can go up to 40 percent, of the number of poses necessary to model big objects by using our new estimation of unknown data.

IV. Conclusion

Latest improvements to our Next-Best-View algorithm for a humanoid robot have been presented in this paper. The robustness of our approach is enhanced by three main modifications: (i) a cascade NEWUOA search of the best camera pose which splits the search of the best positions and orientations, (ii) a better formulation of the landmark visibility constraint which also takes into account the landmarks' normal vector, and (iii) the addition of a constraint to drive the NBV away from poses which could not be reached by the PG.

Our tests on the modeling of big objects lead us to add another improvement by estimating the amount of unknown using the number of unknown voxels instead of the number of pixels corresponding to unknown data.

Following this work, we are now working on complementary tasks, such as motion planning and other vision topics, in order to confirm the results obtained in simulation and further analyze the pertinence of our formulation by doing real experiments with an HRP2 robot.

Acknowledgment

This work is partially supported by grants from the ROBOT@CWE EU CEC project, Contract No. 34002 under the 6th Research program `www.robot-at-cwe.eu`.

The visualization of the experimental setup relied on the AMELIF framework presented in [15].

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 4, pp. 91–110, 2004.

[2] T. Foissotte, O. Stasse, A. Escande, P.-B. Wieber, and A. Kheddar, "A two-steps next-best-view algorithm for autonomous 3d object modeling by a humanoid robot," in *IEEE ICRA Proceedings*, 2009.

[3] C. Connolly, "The determination of next best views," in *IEEE International Conference on Robotics and Automation*, 1985.

[4] J. Maver and R. Bajcsy, "Occlusions as a guide for planning the next view," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 1993.

[5] J. Banta, Y. Zhien, X. Wang, G. Zhang, M. Smith, and M. Abidi, "A best-nextview algorithm for three-dimensional scene reconstruction using range images," in *Proceedings SPIE*, 1995.

[6] R. Pito, "A solution to the next best view problem for automated surface acquisition," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 1999.

[7] K. Yamazaki, M. Tomono, T. Tsubouchi, and S. Yuta, "3-d object modeling by a camera equipped on a mobile robot," in *IEEE ICRA Proceedings*, 2004.

[8] K. Tarabanis, P. Allen, and R. Tsai, "A survey of sensor planning in computer vision," in *IEEE Transactions on Robotics and Automation*, 1995.

[9] W. Scott, G. Roth, and J. Rivest, "View planning for automated three-dimensional object reconstruction and inspection," *ACM Comput. Surv.*, 2003.

[10] Y. Yong and K. Gupta, "An information theoretical approach to view planning with kinematic and geometric constraints," in *IEEE ICRA Proceedings*, 2001.

[11] O. Stasse, D. Larlus, B. Lagarde, A. Escande, F. Saidi, A. Kheddar, K. Yokoi, and F. Jurie, "Towards autonomous object reconstruction for visual search by the humanoid robot hrp-2." in *IEEE RAS/RSJ Conference on Humanoids Robots, Pittsburg, USA, 30 Nov. - 2 Dec.*, 2007.

[12] T. Foissotte, O. Stasse, A. Escande, and A. Kheddar, "A next-best-view algorithm for autonomous 3d object modeling by a humanoid robot," in *IEEE RAS/RSJ Conference on Humanoids Robots, Daejeon, South Korea, 1-3 Dec.*, 2008.

[13] M. Powell, "The newuoa software for unconstrained optimization without derivatives," University of Cambridge, Tech. Rep. DAMTP Report 2004/NA05, 2004.

[14] A. Escande, A. Kheddar, and S. Miossec, "Planning support contact-points for humanoid robots and experiments on hrp-2," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 2974 – 2979.

[15] P. Evrard, F. Keith, J.-R. Chardonnet, and A. Kheddar, "Framework for haptic interaction with virtual avatars," in *17th IEEE International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN 2008)*, 2008.