TOWARDS HUMAN INSPIRED SEMANTIC SLAM

Dominik Maximilián Ramík, Christophe Sabourin and Kurosh Madani
Signals, Images, and Intelligent Systems Laboratory (LISSI / EA 3956), Université Paris-Est
Senart Institute of Technology, Avenue Pierre Point, 77127 Lieusaint, France

Keywords: SLAM, Semantics, Humanoid robotics, Human inspired, Image segmentation, Scene interpretation.

Abstract: Robotic SLAM is attempting to learn robots what human beings do nearly effortlessly: to navigate in an unknown environment and to map it in the same time. In spite of huge advance in this area, nowadays SLAM solutions are not yet ready to enter the real world. In this paper, we observe the state of the art in existing SLAM techniques and identify semantic SLAM as one of prospective directions in robotic mapping research. We position our initial research into this field and propose a human inspired concept of SLAM based on understanding of the scene via its semantic analysis. First simulation results, using a virtual humanoid robot are presented to illustrate our approach.

1 INTRODUCTION

In mobile robotics, the ability of self-localization is crucial. In fact, knowing precisely where the robot is and what kind of objects surround it in a given moment enables it to navigate autonomously. An informal definition of the Simultaneous Localisation And Mapping (SLAM) says it as a process, in which a mobile robot explores an unknown environment, creates a map of it and uses it simultaneously to infer its own position. The real environment is usually complex and dynamic and it is not easy to interpret. This complexity makes SLAM a challenging task. A comprehensive list of nowadays most common SLAM techniques can be found in (Durrant-Whyte, et al., 2006a), (Durrant-Whyte, et al., 2006b) or (Muhammad, et al., 2009). Although from its beginning a significant advance has been achieved (Thrun, et al., 2008), SLAM is not yet a solved problem. Performing SLAM in dynamic environment (Hahnel, et al., 2003) or understanding the mapped environment by including semantics into maps (Nüchter, et al., 2008) are the actual challenges.

In this paper, the state of the art in SLAM is investigated. A relatively new field of research is identified, which is attempts to perform SLAM with the aid of semantic information extracted from sensors. As one of the research interests of our laboratory (LISSI) is autonomous robotics notably in relation to humanoid robots, we are convinced that the research on semantic SLAM will bring a useful contribution. We position our initial research into this field, drawing our inspiration from the human way of navigation. Contrary the precise and "global" approach to most current SLAM techniques, the human way of doing is based on very fuzzy description of the world and it gives preference to local surroundings of the navigation backdrop. A simulation using a humanoid robot Nao is presented to demonstrate some of the proposed ideas. The real Nao will be used in our further work.

The paper is organized in the following way: section 2 focuses on the state of the art in semantic SLAM. In the third section, our approach to image segmentation and scene interpretation is discussed. Section 4 gives an overview of our robotic humanoid platform. The fifth section presents our initial results and the paper concludes with section 6.

2 SEMANTIC SLAM

One of the latest research directions on the field of SLAM is the so-called semantic SLAM. The concept may be perceived as being very important for future mobile robots, especially the humanoid ones, which will interact with humans and perform tasks in human-made environment. In fact, it is this interaction, which is one of important motives for employing semantics in robotic SLAM as humanoids are particularly expected to share the living space with humans and to communicate with them.

Maximilián Ramík D., Sabourin C. and Madani K. (2010).
TOWARDS HUMAN INSPIRED SEMANTIC SLAM.
In *Proceedings of the 7th International Conference on Informatics in Control, Automation and Robotics*, pages 360-363
DOI: 10.5220/0002912303600363

One way of adding semantics to SLAM may be the introduction of human spatial concepts into maps. Humans usually do not use metrics to locate themselves but rather object-centric concepts ("I am near the sink" and not "I am on [12, 59]") and they fluently switch between reference points rather than using global coordinates. Moreover, the presence of certain objects is often an important clue to recognize a place. This problem is addressed in (Vasudevan, et al., 2007). Here, the world is represented topologically; place recognition is performed based on probability of presence of objects in an indoor environment. The work shows a study aimed to understand human concepts of place recognition. It proposes that humans understand places by presence or absence of significant objects. Place classification by presence of objects has been used by (Galindo, et al., 2005), where low-level spatial information is linked to high-level semantics. Their robot has interfaced with humans and performed tasks based on high-level commands involving robots "understanding" of the meaning of place names for path planning. However, object recognition is black-boxed here. In (Persson, et al., 2007) a system is developed to map an outdoor area, generating a semantic map with buildings and non-buildings labelled. In (Nüchter, et al., 2008), a more general system is presented with a robot equipped by a 3D laser scanner evolving in an indoor environment and constructing a 3D semantic map. The processing is based on Prolog clauses enveloping pre-designed prior knowledge about the environment enabling the robot to reason about the environment. In (Ekvall, et al., 2006), object recognition is performed by a robot equipped by a laser range finder and a camera. A semantic structure is extracted from the environment and integrated to robots map. Another semantic mapping technique is shown in (Meger, et al., 2008) including an attention system.

3 IMAGE SEGMENTATION AND SCENE INTERPRETATION

Section 2 showed the pertinence of semantic SLAM for state of the art robotic mapping. It is this field, on which we are focusing our research. Our motivation comes from the natural ability of human beings to navigate seamlessly in complex environments. To describe a place, we use often very fuzzy language expressions and approximation (see (Vasudevan, et al., 2007)) in contrast to current SLAM algorithms. An interesting point is that people are able to infer distance of an object using its apparent size and their experience of object's true size. Recognition of objects and understanding their nature is an integral part of "human SLAM". We believe that application of semantics and human inspired scene description could bring a considerable benefit in development of robust SLAM applications for autonomous robotics.

For scene interpretation, the image has to be segmented first. Although many image segmentation algorithms exist (see (Lucchese, et al., 2001) for a reference), not all are suitable for mobile robotics due to need of real-time processing. We implement a fast algorithm that breaks the input image into parts containing similar colors with less attention to the brightness. We have chosen the YCbCr color model with Y channel dedicated to the luminance component of the image and other two channels Cb and Cr containing respectively the blue and the red chrominance component. Unlike RGB, the YCbCr model separates the luminance and the color into different channels making it more practical for our purposes.

Our algorithm works in a coarse-to-fine manner. First, the contrast is stretched and median filter is applied to the Cr and Cb components. Then the first available pixel not belonging to a detected component is chosen as a seed point. Eq. 1 captures how a seed point is used to extract the segment of interest (S). P stands for all the pixels in the image, whereas p is the actually examined pixel. Predicate C is true if its arguments (p, p_s) are in four-connectivity. I stands for the pixel's intensity. Seed pixel is denoted by p_s. A pixel of the image belongs to the segment S if the difference of intensities of the current and the seed pixel is smaller than a threshold and there exists a four-connectivity between it and the seed pixel

$$\forall p \in P;\ C(p, p_s)\ \&\ |I(p) - I(p_s)| < \varepsilon \rightarrow p \in S\ . \qquad (1)$$

Using this on both chroma sub-images we obtain segments denoted as S_{Cr} and S_{Cb}. A new segment S is then obtained following Eq. 2 as the intersection of segments found on both chroma sub-images without pixels already belonging to an existing segment

$$S = S_{Cr} \cap S_{Cb} - S_{all}. \qquad (2)$$

At the end of the scan, a provisory map of detected segments is available, but the image is often oversegmented. In the second step, all the segments are sorted by their area and beginning with the largest one the segmentation is run again. This time the seed point is determined as the pixel from the skeleton whose distance to its closest contour pixel is maximal. By this step, similar segments from the

previous step are merged. The ultimate step is construction of a luminance histogram of each segment. If multiple significant clusters are found in the histogram, the segment is broken-up accordingly to separate them.

Now, the segments are labeled with linguistic terms describing their adjacency to each other horizontal and vertical position and span on the image. The average color, its variance and the compactness (Q) of the segment is computed following Eq. 3, where n denotes the area of the segment and o the number of contour pixels.

$$Q = 4\pi n / o^2. \quad (3)$$

These features are used in a set of linguistic rules - the prior knowledge about the world. The aim is to determine the nature of segments and their appurtenance to an object of the perceived environment. E.g. a compact segment found in mid-height level surrounded by the wall is considered as a "window", small compact segments adjacent to the floor are denoted a "box", wide span grayish segment adjacent to the ceiling is labeled a "wall" etc.

4 NAO, THE HUMANOID ROBOT

The robotic platform we use is described in this section. It is based on Nao, a humanoid robot manufactured by Aldebaran Robotics[1]. The robot is about 58cm high with weight slightly exceeding 4kg with 25 DOF. Among others sensors it is equipped with two non-stereo 640x480px CMOS cameras. For simulations, a virtual version of Nao is available for the Webots simulation program developed by Cyberbotics[2]. For development purposes, we have chosen URBI language created by Gostai[3] and aimed specially to robotics. It allows fast development of complex behaviours for robots and provides a simple way of managing parallel processes. LibURBI connectors allow user to develop own objects using so called UObject architecture and to plug them into the language. These objects can be developed in C++, Matlab or Java code. For the demo simulation presented in the next section, we used the simulated robot mentioned above and we are going to use the real one in our further research.

The task itself may be not perceived as being strictly specific for humanoid robots. However, the motivation to use humanoid robots comes from the fact, that they are specially designed with the aim to interact with humans and to act in a human-made environment. The concepts we are exploiting here come from human approach to navigation and orientation in the space, thus embedding such human inspired semantic SLAM capabilities onto a humanoid robotic platform seems pertinent to us.

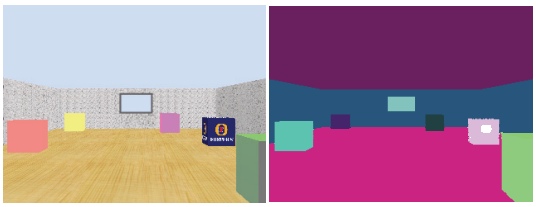

Figure 1: A view of the robot's random walking sequence. The left image is the original one. The right one shows segments detected during the segmentation phase.

5 RESULTS

As a demonstration of some of the mentioned principles, we present a simulation using Webots, where a virtual Nao is walking through a room with objects (cubes) of different colors inside it. The YCrCb image, acquired by Nao's front camera is segmented using our fast segmentation algorithm described in the precedent section (see Fig. 1). The processing speed is several tens of ms for a 320x240 frame on a 2GHz CPU Intel C2D.

After having the image segmented, all segments are labeled and interpreted by a set of prior knowledge rules. Segments can be even merged using these rules to cope with partial occlusions. The "semantic" information is used to approximate the actual distance of objects. Having an object of type "window", its typical size is looked up in the memory (at this stage, the dimensions are known a-priori as the actual learning of object sizes is supposed to be addressed in the future work). The size information is used along with the apparent size of the object to compute its approximate distance (see Fig. 2). This is described by Eq. 4 (simplified for horizontal size only). The distance d to an object is the product of estimated real width w_{real} of the object and tangent of its width in pixels w_{px} on the image multiplied by fraction of the horizontal field of view φ and the width w_{img} of the image in pixels

$$d = w_{real} * \tan (w_{px} * \varphi / w_{img}) . \quad (4)$$

[1] http://www.aldebaran-robotics.com
[2] http://www.cyberbotics.com/
[3] http://www.gostai.com/

The aim of this calculation is absolutely not to infer the exact distance of an object, but rather to determine whether it is “far” or “near” in the context of the world. This can help in further process of creation of the map of the location. Resigning to precise metric position of every object in the mapped world and replacing it only by rough metric and human expressions like “near to” or “beside of” is believed to enable us to create faster and more robust algorithms for SLAM. Using “object landmarks” to navigate in an environment is certainly more meaningful that using e.g. simple points as in case of classical SLAM.

Precise metric information of course has still its role here, but only in some specific cases like close obstacle avoidance or disclosure to grasp an object and notably when the robot is learning typical sizes of objects to enable inference of their distance when they are seen again.

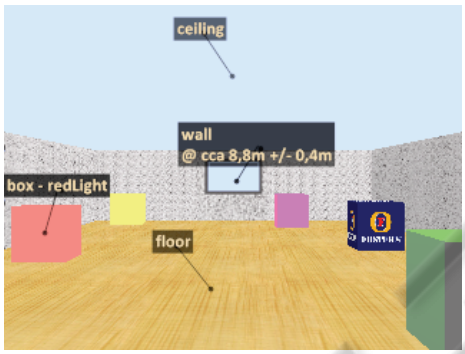

Figure 2: The same view as in case of Fig. 1 after the interpretation phase. Some of the detected objects are labeled. The opposing wall is labeled also with its approximate distance with respect to the robot.

6 CONCLUSIONS

State of the art techniques have been discussed in this paper. In spite of a great advance in past years, a generally usable SLAM solution is still missing. We identify the pertinence of semantic SLAM for the future of mobile robotics and we present our initial research on this field inspired by the human way of navigation and place description. We show a concept of a prospective semantic SLAM algorithm driven by object recognition and the use of human spatial concepts.

For description of a scene by semantic means, a fast and efficient algorithm for image segmentation is an important starting point. A part of our future work will be dedicated to further development of such an algorithm. Another part of our future work will be focused on development of algorithms of semantic SLAM we outlined in this paper. They will be consequently implanted and verified in an indoor environment on the real Nao robot.

REFERENCES

Durrant-Whyte, H., et al. Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. Robotics and Automation Magazine. 2006a, Vols. 13, No 2, pp. 99-110.

Simultaneous Localisation and Mapping (SLAM): Part II State of the Art. Robotics and Automation Magazine. 2006b, Vols. 13, No 3, pp. 108-117.

Ekvall, S., Jensfelt, P. and Kragic, D. Integrating Active Mobile Robot Object Recognition and SLAM in Natural Environments. International Conference on Intelligent Robots and Systems, 2006 IEEE/RSJ. Beijing : IEEE, 2006, pp. 5792-5797.

Galindo, C., et al. Multi-Hierarchical Semantic Maps for Mobile Robotics. International Conference on Intelligent Robots and Systems (IROS 2005). Edmonton : IEEE, 2005, pp. 2278- 2283.

Hahnel, D., et al. Map Building with Mobile Robots in Dynamic Environments. Proceedings of the IEEE International Conference on Robotics and Automation. Taipei : IEEE, 2003, Vol. 2, pp. 1557-1563.

Lucchese, L. and Mitra, S. K. Color image segmentation: A state-of-the-art survey. Proc. Indian Nat. Sci. Acad. (INSA-A). 2001, Vols. 67-A, pp. 207-221.

Meger, D., et al. Curious George: An attentive semantic robot. Robotics and Autonomous Systems. Amsterdam : North-Holland Publishing Co., 2008, Vol. 56, pp. 503-511.

Muhammad, N., Fofi, D. and Ainouz, S. Current state of the art of vision based SLAM. Image Processing: Machine Vision Applications II, Proceedings of the SPIE. 2009, Vol. 7251, pp. 72510F-72510F.

Nüchter, A. and Hertzberg, J. Towards semantic maps for mobile robots. Robotics and Autonomous Systems. Amsterdam : North-Holland Publishing Co., 2008, Vol. 56, pp. 915-926.

Persson, M., et al. Probabilistic Semantic Mapping with a Virtual Sensor for Building/Nature detection. International Symposium on Computational Intelligence in Robotics and Automation. Jacksonville : IEEE, 2007, pp. 236-242.

Thrun, S. and Leonard, J. J. Simultaneous Localization and Mapping. [ed.] B. Siciliano and O. Khatib. Springer Handbook of Robotics. Berlin Heidelberg : Springer-Verlag, 2008, 37.

Vasudevan, S., et al. Cognitive maps for mobile robots-an object based approach. Robotics and Autonomous Systems. Amsterdam : North-Holland Publishing Co., 2007, Vol. 55, pp. 359-371.