A New Landmark Detection Approach for Slam Algorithm Applied in Mobile Robot

Xuan-Ha Nguyen[1*], Van-Huy Nguyen[2], Thanh-Tung Ngo[1]

[1]*Hanoi University of Science and Technology, No.1 Dai Co Viet str., Hai Ba Trung dist., Hanoi, Vietnam*

[2]*CMC Institute of Science and Technology, No. 11, Duy Tan, Cau Giay dist., Hanoi, Vietnam*

Abstract

Simultaneous Localization and Mapping is a key technique for mobile robot applications and has received much research effort over the last three decades. A precondition for a robust and life-long landmark-based SLAM algorithm is the stable and reliable landmark detector. However, traditional methods are based on laser-based data which are believed very unstable, especially in dynamic-changing environments. In this work, we introduce a new landmark detection approach using vision-based data. Based on this approach, we exploit a deep neural network for processing images from a stereo camera system installed on mobile robots. Two deep neural network models named YOLOv3 and PSMNet were re-trained and used to perform the landmark detection and landmark localization, respectively. The landmark's information is associated with the landmark data through tracking and filtering algorithm. The obtained results show that our method can detect and localize landmarks with high stability and accuracy, which are validated by laser-based measurement data. This approach has opened a new research direction toward a robust and life-long SLAM algorithm.

Keywords: Deep neural network, mobile robot, object detection, stereo camera, landmark-based slam

1. Introduction

Mobile robots have received much attention in recent years due to their potential application in many fields, for example, logistics, exploration of hazardous environments, self-driving cars and services [1]. Research topics on mobile robots cover a wide range of technologies including, navigation, perception, learning, cooperation, acting, interaction, robot development, planning, and reasoning. For the task of navigation, the robot must know the map of environments and its position simultaneously. This technique is called SLAM – Simultaneous Localization and Mapping.

The SLAM contains the simultaneous estimation of the state of a robot equipped with onboard sensors, and the generation of a model (the map) of the environment that the sensors are perceiving. In simple instances, the robot state is described by its pose (position and orientation), although other quantities may be included in the state, such as robot velocity, sensor biases, and calibration parameters. The map, on the other hand, is a representation of aspects of interest (e.g., position of landmarks, obstacles) describing the environment in which the robot operates [1].

The architecture of a SLAM system includes two main components as shown in Fig. 1 [2]: the front-end and the back-end. The front-end abstracts sensor data into models that are amenable for estimation, while the back-end performs inference on the abstracted data produced by the front-end to estimate the map. The data association module in the front-end includes a short-term data association block and a long-term one. Short-term data association is responsible for associating corresponding features in consecutive sensor measurements; On the other hand, long-term data association (or loop closure) is responsible for associating new measurements to older landmarks. We remark that the back-end usually feeds back information to the front-end, e.g. to support loop closure detection and validation.

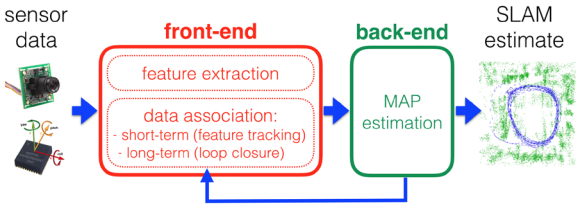

Fig. 1. Architecture of typical SLAM system [1]

*Corresponding author: Tel.: (+84) 946.307.782
Email: ha.nguyenxuan@hust.edu.vn

The SLAM has received noticeable progress over the last 30 years, enabling large-scale real-world applications. A steady transition of this technology to industry has been observed [1]. According to [1], the development of SLAM has overcome three ages. In the *classical age* (1986-2004), researchers focused on the main probabilistic formulations for SLAM, including approaches based on Extended Kalman Filters, Rao-Blackwellised Particle Filters, and maximum likelihood estimation. Subsequently, the *algorithmic-analysis age* (2004-2015) showed the study of basic properties of SLAM, including observability, convergence, and consistency. In this period, developments were based on the suppose that the front-end has been entrusted with establishing correct data association, and thus the back-end is the main focus. Many indoor SLAM applications with quite stable environments were solved with enough accuracy and robustness and thus the algorithmic-analysis trend is considered to mature.

Recently, the robot, environment, performance combinations procure much of fundamental research [1]. Current SLAM algorithms can be easily gone to fail when either the motion of the robot or the environment is too challenging, for example, fast robot dynamics, highly dynamic-changing environments. Also, SLAM algorithms are often unable to deal with strict performance requirements, for example, high rate estimation for fast closed-loop control. These lead us to enter the third era for SLAM, the *robust-perception age, which allows SLAM to be applied* in highly-dynamic-changing environments with improved robustness as well as low computational requirements.

It is inferred that a stable landmark detector is prerequisite for robust data association, especially in dynamic-changing environments. A reliable landmark detector based on object detection would be a promising approach. There have been several investigations dealing with the data association [1-5]. These works have tried to improve the sensor fusion model in order to receive a stable landmark detector. A vision-based method using artificial neural networks is also exploited. However, there is still a lack of a reliable and robust method. In this work, we propose a new approach to detect landmarks for the SLAM algorithm. Based on this method, we use and train a deep-learning-network-based model, called YOLOv3 [6], for the image processing to detect objects in the images captured by cameras of a robot. These objects are considered as landmarks for the SLAM algorithm. Since the vision-based method and deep learning network, which are considered as the state-of-the-art approach, are used, a reliable and robust landmark detector would be obtained.

2. Approach

2.1 Landmark-based SLAM

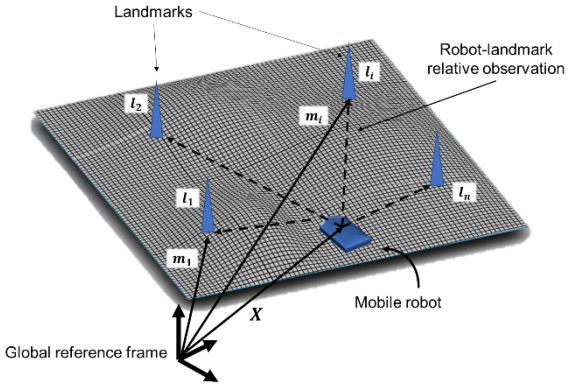

Fig. 2. Concept of landmark-based SLAM

Figure 2 illustrates the basic concept of landmark-based SLAM where a mobile robot and landmarks of the environment are represented in a global reference frame. The goal of SLAM is to know the map of the environment and the robot's position simultaneously. It is stated as follows:

Given:

- The robot's control: $\boldsymbol{U} = \{u_1, u_2, \ldots u_k\}$. These parameters are converted from the odometry system of the robot through encoder and invert kinematic equations;
- Relative observations: $\boldsymbol{Z} = \{z_1, z_2, \ldots z_k\}$. These are relative positions from robot to landmarks of environment which are measured/detected from positioning sensor systems, for example, laser, infrared, ultra-sonic or image-based sensors. In this work, we focus on the landmark detector using a stereo camera, in which we use deep learning models for image processing to detect and measure the position of landmarks.

Wanted: an optimized-mathematic model to estimate

- Map of landmarks: $\boldsymbol{m} = \{m_1, m_2 \ldots, m_n\}$;
- Path/location of robot: $\boldsymbol{X} = \{x_1, x_2, \ldots x_k\}$.

2.2. Data association

The basic idea of our approach is illustrated in a flow diagram in Fig. 3. The flow is described as follows:

- Step 1: from left images of stereo camera equipped with a robot, we run object detection model (re-trained YOLOv3) to detect interested objects and their bounding-box's information

including the height, the width and the coordinates of bounding-box's center;

- Step 2: from the left and right images, we run a deep learning network called PSMNet [8] to calculate the disparity map, which is used to compute the depth map.
- Step 3: from odometry information, intrinsic and extrinsic parameters of the stereo camera, information of detected objects, we calculate the relative position between detected objects and cameras/robot. Position of detected objects are computed;
- Step 4: we perform tracking and filtering algorithms to determine whether the detected objects are the new landmarks and then associate them to the landmark database.

Table 1. Parameters for re-training YOLOv3

Parameters	1st train	2nd train	3rd train
Batch size	64	64	64
Subdivision	32	32	32
Learning rate	0,001	0,001	0,001
Optimizer	Adam	Adam	Adam
Filter	27	27	18
Image size	416x416	1024x512	1024x512
No. of image	2500	2500	2500
Iteration	500k	300k	20k
No. of class	4	4	1

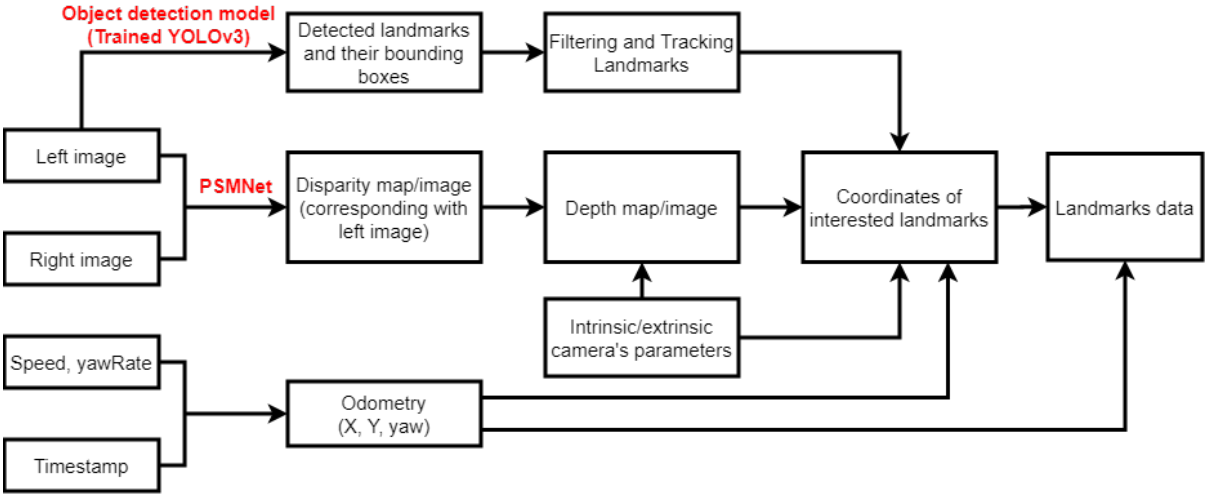

Fig.3. Flow diagram for data association for SLAM algorithm

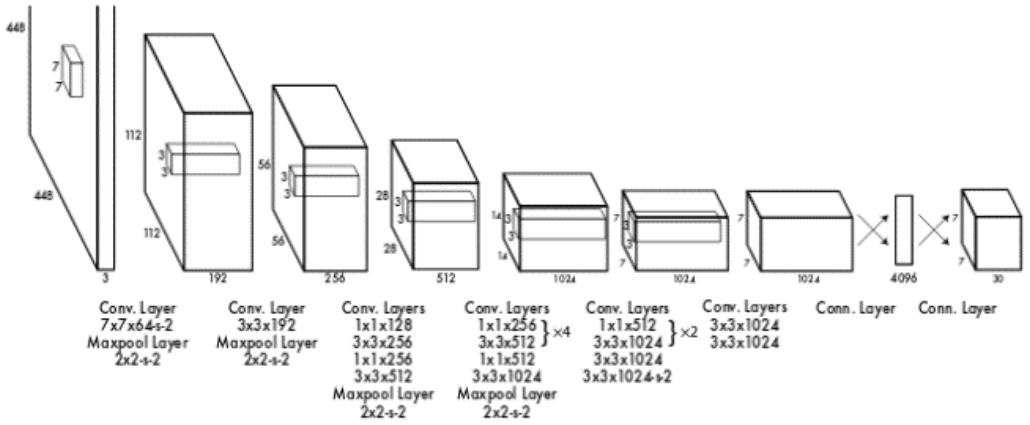

Fig. 4. Architecture of YOLOv3 with Darknet53 backbone [6]

2.3. Landmark detection

You Only Look Once (YOLO) is a state-of-the-art, real-time object detection model. It processes images at very high speed accompanied by a high rate of accuracy of 57,9% at IoU of 0,5 on COCO dataset. As showed in Fig. 4 YOLOv3 is developed basing on Darknet-53 neural network which consists of 53 convolutional layers. In this work, we exploit a pretrained model of YOLOv3 [6] and perform fine-tuning the model to obtain the best-fitted model. Toward applications for autonomous driving, we use the so-called Cityscapes dataset [7] for re-training the pre-trained model. The Cityscapes dataset is an open-source dataset, containing 5000 annotated images with fine annotations captured by the stereo camera installed on a car with a high-accuracy GPS system.

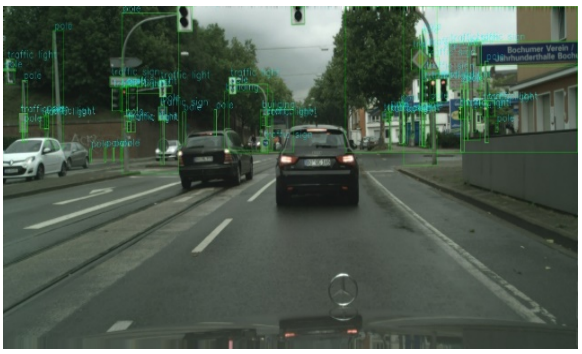

Fig. 5. Image of Cityscapes dataset with bounding boxes of interesting objects [7]

The dataset has 30 classes of objects in traffic infrastructure and road. In this work, we consider only poles, traffic lights, traffic signs as landmarks for the SLAM algorithm. Therefore, we build a program to refine the dataset keeping only interesting objects. Fig.5 shows an example of images having interesting objects accompany with their bounding box. We re-trained the model on a high-performance server with 4 GPUs Pascal Titan X. The configuration of the training parameters is shown in Table 1.

2.4. Landmark localization

After detecting landmarks and their corresponding bounding boxes we compute the position of landmarks relative to the cameras/robot. We can compute the depth from cameras to landmarks using the stereo camera concept with OpenCV. However, the accuracy would be not good enough for practical applications. In this work, we use a deep neural network named PSMNet [8] to compute the disparity of left and right images of the stereo camera system. PSMNet is a pyramid stereo matching network consisting of two main modules including spatial pyramid pooling module and 3D as shown in Fig. 6. From the disparity, we calculate the depth as well as the position of landmarks as shown in Fig. 7. The position of the center of each landmark is computed according to Eq. (1), where: $(x_l\ y_l)$ and (x_r, y_r) are the coordinates of landmark's center; d is the disparity between the left and the right images; f is the focus of the camera; T is the baseline of the stereo camera system.

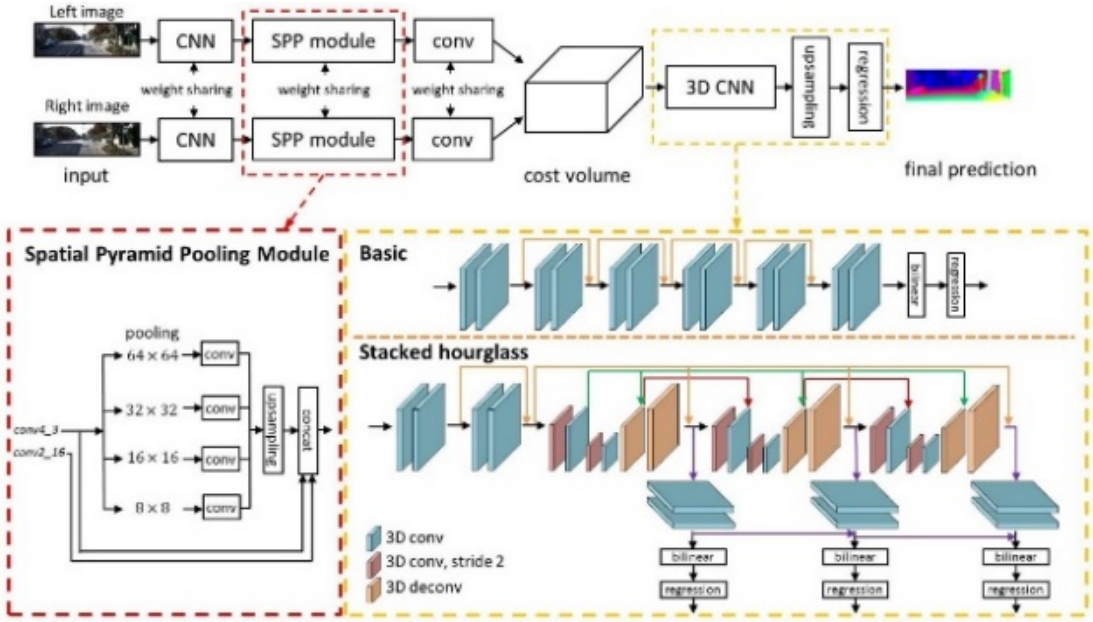

Fig. 6. Architecture of PSMNet model [8]

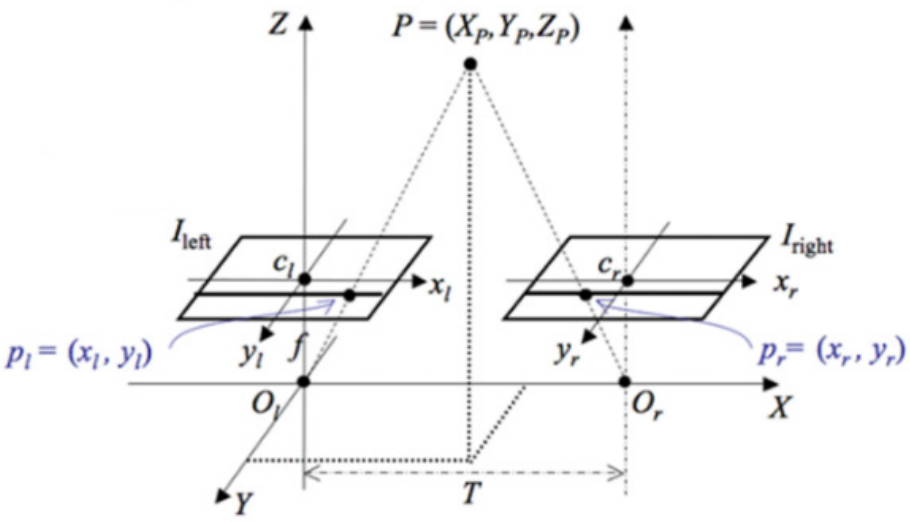

Fig. 7. Drawing of computation of landmark's position

$$Z_P = f\frac{T}{d}; X_P = x_l\frac{T}{d}; Y_P = y_l\frac{T}{d}; d = x_l - x_r \quad (1)$$

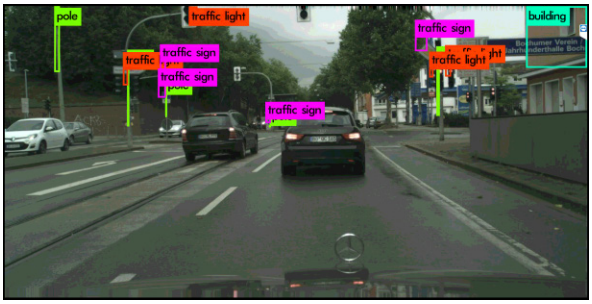

Fig. 8. Result of landmark detection with trained YOLOv3

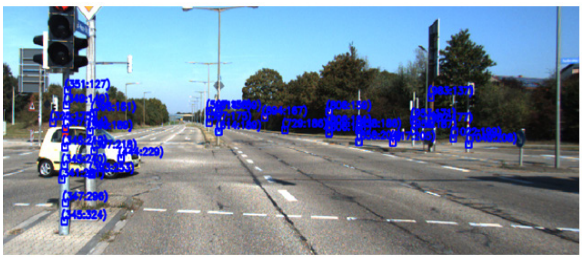

Fig. 9. Computation of pole positions as landmark using YOLOv3 and PSMNet

3. Results and discussion

We performed three trainings for YOLOv3 with parameters listed in Table 1. We choose four types of objects as landmarks including poles, traffic lights, traffic signs, buildings. With the first training, we have the Intersection over Union (IoU) only 50-60% at the iteration of 50k. The reason would be the size of inputted images is too small of 416x416 pixels leading to loss of information when scaling.

In the second training, we increased the size of images to 1024x512 pixels and obtain much better results with IoU rate as high as 80-90% and a precision of 91-99%. As exemplarily shown in Fig. 8, the re-trained model works qualitatively well where almost all poles, traffic lights, traffic signs, and buildings are detected with a high rate of IoU.

For the third train, we considered only pole and configured the filter to 18. The result does not show a better result. Therefore, we use the result of the second training covering more types of objects.

To validate PSMNet we use an open-source dataset named KITTI [9] for the comparison. The KITTI dataset has many images accompanied by laser-based measurement, where the relative position from robot/laser-based sensors to objects in captured images is measured. The poles can be considered as landmarks for practical applications.

Fig. 9 shows exemplarily the computation of pole positions using our proposed method. The results are then compared with laser-based data of KITTI dataset as shown in Table 2.

It is seen that the discrepancy between the depth from our method and the depth from laser-based measurement is small. Especially, when the distance is smaller than 30m the PSMNet-based results match very well with the laser data. This can be explained that, with objects lying within the distance below

30 m, the quality of the pixels is quite good to be inferred. Over this range the results cannot be believed; The error increases to 20%. With a distance further than 30m, the calculation is not exact anymore. In practice we have the tracking and filtering algorithm for landmarks, therefore we do not need to detect and localize landmarks which are very far from the robot. As a result, landmarks in the distance below 30m can be used. Thus, our method can be used efficiently in practice with a high rate of reliability and accuracy. Since we use the vision system, the ambiguity of the environment which usually happens with the laser system can be eliminated.

Table 2. Validation of depth computed by our method and laser-based measurement data

Laser-based measurement data (m)	Depth computed by our method (m)	Discrepancy (%)
6,872	7,12	1,86%
8,863	9,03	1,55%
13,761	14,23	1,74%
22,786	23,143	1,57%
24,033	23,143	-3,70%
30,293	27,98	-7,64%
37,108	33,98	-8,43%
43,748	38,163	-12,77%
59,793	46,531	-22,18%
71,019	56,224	-20,83%

4. Conclusions

In this paper, we have presented a new approach to detect and localize landmarks for the landmark-based SLAM algorithm using deep neural networks for processing images. The YOLOv3 and PSMNet were trained and exploited for the issue of object detection and object localization. Our method shows qualitatively high quality both in the capacity of landmark detection and landmark localization where the average IoU and precision are as high as 80-90% and 91-99%, respectively. Our method can be used for a vision-based SLAM system, where landmarks in a distance of below 30m can be detected and localized at a very high rate of accuracy. The results of this work are good bases for further research in the future. This would be a hot research topic for the community of mobile robot scientists in the era of artificial intelligence with deep neural networks.

In the future, we will continue improving our model and transfer the result to a real system toward the practical application of autonomous driving.

Acknowledgements

This research is funded by Hanoi University of Science and Technology (HUST) under project number TC2018-PC-022.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, J. J. Leonard. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. IEEE Transactions on Robotics, 32(6): 1309 – 1332, 2016.

[2] Y. Latif, C. Cadena, and J. Neira. Robust loop closing over time for pose graph slam. The International Journal of Robotics Research (IJRR), 32(14):1611–1626, 2013.

[3] N. Suenderhauf and P. Protzel. Towards a robust back-end for pose graph SLAM. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pages 1254–1261. IEEE, 2012.

[4] F. Zhong, S. Wang, Z. Zhang, C. Zhou, and Y. Wang. Detect-SLAM: Making Object Detection and SLAM Mutually Beneficial. IEEE Winter Conference on Applications of Computer Vision, 12-15 March 2018.

[5] S. L. Bowman, N. Atanasov, K. Daniilidis and G. J. Pappas. Probabilistic Data Association for Semantic SLAM. 2017 IEEE International Conference on Robotics and Automation (ICRA) Singapore, May 29 - June 3, 2017.

[6] https://pjreddie.com/darknet/

[7] https://www.cityscapes-dataset.com/

[8] Chang, J. R., & Chen, Y. S. (2018). Pyramid Stereo Matching Network. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 5410–5418. https://doi.org/10.1109/CVPR.2018.00567

[9] http://www.cvlibs.net/datasets/kitti/eval_depth.php?benchmark=depth_completion