

# Multi-sensor semantic mapping and exploration of indoor environments

Islem Jebari, Stéphane Bazeille, Emmanuel Battesti,  
Hassene Tekaya, Marius Klein, Adriana Tapus, David  
Filliat  
Cognitive Robotics Theme, ENSTA ParisTech  
Paris, France

islem.jebari@ensta-paristech.fr

Cédric Meyer, Sio-Hoï Ieng, Ryad Benosman  
Institut Des Systemes Intelligents et de Robotique (ISIR-  
CNRS)  
Paris, France

Eddy Cizeron, Jean-Charles Mamanna, Benoit Pothier  
GOSTAI S.A.S  
Paris, France

*Abstract*—The human perception of the external world appears as a natural, immediate and effortless task. It is achieved through a number of “low-level” sensory-motor processes that provide a high-level representation adapted to complex reasoning and decision. Compared to these representations, mobile robots usually provide only low-level obstacle maps that lack such high-level information. We present a mobile robot whose goal is to autonomously explore an unknown indoor environment and to build a semantic map containing high-level information similar to those extracted by humans and that will be rapidly and easily interpreted by users to assess the situation. This robot was developed under the Panoramic and Active Camera for Object Mapping (PACOM)<sup>1</sup> project whose goal is to participate in a French exploration and mapping contest called CAROTTE<sup>2</sup>. We will detail in particular how we integrated visual object recognition, room detection, semantic mapping, and exploration. We demonstrate the performances of our system in an indoor environment.

*Keywords:* SLAM; object detection; autonomous exploration; semantic mapping;

## I. INTRODUCTION

The Panoramic and Active Camera for Object Mapping (PACOM) project addresses the understanding of how an autonomous embodied system can build and extract information from sensory and sensory-motor data and generates plans and actions to explore and navigate in typical indoor environmental settings. In particular, we seek to extract high-level semantic information that is easy to understand and interesting to the robot users such as surrounding objects and the environment structure. To achieve this goal, the system

---

<sup>1</sup> The PACOM project is supported by DGA in the frame of the “CAROTTE” competition and funded by ANR under the subvention 2009 CORD 102.

<sup>2</sup> CAROTTE is organized by the French research funding agency (ANR) and the French armament procurement agency (DGA). Website: <<http://www.defi-carotte.fr>>

requires different sensing modalities and also needs to act in order to improve its understanding of the environmental situation or to disambiguate its interpretation.

The project goal is to participate in the CAROTTE challenge that proceeds over three years with an increase in the difficulty over the years. The competition between 5 selected teams takes place in an arena of approximately 120m<sup>2</sup> where objects are laid. The environment contains several rooms typically 10 or more, with variable grounds and various difficulties (fitted carpet, tiling, grid, sand, stones,...). Several kinds of objects are present, either isolated or gathered, in multiple specimens, which must be detected, located, and identified or characterized by the robot. The objects used in the competition are: computers, boxes, paperboards, journals, books, telephones, keys, pens, staplers, cables, lamps, weapons, ammunition, cans, bottles, plants, cameras, radios, ventilators, balloons, robots. The complete description of the challenge can be found on the website<sup>2</sup>.

After a review of the related work, this paper will detail the hardware and software architecture of our robot in Section III, describe the software components used in Section IV before giving experimental results in Section V.

## II. RELATED WORK

This project involves the integration of several research areas for which we will make a short survey: localization and mapping for mobile robots, object detection and recognition and exploration of unknown environments.

The problem of Simultaneous Localization and Mapping (SLAM) of an unknown environment by a mobile robot is the subject of a lot of research for more than 20 years. Today, very robust solutions exist for SLAM in planar environments using 2D scanning laser sensors to the point where several effective commercial or open source software packages are available. Research on SLAM using laser sensors is now focused on 3D SLAM [1] which is made possible by combining multiple planar laser scanners with different orientations or by using a planar laser sensor rotating around a second axis. The most active research area in SLAM is probably visual SLAM with

different approaches including metric landmark-based SLAM using monocular vision [2], panoramic vision [3], or stereovision [4] and topological SLAM using panoramic [5] or monocular vision [6]. However, in an autonomous exploration context, these approaches are limited, as they do not usually provide an estimation of free space for robot movement when using panoramic or standard camera, and are not reliable when facing texture-less walls in indoor environment when using stereovision. Vision in general is also limited by the need of ambient light, which could be absent in some environments of interest. Finally, several researchers integrated laser sensor and panoramic vision [7], which provides the robustness of obstacle sensing by laser and very informative visual data. Beside these approaches that are mainly directed toward low level robot localization and navigation, several other approaches have been proposed to introduce higher-level semantic information into maps. This includes the classification of space into different categories such as rooms, corridors [8], roads, buildings [9] and the addition of objects in a hierarchical map representation. One related problem in computer vision is the use of contextual information to enhance object detection and recognition [10].

In the field of object detection, several approaches enable the robot to recognize objects using vision in complex and realistic environments. Many of these approaches are based on robust, scale invariant keypoint detection and recognition (e.g., SIFT [11], SURF [12]) or on machine learning to achieve fast and reliable keypoint recognition [13]. Beside individual object recognition, these features can be used for object class recognition using the “bag of keypoints” approach [14]. Object recognition is also possible using 3D information taken from a laser scanner [1].

The exploration of an unknown environment using a 2D scanning laser sensor is often performed using the frontier-based exploration algorithm [15] or an exploration strategy choosing the next best position for the robot given the utility of this position for the mapping problem [16]. In the context of this project, the exploration should also take into account the need to search for objects. This general problem of optimally covering the whole environment with several viewpoints is referred to as the “art gallery problem” [17] and is known to be NP-hard. However, several practical solutions exist, for example by optimizing the probability of finding the target given a limited number of actions [18]. It should be noted that several work on semantic mapping do not mix the problems of unknown environment exploration and object search, either by relying on a human interaction to guide exploration or by using a previously built map [19].

### III. SYSTEM OVERVIEW

We developed a robot (see Fig.1 Right) based on a pioneer 3 dx from Mobile Robots Inc. The robot was fitted with 2 scanning laser range finders (one horizontal SICK LMS 200 and one vertical Hokuyo UTM 30 LX), a ring of sonar sensors, a Pan-Tilt-Zoom camera and three on-board computers. Our robot has not been optimized for the competition only as it can be used for assistive applications for which the functionalities of semantic mapping are particularly interesting.

The software architecture (see Fig.1 Left) uses the Urbi framework; a middleware for programming complex robotic systems developed and supported by Gostai<sup>3</sup>. Urbi is composed of a distributed component architecture (UObject), and an innovative orchestrator language (urbiScript) to coordinate all components. This language incorporates high-level features that facilitate the development of parallel and event-based applications.

For the project, we thus developed a set of UObjects in C++ carrying out the various necessary functionalities. The whole mission of the robot is implemented in urbiScript which uses these UObjects functionalities and coordinates their activation.

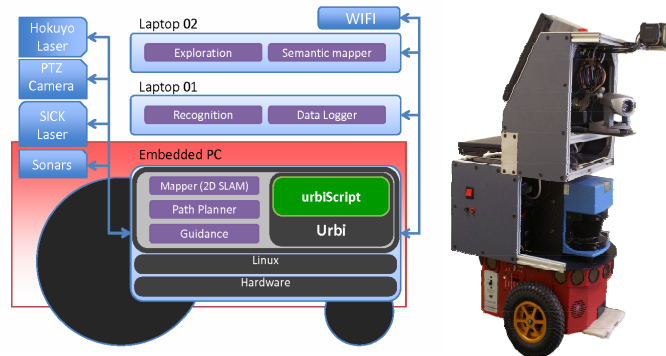


Figure 1. Left: The structure and distribution of the various processes in the robot. Right: The hardware architecture of the robot

A simplified version of the urbiScript code for the competition’s mission is shown on Figure 2. It uses the UObjects described in the next section.

```

every(500ms) { logger.savePosition();}

every(300ms)
{ if(robot.isMoving())
  { image_buffer.pushBack(camera.val);}}

whenever(!image_buffer.empty())
{recognition.localizeObjects(image_buffer.removeFront());}

while(execution time<max_time &&
!exploration.complete())
{ exploration.computeNextPosition();
robot.goToPosition(exploration.nextPosition);
for( var pan = pi; pan >= -pi; pan -= pi/4)
{ robot.orientCamera(pan,tilt);
image_buffer.pushBack(camera.val);}
mapper.updateMap();
exploration.updateMap();
}

robot.goToPosition([0,0,0]);

semantic.addInformationToTheMap(map);

```

Figure 2. Part of the urbiScript code implementing the robot mission. Interesting features are the use of “every” for background position and image logging during robot movement and the use of “whenever” for asynchronous image processing in order to be able to move the robot during object detection

<sup>3</sup> Gostai S.A.S. Website : <<http://www.gostai.com>>

#### IV. SYSTEM OVERVIEW

The main software components developed within our project are: Exploration, Object recognition and Semantic mapping. In addition to these main parts, we also integrated other components for Mapping, Path planning, Guidance and Data Logging.

Mapping performs 2D SLAM using the horizontal laser scanner and the Karto software library<sup>4</sup>, which provides good performance and robustness in indoor environments. This library uses scan matching to correct the robot odometry drift and provides a 2D occupancy grid map. Path planning performs global path planning given the current map of the environment and carries out local obstacle avoidance of dynamic objects taking into account lasers and sonars. This module is also based on the Karto library. Guidance controls the robot to execute the computed path. Data Logging records the data and the results of the mission.

##### A. Exploration

In our context, exploration is required to completely map the environment using the horizontal laser scanner and to search for objects detected using the pan-tilt camera. We took advantage of the fact that the two sensors have similar field of view (a semi-circle in front of the robot) to integrate these two objectives into a single algorithm.

Computing an optimal set of positions to cover the environment is a NP-hard problem. We therefore use a stochastic sampling strategy inspired by the Randomized Art-Gallery Algorithm [20] to search for the next robot position that discovers the more unseen area. In order to minimize robot movements, N positions are sampled inside the free space in a semi-circled area in front of the robot (N=50 in this paper). If no reachable position or no unseen area can be found in this area, positions are sampled from the whole map to ensure global exploration.

Each sampled position is evaluated using a scoring function that takes object search and mapping into account. For object search, a score  $S_{obj}$  is computed as the size of the area visible through the camera that has not been observed yet. The visible area takes the pan-tilt capability of the camera and the minimal resolution needed for object recognition into account. The unobserved area is computed using a 2D view-map that records the area previously perceived by the robot camera. For mapping, a score  $S_{map}$  is computed as the number of frontier cells between free and unknown area [21] that are visible through the laser sensor from the position. Finally, a score  $S_{dist}$  is computed as the inverse of the travel distance from the current position. The final score is a weighted sum of these components:

$$S = S_{obj} + \lambda_{map} S_{map} + \lambda_{dist} S_{dist} \quad (1)$$

The sampled position with the highest score is taken as the next robot position. The exploration is performed until a given percentage (98% in this paper) of the free area has been observed by the camera.

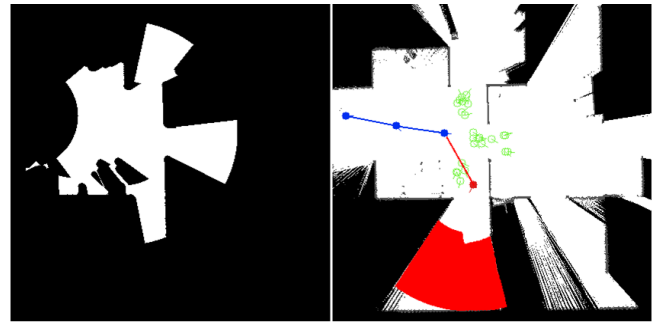


Figure 3. Illustration of the exploration strategy. The left part shows the view-map recording already seen areas. The right part shows the exploration trajectory, the sampled positions (in green) and the selected next position with the next observed area (in red)

##### B. Object recognition

The goals of the object recognition task are: first, to detect the objects in the robot’s visual environment and secondly, to estimate their location in the camera reference frame. Object detection is performed using a two step method. The first step includes a fast extraction of the salient regions [22] in each image in order to segment the image to allow multiple object detection and to improve detection and localization speed by reducing the regions of interest. The second step is based on local features extraction [12, 23] and a bag of visual words approach [24]. Once an object is detected, its localization is estimated using additional information on objects or with a second step of dense features extraction [11] followed by the computation of the geometric transformation between the real object and a previously learned image. Objects to detect are known “a priori” and a database has been built containing each object over different points of view and different known distances.

###### a) Image segmentation

Segmentation is performed by computing the entropy based on the measure of color and gradient rarity [22]. Each image (Fig.4-1) is subsampled 10 times to 64x48 pixels in order to accelerate computation. The saliency map is computed (Fig.4-2) and is used to find local maximums that are passed through a region growing and fusion algorithm to find stable salient regions (Fig.4-3). The smallest regions are then filtered in order to find the candidate objects (Fig.4-4) that are represented by a rectangular bounding box.

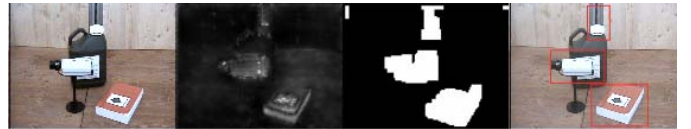


Figure 4. Steps of image segmentation : the first image is the original image taken by the robot, second is the saliency map computed on the subsampled image, third is the saliency map once thresholded and passed through region growing and fusion algorithm. The last one shows the salient regions drawn on the original image

<sup>4</sup> Karto Robotics : Website <<http://www.kartorobotics.com/>>

### b) Object detection

The detection algorithm is based on a bag of visual words approach [24] (Fig.5) in which images are only represented by the occurrence frequency of local features taken from a dictionary. In our approach, two dictionaries are built by extracting SURF keypoints [12] and local color histograms [23] from each image of the learning database. Descriptors are then clustered using hierarchical k-means. These tree structures [25] are used to improve the matching speed between descriptors and visual words.

Object learning is performed by computing visual words corresponding to the features extracted in the object images, and by associating to each visual word its occurrence frequency for each object. The structure obtained is a reverse index allowing fast object detection from image descriptors.

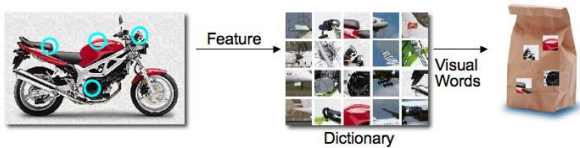


Figure 5. Bag of visual words representation

For object recognition during the robot exploration, visual words are extracted from salient regions of each image taken by the camera. The inverse index is then used with a Bayesian voting method [23] which indicates the presence of a learned object in each of the salient region.

### c) Object localization

When an object has been detected in a sub-image, it has to be localized in order to be added to the semantic map. For this, we either compute the geometric transformation between the detected object and its image in the database when the object is textured, or we suppose that the object is put on the ground when the object is colored but textureless.

The geometric transformation calculates a set of matching points between the object to localize and a learned image using SIFT descriptors [11] (SURF descriptors are not dense enough to get enough correct matches on our images). Pairs of locations of matched points are given as an argument to an algorithm that computes the best homography using RANSAC method [26] in order to eliminate outliers, which increase the accuracy of the detected position of the object in the captured image. The distance of the recognized object is computed by using each inter-pair distance between the matched points to estimate scale difference between the two images. The matching method allows object localization even if they are partially hidden. Given that the density of SIFT descriptors is very low on uniformly colored objects, we assume that colored objects are put on the ground, which is quite often in the context of CAROTTE. We just have to determine the object position in the image and to intersect a line of view with the ground plane. In order to accelerate this step, we use SURF visual words already extracted by the object detection algorithm previously discussed in order to find the center of object (pink crosses in Fig.6).

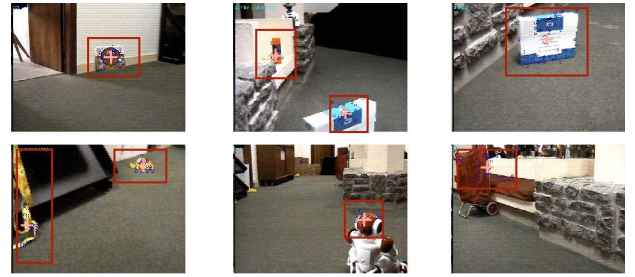


Figure 6. Result of the object localization during an exploration. Red rectangles correspond to salient regions where objects have been detected (from a set of 10 previously learned objects), blue circles are the visual words extracted in the salient regions (SURF descriptors and local color histograms) and pink crosses are the estimated position of object center used to compute object distance

As we can see in Figure 6, we have a reasonable localization of the object in the image. However, if the object is big or not totally settled on the ground, the estimated distance of the object has greater errors which could be corrected using Kalman filters on the estimated objects position as shown in the next section.

### C. Semantic mapping

“Semantic mapping” adds meaningful information to the 2D map such as the 3D structure of the environment, the position, name and images of the detected objects and the rooms. The 3D structure is built as a point cloud using the second laser sensor, assuming that the “mapping” module gives a correct position. We estimate the object position by integrating multiple detections using Kalman filter [27]. The application of this method is particularly indicated in our case because we have an external estimate of the robot position. Rooms are detected in the occupancy grid through an algorithm that detects doors and analyses the resulting connected components of open space.

The 2D semantic mapping consists in calculating the coordinates of the object detected in the absolute reference frame. The object detection module gives the identifier corresponding to the object category, but it should be decided if this object has already been seen or if it is a new exemplar of the same object. For this, we calculate the jacobian matrix associated to the transformation from the image position to the map position in order to estimate the uncertainty of the position of the object in the world frame. We use the Mahalanobis distance to decide if a perception corresponds to an object already present in the map or not. Thus, for each object already placed in the map corresponding to the same category as the detected object, we compute the Mahalanobis distance with the detected object and if the minimal value of the Mahalanobis distance is higher than a certain threshold, we add the object detected as a new object; If not, we update the position of the object recognized with the Kalman filter.

For the 3D semantic mapping, we chose an approach rather similar to the object mapping. The positions of the obstacles detected by the second laser, directed vertically (with a slope of 60° compared to the horizontal one) are calculated in the world reference and are memorized. This position is calculated by supposing exact the position of the robot given by the 2D



SLAM. We thus obtain a 3D point cloud representing partially the environment according to the movements of the robot that is used to categorize the various types of obstacles and to assist the target detection. We use the PCL (Point Cloud Perception) library to characterize the structure of the environment (corridors, walls,...) by selecting a relevant volume, filtering the noise, reducing the density of the points used and extracting the planes corresponding to walls with RANSAC algorithm.

An illustration of the algorithm used to detect horizontal planes and extract objects is shown on Figure 7.

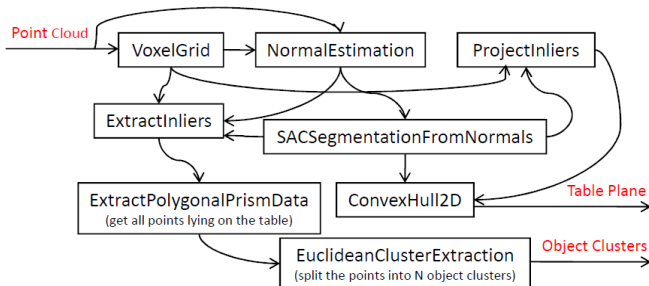


Figure 7. Illustration of table detection and object extraction with PCL library

### V. EXPERIMENTAL RESULTS

For the experimental part, we carried out tests in an indoor environment of an approximate surface of 61,2 m<sup>2</sup>, containing 3 rooms and a corridor. Our robot needed about 733s to explore the environment and come back to the starting point. During the mission, the Pan-Tilt camera took about 1400 images during the movement and 10 images at each exploration point.

The final result corresponding to the semantic map obtained at the end of the run is shown on Figure 8. Green color corresponds a correct detection with a correct label. Blue color corresponds to a correct detection with a false label. Red color refers to a false detection and pink color refers to missed objects. We note that the missed objects are small objects that are located in the corner of the rooms. Some statistics related to the detection module are shown below.

correct detection / correct label	61,55 %
correct detection / false label	7,69 %
false detection	15,38 %
missed objects	15,38 %

To explore the whole environment, the robot needed 20 exploration points which are shown and numbered in Figure 9. The robot trajectory is quite complex because of the visual exploration strategy which tries to discover a big unknown place first. The visual area covered by the camera is shown on Figure 10 (Right) and room detection is shown on Figure 10 (Left).

An interface (see Fig.11) displaying the map, the robot trajectory, the object detection in 3D and the mission replay has also been developed using the Peekabot library.

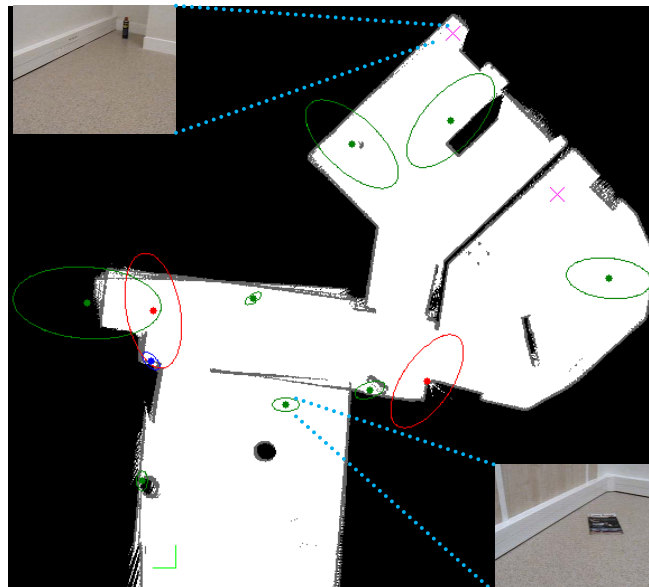


Figure 8. Semantic map showing the environment structure and the objects position with their ellipses of uncertainty. The starting point is shown as a green set of axes. One detected object and one missed object are shown

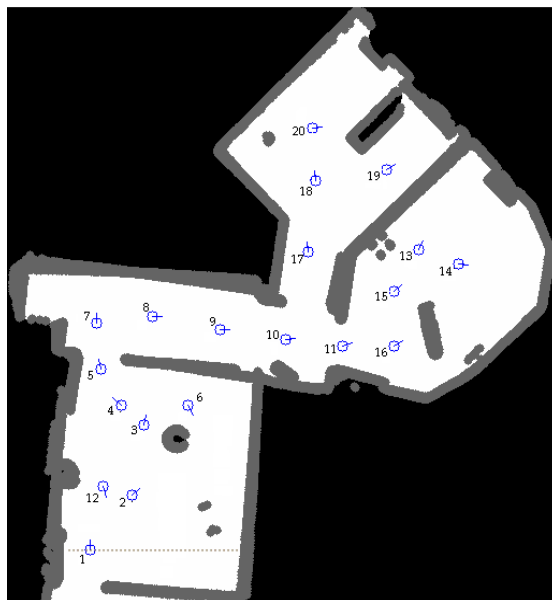


Figure 9. Exploration map: blue points correspond to exploration points with the robot orientation

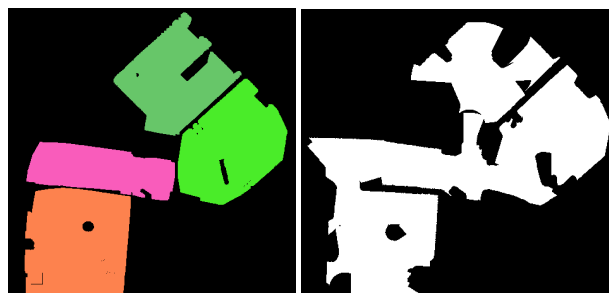


Figure 10. Left: Room detection. Right: view-map showing areas perceived by the camera (91% of the entire surface)

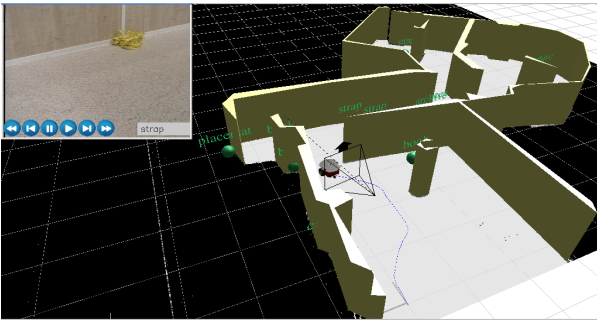


Figure 11. Interface displaying the map and the object detection

## VI. CONCLUSION & PERSPECTIVES

In this paper, we presented our solution to participate in the CAROTTE competition. This first year enabled us to design a competitive mechanical system and a robust hardware and middleware platform. Moreover, concerning the software part, we have validated the mapping and the exploration modules. The laser resulting maps are clean and accurate and the area covered by vision is nearly complete. The object detection module gave good results thanks to several approaches of object detection, using color and keypoints. As a result, semantic maps are mainly faithful to the real world. The guidance module revealed some problems relative to the avoidance of some obstacles like see-through walls. To overcome this default, we are working today to combine effectively laser and sonars data.

For the second year of the challenge, the software robustness will be more deeply evaluated because many new constraints will be added to the competition like a composite ground (carpet, grid, sand, grass, gravel...), more objects (real and virtual) and more complex environments (multi-level, mirror wall...).

We are currently improving the 3D mapping using the vertical laser combined with a high resolution color camera and a time of flight camera, which have the advantage of fast imaging and high lateral resolution combined with the depth information of the captured scene. We are also integrating depth perception in order to improve object recognition.

## REFERENCES

- [1] R.B. Rusu, N. Blodow, Z. Marton, A. Soos and M. Beetz, "Towards 3D object maps for autonomous household robots," *Intelligent Robots and Systems (IROS 2007)*, pp. 3191-3198, 2007.
- [2] A.J. Davison, I.D. Reid, N.D. Molton and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," *IEEE Trans. PAMI*, 2007.
- [3] T. Lemaire, S. Lacroix, "Long term SLAM with panoramic vision," *Journal of Field Robotics*, 24(1): 91-111, 2007.
- [4] B. Morisset, R.B. Rusu, A. Sundaesan, K. Hauser, M. Agrawal, J.-C. Latombe and M. Beetz, "Leaving Flatland: Toward Real-Time 3D Navigation," In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, May 2009.
- [5] O. Booij, B. Terwijn, Z. Zivkovic and B. Kröse, "Navigation using an appearance based topological map," In *proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2007.
- [6] A. Angeli, D. Filliat, S. Doncieux and J.-A. Meyer, "A Fast and Incremental Method for Loop-Closure Detection Using Bags of Visual Words," *IEEE Transactions On Robotics*, Special Issue on Visual SLAM, 2008.

- [7] A. Tapus, N. Tomatis and R. Siegwart, "Topological Global Localization and Mapping with Fingerprint and Uncertainty," In *Proceedings of the International Symposium on Experimental Robotics (ISER)*, Singapore, 2004.
- [8] O.-M. Mozos, R. Triebel, P. Jensfelt, A. Rottmann and W. Burgard, "Supervised semantic labeling of places using information extracted from laser and vision sensor data," *Robotics and Autonomous Systems Journal*, 55(5):391-402, 2007.
- [9] D.F. Wolf and G.S. Sukhatme, "Semantic Mapping Using Mobile Robots," *IEEE Transactions on Robotics*, vol. 24, Issue 2, pp. 245-258, April 2008.
- [10] A. Torralba, K. Murphy, W. Freeman and M. Rubin, "Context-based vision system for place and object recognition," *ICCV'03 (Intl. Conf. on Computer Vision)*, 2003.
- [11] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 60, 2, pp. 91-110, 2004.
- [12] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "SURF: Speeded Up Robust Features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, No. 3, pp. 346-359, 2008.
- [13] S. Hinterstoisser, S. Benhimane, N. Navab, P. Fua and V. Lepetit, "Online Learning of Patch Perspective Rectification for Efficient Object Detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska (USA), 2008.
- [14] G. Csurka, C. Dance, L. Fan, J. Williamowski and C. Bray, "Visual categorisation with bags of keypoints," *ECCV04 workshop on Statistical Learning in Computer Vision*, pp. 59-74, 2004.
- [15] B. Yamauchi, "Frontier-based exploration using multiple robots," In *Proceedings of the Second international Conference on Autonomous Agents*, Minneapolis, Minnesota, United States, May 1998.
- [16] B. Tovar, L. Muñoz-Gomez, R. Murrieta-Cid, M. Alencastre-Miranda, R. Monroy and S. Hutchinson, "Planning Exploration Strategies for Simultaneous Localization and Mapping," *Journal on Robotics and Autonomous Systems*, vol. 54(4), pp. 314-331, 2006.
- [17] T.C. Shermer, "Recent results in art galleries," In *Proceedings of the IEEE*, pp. 1384-1399, 1992.
- [18] J.K. Tsotsos and K. Shubina, "Attention and Visual Search : Active Robotic Vision Systems that Search," *The 5th International Conference on Computer Vision Systems*, 2007.
- [19] K. Sjö, D.G. Lopez, C. Paul, P. Jensfelt and D. Kragic, "Object Search and Localization for an Indoor Mobile Robot," *Journal of Computing and Information Technology - CIT 17*, vol.1, pp. 67-80, 2009.
- [20] H. Gonzalez-Banos and J. Latombe, "A randomized Art-Gallery Algorithm for Sensor Placement," In *Annual Symposium on Computational Geometry*, 2001.
- [21] B. Yamauchi, "A frontier-based approach for autonomous exploration," In *proceedings of the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 1997.
- [22] N.J. Butko, L. Zhang, G.W. Cottrel and J.R. Movellan, "Visual Saliency Model for Robot Cameras," In *International Conference on Robotics and Automation*, 2008.
- [23] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2007.
- [24] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," In *Proceedings of the International Conference on Computer Vision*, 2003.
- [25] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006.
- [26] R.C. Bolles and M.A. Fischler, "A RANSAC-Based Approach to Model Fitting and Its Application to Finding Cylinders in Range Data," *IJCAI 1981*: 637-643, 1981.
- [27] R.C. Smith and P. Cheeseman, "On the Representation and Estimation of Spatial Uncertainty," In *The International Journal of Robotics Research* 5 (4): 56-68, 1986.