Online Object Search with a Humanoid Robot

Francois Saidi*, Olivier Stasse*, Kazuhito Yokoi* and Fumio Kanehiro†
* ISRI/AIST-STIC/CNRS Joint Japanese-French Robotics Laboratory
Central 2, 1-1-1 Umezono, Tsukuba, 305-8568 JAPAN
Email: francois.saidi,olivier.stasse,kazuhito.yokoi@aist.go.jp
† Intelligent Systems Institute
National Institute of Advanced Industrial Science and Technology
Central 2, 1-1-1 Umezono, Tsukuba, 305-8568 JAPAN
Email: f-kanehiro@aist.go.jp

***Abstract*— This paper presents an object active visual search behavior in a 3D environment performed by a HRP-2 humanoid robot. The search is formalized as an optimization problem in which the goal is to maximize the target detection probability while minimizing the energy/distance and time to achieve the task. Natural constraints on the camera parameter space based on the characteristics of the recognition system are used to reduce the dimension of the problem and to speed up the optimization process to achieve a real time behavior. We present simulation and real experimental results using an HRP-2 robot.**

I. INTRODUCTION

A. The visual search behavior

Looking for our car keys in the whole house or just scanning the top of the table before locating and grasping the pencil are some common search behavior we, as human, perform easily. With our precise and robust vision system and its amazing recognition ability along with the mastering of our complex kinematics and the 3D motions it allows, active visual search is an easy task.

With a search ability a robot doesn't need to keep a record of the precise 3D coordinates of objects with which it has to interact. And even if such a record is maintained, what happens if these objects are moved? Humanoid robots are multipurpose platforms and will need to use generic tools to extend their capacities. They must thus be able to look for objects, to localize and use them. A search behavior would be a great improvement in humanoid autonomy and a step forward toward their rise outside laboratories.

Before starting a search behavior, the robot needs a model of the desired object. This model could be provided by an external mechanism, but a humanoid has all the required abilities to build that model by its own. An undergoing project in our laboratory, called the "Treasure hunting" aims at integrating in a unique cycle, the model building of an unknown object, and the search for that object in an unknown environment. With such a combined skill, the robot may incrementally build a knowledge of its surrounding environment and the object it has to manipulate without any a-priori models. Latter the robot would be able to find and recognize that object. The time constraint is crucial, as a reasonable limit has to be set on the time an end user can wait the robot to achieve its mission. This paper will focus on the search behavior and we assume that the object model is already created.

B. Problem statement and contributions

Object search is a sensor planning problem which is proven to be NP-complete [1] thus a heuristic strategy is needed to overcome that task. Because of the limited field of view, the limited depth, the lighting condition, the recognition algorithm limitation, and possible occlusion, many images from different points of view are necessary to detect and locate a given object.

The initial knowledge of the target position is encoded in a discrete presence probability [2] which will be updated after each detection attempt. By combining the target distribution knowledge and a model of the recognition system accuracy, we are able to calculate the likelihood of detecting the target for a given sensor parameter. The proposed planning strategy consists in optimizing a rating function at each sensing step. The rating function analyzes the expected field of view (according to already mapped environment) for a given configuration according to various criteria defined further in this paper. In [3], we introduce the concept of *Visibility Map* a statistical accumulator in the sensor configuration space which takes into account the characteristics of the recognition system to constrain the sensor configuration space and speed up the optimization. This paper presents the full search planning strategy along with experiments on the HRP-2 humanoid robot.

C. Related works

Few works on active 3D object search are available, fortunately the sensor planning research field provides us with some hints.

Wixon [4] uses the idea of indirect search (in which one first finds an object that commonly has a spatial relationship with the target, and then restricts the search in the spatial area defined by that relationship) he proposes a mathematical model of search efficiency, which shows that indirect search can improve the search.

Works done by Ye and Tsotsos [2] tackle the field of sensor planning for 3D object search. The search agent's

knowledge of object location is encoded as a discrete probability density which is updated after each sensing action. The detection function uses a simple recognition algorithm, and all factors which influence the detection ability such as imaging parameters, lighting condition, complexity of the background, occlusions etc. are included in the detection function value by averaging experimental results done under various conditions. The vision system uses one pan tilt zoom camera and a laser range finder to build a model of the environment. The search is not really 3D as, the object is recognized using a 2D technique, and the height of the camera is fixed.

Works by Sujan [5] are not focused on object search but on accurate mapping of unknown environment by the mean of sensor planning. The author proposes a model based on iterative planning, driven by an evaluation function based on Shannon's information theory. The camera parameter space is explored and each configuration is evaluated according to the evaluation function. No computational timing tests are provided, but the algorithm seems to focus on configurations which are close to obstacles or to unknown areas to improve the algorithm efficiency, this latter constraint will be formalized with the notion of visibility map introduced in II-B.

The operational research community [6] has extensively studied the problem of optimal search, they came up with interesting theoretical results on search effort allocation which served as a basis for Tsotsos's work.

The Next Best View (NBV) research field [7] studied the sensor planning problem mainly for C.A.D. model building. These works, although sharing some common aspects with the present topic, rely on the fundamental assumption that the object is always in the sensor field.

II. Constraints on the sensor

A. Model of the recognition system

All recognition algorithms have some restrictions regarding the imaging condition (lighting, occlusion, scale...). One of the main assumption which can be easily controlled by active vision is the scale limitation: the smallest scale at which the object can still be recognized constitute a maximum distance limit for the recognition algorithm (R_{max}). It is also suitable to have a sensor configuration in which the whole object is projected inside the image in order to maximize the number of imaged features, this imposes a lower limit for the sensor distance to the object (R_{min}).

Without any loss of generality regarding the recognition algorithm, we can assume that these bounding values (R_{min} and R_{max}) are determined theoretically or experimentally during the model building and are stored with the object model. These limit values depend on the recognition algorithm and on the characteristics of the searched object and are used to further constrain the sensor parameters in order to improve optimization time.

We also assume that a model of the recognition system, which gives the accuracy of the recognition depending on the position of the target relative to the sensor, is available. For instance, in this paper we use a gaussian formulation of the recognition accuracy (equation 1), in which z is the distance of a given voxel to the camera optical center.

$$\rho(c, v_i) = \exp \frac{1}{2}\Big(\frac{z-m}{\sigma}\Big)^2 \tag{1}$$

with

$$m = \frac{R_{max} + R_{min}}{2} \quad and \quad \sigma = \frac{R_{max} - R_{min}}{2}$$

B. The visibility map

The configuration space of the stereoscopic head has initially 6 DOF, but because the roll parameter (rotation around the line of sight) has a small influence on the visible area (the stereoscopic field of view is square), the problem is reduced to 5 dimensions.

The sensor configuration space is discretisized with the same resolution as the occupancy grid for the x, y and z parameters (5 cm). Whereas for pan and tilt, a resolution of half the stereoscopic field of view value, which is 33 ° horizontally and vertically, is used. With such a resolution and a typical environment size of 6x12x2 meters, the configuration space of the sensor has around 24 millions configurations. A greedy optimization approach is impossible to achieve in a reasonable time. To overcome that problem, we propose an adaptative subsampling of the sensor configuration space which takes into account the limitations of the recognition system.

The basic idea of the treatment is to provide the rating function with configurations which meets certain requirements:

- For each configuration, a certain amount of points of interest must be visible.
- Points of interest must be seen under imaging conditions which allow a reliable recognition.
- Configuration must have a low coupling (their view field must weakly intercept).
- The set of all configurations must partition the visible space.

In order to achieve these criteria, we use the concept of visibility map introduced in [3]. Here we describe the steps leading to the construction of this map.

A given 3D point in the environment votes in the sensor configuration space for all the configurations from which it can be imaged under good conditions (conditions allowing a reliable recognition given a recognition method), this is what we call the visibility sphere of a point. This hollow sphere has an inner radius of R_{min} and an outer radius of R_{max} as defined in II-A. Figure 1 shows a 2D representation of a visibility sphere.

All the points of the visible 3D surface of the environment (unknown or solid voxels with an empty neighbor) create their own visibility sphere. The contribution of all the visibility spheres are summed up in an accumulation map we call the visibility map. Figure 2 shows a horizontal cut of the visibility map on which the two rotation dimensions are projected in order to allow a 2D representation.

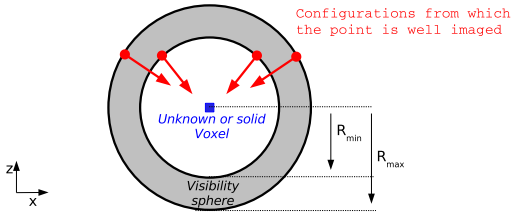

Fig. 1. The visibility sphere represents the 5D configuration set of the stereoscopic sensor in which a particular 3D point can be well recognized by a given recognition algorithm.

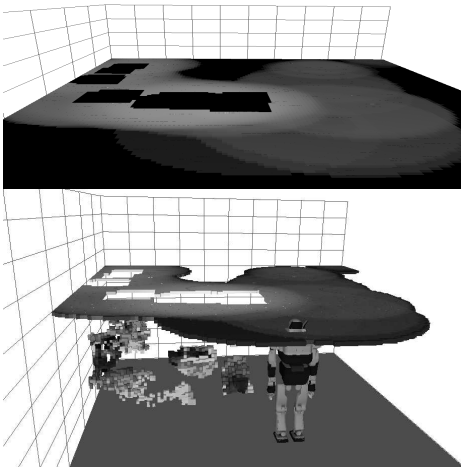

Fig. 2. This visibility map is only computed for reconstructed solid points (gray points under the plane). Each point is creating a visibility sphere around it. Lighter area on the plane represent configurations in which the solid points can be well imaged

The visibility map can be seen as a 5 dimension, gray values map:

- The value of each configuration in the visibility map is called the visibility of the configuration. A candidate is a configuration which has a non zero visibility.
- The set of candidates which have the same x and y parameter is called a cluster (the cluster visibility is the sum of all its candidates visibility). Figure 2 shows in fact the clusters of the visibility map.

The visibility sphere of a point is precomputed according to the R_{min}, R_{max} values and stored in a look-up-table (LUT). The visibility map update is done incrementally, which means that only points which have a change in their state will be considered: new boundary points add their votes to the visibility map and votes of removed points are subtracted. Because of its incremental nature, the visibility map computation gets faster (in average).

C. Local maxima extraction

In order to achieve the criteria listed in the previous section the visibility map is filtered: The coupling (figure 3) inside the same cluster is low because a change in the pan, tilt parameters will bring a lot of new information in the field of view. On the other hand, a change in the x, y, z parameters will most likely produce a small change in the field of view.

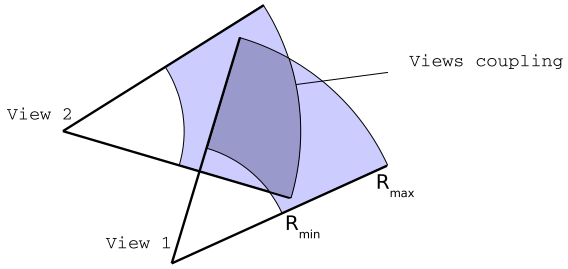

Fig. 3. Illustration of the coupling between views

A local maxima extraction of the visibility map based on a window with different size for the rotation and translation parameters will output the 'locally best' configurations for which a reasonable amount of points is visible. A small size is used for the pan and tilt parameter, reflecting the fact that configurations with close orientation values are weakly coupled. A larger window size is used on the translation parameters. In this paper we use a window of size 3 for rotation and 9 for translation (with a 5 cm resolution for the position and 18 $^\circ$ resolution in orientation).

The greedy exploration of sensor's parameter space is constrained to the local maxima of the visibility map. An interesting feature of the visibility map comes from the fact that solid and unknown points are treated the same way, and generate their visibility sphere, thus suitable configurations for exploring unknown areas are also created. The constraint achieved by the visibility map and the local maxima extraction drastically reduces the configurations to consider at each step. Typical values are around 1000 candidates (to compare with the 24 millions possible sensor placement). Next section will present the overall algorithm.

III. Algorithm

A. Overview

The flowchart of the next best view selection process is depicted in figure 4. When a new world model is available, the corresponding visibility map is computed and the local maxima extraction is performed providing a candidate list. The followings sections give the formulation of the rating function, and describe the different steps of the next view selection. More details regarding the rating function can be found in [3].

B. The probability world map

A discrete occupancy grid is generated by the stereoscopic sensor of the robot. Localization is done through a SLAM process [8] which merges odometric information provided by the walking pattern generator and visual information to provide accurate positioning.

The target presence is represented by a discrete probability distribution function p. Since this probability will be updated after each recognition action, it is a function of both position and time. $p(v_i, t)$ represents the probability that the voxel v_i

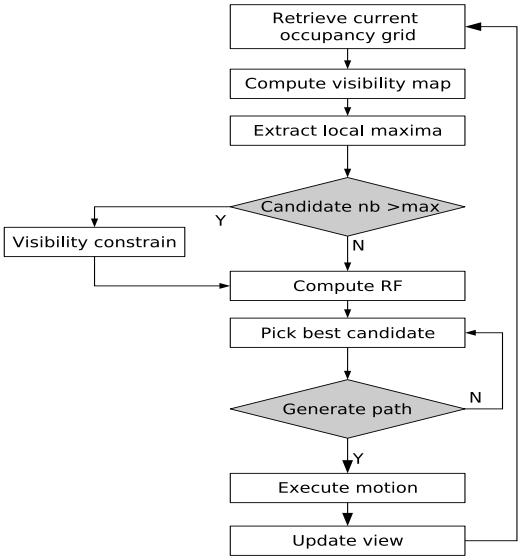

Fig. 4. Flowchart of the next view selection

is a part of the target. For a given camera configuration c,

$$P(c) = \sum_{\Psi(c)} p\big(v_i, t\big), \tag{2}$$

represents the probability that the object is inside the current field of view Ψ. The field of view Ψ takes into account occlusions for already mapped obstacles as well as the depth of field.

C. The rating function

The rating function must evaluate the interest of a given configuration according to different criteria:

1) the probability of detecting the object: the detection probability (DP),
2) the new volume that will be seen: the new information (NI),
3) the cost in time/energy to reach that configuration: the motion cost (MC).

The DP, NI and MC are combined in the rating function (3):

$$RF(c) = \lambda_{DP} \cdot DP(c) + \lambda_{NI} \cdot NI(c) - \lambda_{MC} \cdot MC(c), \tag{3}$$

where λ_{DP}, λ_{NI} and λ_{MC} are scaling factor to balance the contribution of each member of the rating function. This function will be optimized to select the next view.

The weights selection can modify the current strategy of the search:

- a high λ_{NI} will support a wide exploration of the environment,
- a high λ_{DP} will support a deep search of each potential target.

The table below gives the total distance traveled by the robot for 50 different views, and the remaining unknown voxels in the environment for different values of λ_{MC} ($\lambda_{NI} = 1000$).

λ_{MC}	0.01	0.1	0.2	0.5	2	3
Total distance (m)	91.3	71.4	56.3	45.7	21	16
Unknown (%)	13.8	13.7	13.7	16	21	19

a) The detection probability: From equation 1 and 2 we define the detection probability (DP) for a given camera parameter c as:

$$DP(c) = \sum_{\Psi(c)} p\big(v_i, t\big)\rho\big(c, v_i\big). \tag{4}$$

b) The new information: This concept already introduced by [9] and [5] is also used in the overall configuration rating process with a different formulation:

$$NI(c) = \frac{\sum\limits_{\Psi(c)} (v_i = unknown)}{\sum\limits_{Environment} (v_i = unknown).} \tag{5}$$

c) The motion cost: In addition to maximizing the NI and DP, it is also interesting to minimize the distance traveled to reach the configuration. The motion cost computation is based on an Euclidean metric in the configuration space of the sensor for the rotation parameters and on a navigation function (NF) based on a 2D projection of the occupancy grid for the the translation parameters of the sensor:

- $NF(c) = 0$ and the configuration is reachable without moving the waist of the robot:
  $\Rightarrow MC(c) = \sqrt{\alpha_{pan}\big(p' - p\big)^2 + \alpha_{tilt}\big(t' - t\big)^2}$
- $NF(c) = 0$ and the robot must rotate its waist:
  $\Rightarrow MC(c) = \alpha_{WR} \cdot (\theta_{waist} - \theta'_{waist})$
- $NF(c) \neq 0$:
  $\Rightarrow MC(c) = \alpha_{NF} \cdot NF(c)$

where α_{pan}, α_{tilt}, α_{waist} and α_{NF} are weights on each DOF. In this paper, α_{pan}, α_{tilt} are low and α_{NF}, α_{waist} are higher because moving the whole robot takes more time and energy than moving only the head.

Next section presents the optimization of this rating function in order to determine the next sensor placement.

D. Candidates examination

The local maxima extraction presented in section II-C provides us with a list of candidates. If the candidates are too numerous, a visibility constraint is applied and the best candidates are taken (i.e. candidates which received a maximum amount of votes). The number of candidates that can be sent to the rating function depends on the reaction time we want to achieve. Typically we set a limit of 1000 candidates to rate. The current implementation of the rating function takes (initially) 3 ms per candidate, thus in the worst case, it takes up to 3 sec to plan the next view. These steps are depicted in figure 4.

E. The path planning & the recognition function

Once the next sensor placement is decided, a path is planned to reach the desired position. Because the navigation function only gives an optimistic evaluation of reachable configurations, some target locations are rejected by the planner. In such a case, the second rank candidate is picked up and a new path is computed. We currently use an A^* planner which only takes into account the bounding box of the robot, uses a discreet set of orientations and does not allow any backward motion. Theses limitations are clearly visible in the experiment we present at the end of the paper but does not interfere with the proposed search model.

Once the target configuration is reached, the world model is updated, and the recognition of the object is attempted. Few assumptions are made on the underlaying recognition system and the output of the recognition is supposed to be a list of object poses with their associated likelihoods. Each object pose is then converted into the corresponding voxel set and their probabilities are merged with the target presence probability map through the update process. The update process will then normalize the distribution probability in order to have: $\sum_{Environment} p(v_i, t) = 1$. The process is then reiterated until the object position knowledge reaches locally a given threshold in the probability distribution.

IV. Experiments

V. Simulation results

A full search behavior has been tested in simulation (Figure 5). The environment is a 6x4x1.5 meter room with two obstacles, the target is hidden behind the large obstacle. A simulation of the recognition system has been implemented, although simple, it has the main characteristics of a real recognition system with false target detection that adds some noise to the probability map. In the simulation, the robot finds the object after 15 views. Depending on the settings (the $\lambda_{NI}/\lambda_{DP}$ ratio) the robot will lock the target after the first view or will do some remaining exploration before focusing its attention on the target.

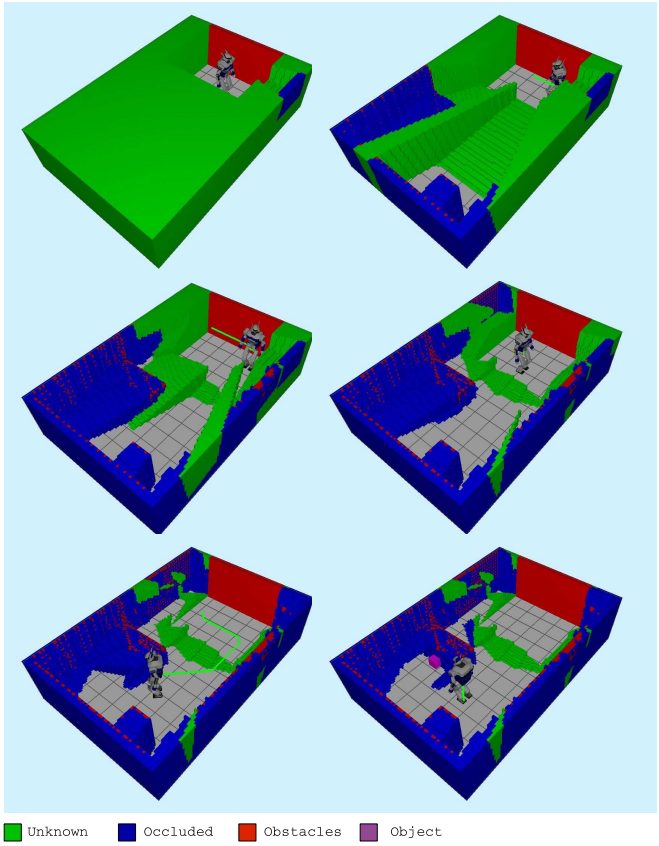

Fig. 5. Image sequence of the search behavior, the object is hidden behind the large obstacle

VI. Experimental setup and results

Real experiment using HRP-2 humanoid robot has been successfully achieved. The recognition system is a color detector based on a normalized color histogram. The 3D position of the center of the color region detected in an pair of image is computed using the camera calibration information. The matching score is proportional to the size of the segmented color region, the closer this size is to real object size, the higher the matching score will be.

The experimental room (Picture in figure 6) is 6 by 4 meters and is divided in two parts by a 2 meters wide panel. Figure 7 shows an image sequence captured form the control interface during the experiment. The environment reconstruction is only based on disparity information and needs to be well textured. The aim of the experiment we present here, was to have a full exploration and mapping of an unknown environment, in order to validate the model in a real world experiment in presence of heavy reconstruction noise and localization error. There was no target object hidden in the room, thus the planning was mainly driven by the new information retrieval even though the detection probability was taken into account in the optimization process. The whole exploration is done in 29 views, the robot finishes exploring the first part of the room in 23 views after mapping enough environment to allow a planning to explore behind the wall for the 6 remaining views.

VII. Conclusion

This paper exposed the framework for a search behavior developed for the humanoid robot HRP-2. The problem, which falls in the sensor planning field, is formulated as an optimization problem. The concept of visibility map introduced in [3] to constrain the sensor parameter space according to the detection characteristics of the recognition algorithm is used to reduce the dimension of the sensor parameter space. Simulation and real experiments using the HRP-2 robot have been achieved to validate the proposed search model. A more powerful path planner such as KineoWorks is on the way to be integrated to provide full body, 3D planning for the robot, and will allow sensor placement to be less constrained. Moreover, a better recognition system based on feature point matching is under development and will allow to predict the pose of a partially imaged object.

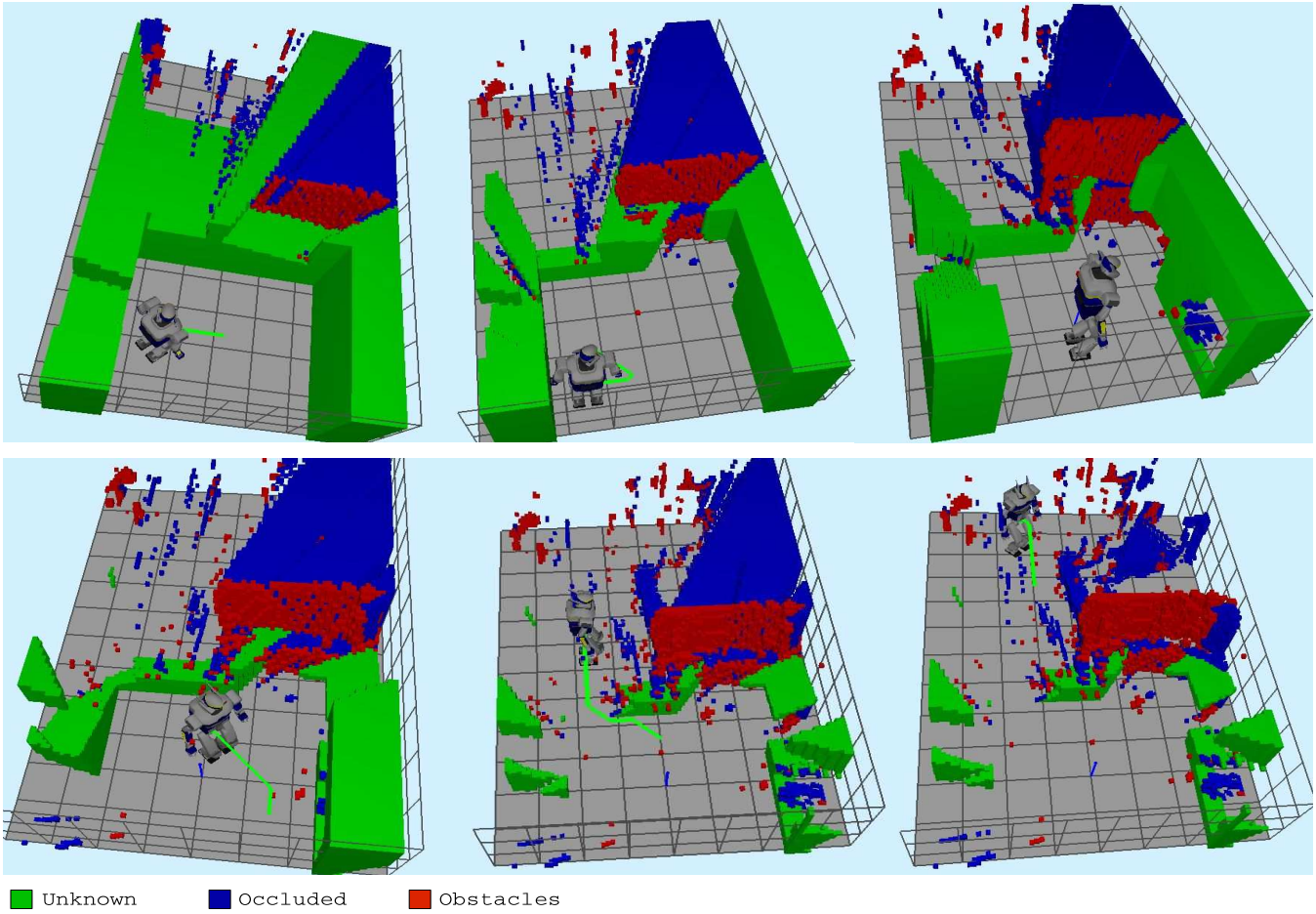

Fig. 7. Images of the real environment exploration as seen through the control interface during the experiment. We notice the heavy reconstruction noise mainly due to false point matching

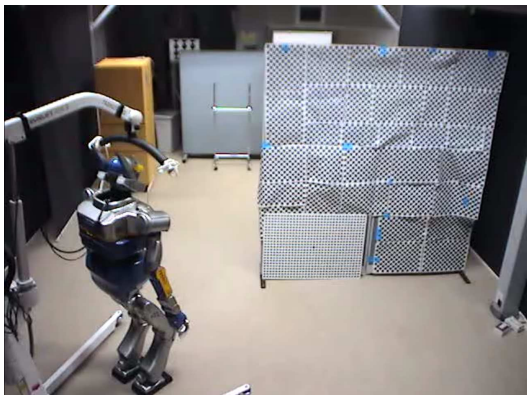

Fig. 6. A picture of the experimental room during HRP-2 exploration

This predicted pose will then be merged in the probability map in order to generate sensor configuration from which the prediction will be confirmed or not.

ACKNOWLEDGMENT

This research was partially supported by a Post-doctoral Fellowship of Japan Society for Promotion of Science(JSPS) and JSPS Grand-in-Aid for Scientific Research.

REFERENCES

[1] Y. Ye and J. K. Tsotsos, "Sensor planning in 3d object search: its formulation and complexity," in *Fourth International Symposium on Artificial Intelligence and Mathematics*, Florida, U.S.A., January 3-5 1996.

[2] ——, "Sensor planning for 3d object search," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 145–168, Feb. 1999.

[3] F. Saidi, O. Stasse, and K. Yokoi, "A visual attention framework for a visual search by a humanoid robot," in *IEEE-RAS International Conference on Humanoid Robots*, Genova, Italy, December 4-6 2006, 346-351.

[4] L. E. Wixson, "Gaze selection for visual search," Ph.D. dissertation, Department of Computer Science, Univ. of Rochester, 1994.

[5] V. A. Sujan and S. Dubowsky, "Efficient information-based visual robotic mapping in unstructured environments," *The International Journal of Robotics Research*, vol. 24, no. 4, pp. 275–293, Apr. 2005.

[6] B. O. Koopman, *Search and Screening*. Pergamon Press, 1980.

[7] C. J. Connolly, "The determination of next best views," *IEEE Int. Conf. on Robotics and Automation*, pp. 432–435, 1985.

[8] O. Stasse, A. Davison, R. Sellaouti, and K. Yokoi, "Real-time 3d slam for humanoid robot considering pattern generator information," in *International Conference on Intelligent Robots and Systems, IROS*, Beijing, China, October 9-15 2006, to appear.

[9] A. Makarenko, S. Williams, F. Bourgault, and H. Durrant-Whyte, "An experiment in integrated exploration," in *IEEE/RSJ International Conference on Intelligent Robots and System*, vol. 1, 2002, pp. 534 – 539.