Semantic SLAM Model for Autonomous Mobile Robots using Content Based Image Retrieval Techniques: A Performance Analysis

Choon Ling Tan, Simon Egerton¹ and Velappa Ganapathy²

¹School of Information Technology, Monash University, Jalan Lagoon Selatan, Bandar Sunway, 46150, Selangor Darul Ehsan, Malaysia
cltan4@student.monash.edu, simon.egerton@infotech.monash.edu.my
http://www.monash.edu.my

²Department of Electrical Engineering, University of Malaya, 50603, Kuala Lumpur, Malaysia
velappa.ganapathy@um.edu.my
http://www.um.edu.my

Abstract. In the image retrieval domain, Content-Based Image Retrieval (CBIR) methods have been proven to be a fast, efficient and reliable technique of indexing a large number of images without necessarily understanding the contents of the images. These methods are therefore; quite useful in the domain of Simultaneous Localization and Mapping (SLAM) as it handles a large amount of data in real-time. It also provides a convenient initial approach to conducting semantic-based SLAM as object recognition methods applied to SLAM have not been proved to be effective either. Therefore, we have developed a model to conduct SLAM on an autonomous mobile robot equipped with vision sensors. In this paper, we give a brief overview of the model and describe some of our initial experimental results.

Keywords: Semantic SLAM, CBIR, Tamura Textures, image patches, semantic signatures.

1 Introduction

Simultaneous Localization and Mapping (SLAM) is a problem within the field of robotics where a robot or autonomous vehicle – equipped with one or more low-level sensors – is given the tasks of (a) mapping, where construction of a map is performed within an unknown environment, and (b) localization, where the robot/vehicle constantly maintains an estimate of its current location relative to other “visible” objects.

The problem of SLAM originates from [12] which brings up the issue of representing spatial information during the application of robotics. Over the years, traditional methods of conducting SLAM have been developed by researchers such as Thrun [6], [14], [16], [15] Davison [1], [11], and Little [3], [4]. While such methods of SLAM have different forms of implementation when compared against each other, they all have the common feature of tracking topological information on a geometric level.

Extraction of low-level geometric information (such as lines and shapes) is not the only viable form of performing SLAM. Recently, there have been several research efforts that have focused upon conducting semantic SLAM on a semantic level [2], [5], [8] where high-level information is inferred from regular topological maps as suggested in [2]. A model based on this form of SLAM would allow a greater amount of information pertaining to the environment (i.e. such as the navigability or the nature of certain areas) to be recorded, in addition to representing the presence of obstacles. To serve as an example, [9] demonstrates the possibility of certain occupied cells within an occupancy grid that are labeled to indicate the presence of buildings. Also, [10] and [5] provide instances where certain regions of the map are classified according to different high-level objects that have been tracked.

However, research into the field of semantic mapping is still considered to be new. A relatively small number of papers have been published regarding the subject and the majority of them are very recent. It is therefore in the interest of contributing towards this field that we propose conducting semantic SLAM though content-based image retrieval (CBIR). The implementation of CBIR in order to achieve a semantic form of SLAM is also new in itself and has the potential to provide certain advantages over traditional feature-based methods. For instance, the semantic information has the potential to be queried by a human using natural language. There also exists the possibility that such semantic methods have tenable biological parallels.

2 Semantic SLAM Model

As a detailed description of our semantic SLAM model can be found in our earlier work [13], only a brief summary will be provided here. The model that has been developed processes a video stream as illustrated in Fig. 1a., where the processing pipeline comprises of a series of stages: segmentation and signature extraction, categorization and clustering,
and extraction of semantic information for processing into locations and area. This pipeline has been realized using CBIR techniques [13] for feature extraction and signature descriptions of individual image frames. These are then processed into semantic descriptors and these sequences are submitted to a fuzzy inference system which produces a similarity score relative to a specifically memorized semantic location descriptor related to one of the image frames. Fig. 1b. illustrates our implementation and Fig. 2. shows a typical semantic location descriptor generated from this process.

Fig. 1. The various stages within our semantic SLAM model showing (a) the complete SLAM pipeline, and (b) a possible implementation from which the results are reported below.

Volume 12, No. 4
Australian Journal of Intelligent Information Processing Systems
2 Experiments

We have implemented our model in MATLAB, using a standard webcam with a resolution of 640 * 480 pixels, where the images are segmented in a similar manner to that in [7]. However, in order to obtain a meaningful amount of semantic data, each image patch consists of 96 * 80 pixels, resulting in 40 separate image patches per image. The size of image patches is determined through the use of the bestblk command within the MATLAB environment, which specifies the optimal patch size for processing.

In this paper, we reproduce the results of 3 experiments that were previously conducted. The first two experiments involve a camera moving between two areas with different semantic contexts, while the third experiment involves a camera moving on a lateral plane within the same semantic area.

The main purposes of these experiments are (a) to determine if the proposed calculation of the similarity score is capable of differentiating the semantic context between 2 areas, and (b) to determine if the similarity score related to a reference image is consistently bounded within an acceptable value range when images are restricted to the same semantic context. As the first sequence has no previous image to refer to, we set this particular image to be the reference image. However, once any future image generates a similarity score of 5 or less, then that image is considered to be the new reference image. The highest and lowest score an image can possibly generate is 8.31 and 1.69, respectively.

The first experiment consists of 13 images as shown in Fig. 6, where the first 9 images are taken in a corridor and the remaining 4 images are those of a room. Fig. 7 shows the 8 images involved in the second experiment, with 4 corridor images and 4 room images. Images for the third experiment involving lateral plane movement are shown in Fig 8., which consists of 6 images in total.
After conducted multiple runs based on different combinations of weight values for each rule set, we determined that
the values that generate the most promising results were 1.0 for the factors of size and relationship, and 0.0 for the
cluster position. The results are displayed in Fig. 9, which is shown below.

It should be noted that the similarity score in Experiment 1 experienced a sharp dive in value between Image 8
(where the camera is immediately in front of an open doorway) and 9 (where the camera has rotated approximately 90
degrees to face the room), and then increases in value as the camera traverses into the room. It should also be of note
that a similarly large variance is observable between Image 4 and 5 for Experiment 2, though in this instance, it is of an
upward, not downward trend. And in Experiment 3, the similarity scores are consistently above an approximate value of
6.25, which is still considered a relatively high value.

Therefore, future experiments will be conducted to determine if the phenomenon of large value changes over the
course of 2 consecutive images is repeatable for a significant amount of image sets.
3 Conclusion

In this paper, we have presented three experiments which run on separate image sets and their related results based on a proposed semantic based SLAM model with the best possible test parameters. The results show the similarity score values of each image when compared to a reference image, which changes whenever a similarity score yields a value below a certain threshold.

These results raise the possibility that a semantic form of SLAM is indeed possible, provided that certain future modifications are made in order to further improve accuracy. Therefore, further experimentation is deemed necessary to determine if a large variance in value at a particular image moment is able to be observed consistently. The threshold score to determine a difference in semantic context should also be re-evaluated, and modified if deemed necessary. Therefore, we see our model as a potential contribution towards semantic, and especially ontologically based methods that play a key role in future developments of both SLAM and generalized mapping models.

References


