KHARITEH: Toward Simultaneous Localization and Mapping Through Deep Learning

Ashkan Aleali and Saman Biookaghazadeh

Abstract—We aim to present a method to boost robot localization and mapping by using deep learning algorithms. In robotics, the particle filter-based SLAM (Simultaneous Localization and Mapping) algorithm has many applications. SLAM consists of multiple parts; Landmark extraction, data association, state estimation, state update and landmark update. There are many ways to solve each of the smaller parts. Traditinally, laser beams (for indoor environment) and sonar (for marine environment) have been utilized to construct a real time map. A promising approach for solving the SLAM problem, is using vision techniques, which provides higher accuracy compared to laser and sonar and also can provide much more information compared to classical approaches. In this work, we are going to explore applicability of vision approaches using deep learning algorithms on SLAM problem.

1 Introduction

In robotic mapping, simultaneous localization and mapping (SLAM) [8] is the computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it.

SLAM systems consists of important characteristics such as: (1) detecting features, which is able to find landmarks in an image, (2) matching landmarks in consecutive frames, (3) constructing a map of surrounding, given landmarks, and (4) memorizing previously visited areas in the map, aka the loop closure problem.

We aim to present a method to boost robotlocalization and mapping by using deep learning algorithms. In robotics, the particle filter-based SLAM (Simultaneous Localization and Mapping) algorithm has many applications. SLAM consists of multiple parts; Landmark extraction, data association, state estimation, state update and landmark update. There are many ways to solve each of the smaller parts. Traditionally, laser beams (mostly for indoor environment) and sonar (for marine environment) have been utilized to construct a real time map.

Another option is to use vision. We believe using computer vision techniques will provide more informative results and also is more intuitive as it resembles the way humans look at the world. Traditionally, using computer vision techniques were not feasible due to intensive computational overhead and inaccuracy in dark environments. Recent advances in hardware, computational power, and object detection techniques have shown promising results in this field . For instance, in [11], the authors used stereo and triclops cameras to come up with a vision based solution using SIFT feature detection [7].

2 Previous Work

Recently, there has been interest in solving the loop closure problem using the deep learning techniques. This is mainly due to its similarity to the well-studied image classification problem. Also, many image description techniques exist for visual loop closure detection, and they mainly use traditional hand-crafted features.

Cummins et al. in [4] use Bag-of-Visual-Words (BoVW), initially introduced in [12], to solve the loop closure problem. BoVW transforms the image to a histogram of visual words. The visual words are vector-quantized versions of the local keypoint descriptors such as SIFT [7] and SURF [3]. The authors of [4] use BoVW mainly to compute image similarity. So far, this has been the most popular technique in visual loop closure detection. Other image descriptors, like Fisher Vector (FV) [9], [10] has also been used successfully. FV uses a Gaussian mixture model to build a word dictionary where the means of the Gaussian components are cluster centers. Vector of locally aggregated descriptors (VLAD) [2], [6] has also been used. VLAD is specifically useful when the memory and computational power is limited. VLAD, compared with FV, only uses the means, and hence is a simplification of FV.

Hou et al. in [5] use feature descriptors based on a convolutional neural network (CNN) to solve the loop closure problem. They use the output of each layer as a descriptor and construct a vector from the image after feeding the image to a pre-trained CNN.

Zhou et al. in [13] demonstrate the importance of availability of a specific dataset, consists of different dense scenes. This will help scene recognition area to deal with better and more useful datasets. They also represent methods to compare the places databases, with other databases such as ImageNet. Using CNN, they are able to learn deep features for scene recognition tasks. Authors make comparison between ImageNet and Places dataset, in order to prove the diversity and generalization of the places dataset compared to ImageNet, while training scene recognition deep networks. Arandjelovic et al. in [1] combines CNN and VLAD techniques in order to represent a better way for place recognition application. In general, it has a CNN component which is designed to extract D-dimensional features of the image, on different spatial locations. Then this output will be used in a VLAD layer.

3 PROJECT

We plan to implement the discussed approaches and investigate the possiblity of implementing other parts of the SLAM, e.g. the landmark detection, using deep learning techniques since all the work done so far is focused on solving the loop closure problem. We will also try to improve on the solutions. We will also try to label objects captured by the camera as a by-product of the algorithm. We are also thinking to test our implementation on a Turtlebot.

3.0.1 Timeline

We plan to collect sample data using RGBd camera (kinect), through mid-march. We plan to implement the CNN through end of march and finish the experiments by then. We will compare the perfomance of our algorithm in term of accurcy and computation time.

Finally, Ashkan Aleali is currently working on a similar project in the course "Perception in Robotics". The project is to implement and test a SLAM algorithm on a turtlebot.

REFERENCES

[1] ARANDJELOVIC, R., GRONAT, P., TORII, A., PAJDLA, T., AND SIVIC, J. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 5297–5307.

[2] ARANDJELOVIC, R., AND ZISSERMAN, A. All about vlad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 1578–1585.

[3] BAY, H., TUYTELAARS, T., AND VAN GOOL, L. Surf: Speeded up robust features. In *European conference on computer vision* (2006), Springer, pp. 404–417.

[4] CUMMINS, M., AND NEWMAN, P. Highly scalable appearance-only slam-fab-map 2.0. In *Robotics: Science and Systems* (2009), vol. 5, Seattle, USA, p. 17.

[5] HOU, Y., ZHANG, H., AND ZHOU, S. Convolutional neural network-based image representation for visual loop closure detection. In *Information and Automation, 2015 IEEE International Conference on* (2015), IEEE, pp. 2238–2245.

[6] JÉGOU, H., DOUZE, M., SCHMID, C., AND PÉREZ, P. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 3304–3311.

[7] LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110.

[8] MOUNTNEY, P., STOYANOV, D., DAVISON, A., AND YANG, G.-Z. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2006), Springer, pp. 347–354.

[9] PERRONNIN, F., AND DANCE, C. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8.

[10] PERRONNIN, F., SÁNCHEZ, J., AND MENSINK, T. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision* (2010), Springer, pp. 143–156.

[11] SE, S., LOWE, D., AND LITTLE, J. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The international Journal of robotics Research 21*, 8 (2002), 735–758.

[12] SIVIC, J., ZISSERMAN, A., ET AL. Video google: A text retrieval approach to object matching in videos. In *iccv* (2003), vol. 2, pp. 1470–1477.

[13] ZHOU, B., LAPEDRIZA, A., XIAO, J., TORRALBA, A., AND OLIVA, A. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (2014), pp. 487–495.