Real-time 3D Reconstruction through Recurrent Neural Network

Alex Fu
Stanford University
Computer Science Department
alexfu@stanford.edu

Lingjie Kong
Stanford University
Mechanical Engineering Department
ljkong@stanford.edu

Abstract

Inspired by the recent success of neural networks (NN) in 3D reconstruction as well as the application of dense data approach by using ORB-SLAM and RGB-D SLAM, we propose a better solution by combining RNN structure as well as ORB-SLAM. Different from previous RNN 3D reconstruction approach in feeding raw Multiview image as the network input at each time step, we are presenting a solution which use ORB-SLAM to give an initial rough estimation of point location in 3D space from multi-view. Estimated point location will be converted to occupancy of 3D grid. Rough occupancy of 3D grid be past into the network at each time step and eventually map to predicted 3D grid occupancy. The error between the true 3D occupancy and predicted one will be minimized through loss function and stochastic gradient decent. We propose our solution should outperform both pure RNN and SLAM framework because our recurrent network is inspired by a residual network, rather than figure up the mapping, our network will refine the estimated 3D occupancy grid.

1. Introduction

Rapid and reliable 3D reconstruction has become a major innovation in many application such as autonomous driving, 3D printing, and virtual reality. However, fast as well as reliable 3D reconstruction is still under research.

Most 3D objection reconstruction methods work from case to case. Overall, there are two different major approaches. One relies on dense data SLAM to figure the best matching from frame to frame in order to reconstruct the overall 3D. The other one depends on neural networks and let neural network to learn the best matching from input Multiview images to output 3D scene.

Therefore, we are going to introduce our method which combine both SLAM pipeline and neural network.

2. Problem Statement

There are two major 3D reconstruction techniques as mentioned above by using SLAM or neural networks. ORB-SLAM and RGB-D SLAM not only requires a large amount of data, but also might fail due to problematic feature correspondences from local appearance changes or self-occlusions. Meanwhile, CNN or 3D-R2N2 approach using neural networks might require long training time as well as low resolution accuracy. Therefore, there is not a single state-of-art approach which can be applied to all 3D reconstruction case.

Inspired by the idea of boosting which wisely combine several weak algorithms together to eventually outperform all, we are presenting an approach which wisely combine the SLAM approach and NN approach. We hope to present an algorithm which can be applied for most 3D reconstruction problem with a reasonable high accuracy.

We will train our model on PASCAL 3D and ShapeNet dataset.

3. Technical Approach

As mentioned above, we will first apply ORB-SLAM. As mentioned in ORB-SLAM: a Versatile and Accurate Monocular SLAM System [1], an automatic map initialization was used to find the initial correspondence. Then, model is selections by compute a robust heuristic

$$R_H = \frac{S_H}{S_H + S_F}$$

In which S_H and S_F is the score for homography matrix and fundamental matrix. The score is evaluated through find the best correspondence judged by

$$x_c = Hx_r$$
$$x_c^T F x_r = 0$$

Then, we will find the motion and structure from motion recovery. Bundle adjustment will be used again to refine the result. Under the condition that we have already known the camera intrinsic matrix, we should be able to estimate the rough 3D point location from multiple views. Then, 3D point location will be converted into 3D occupancy grid which will be feed into the RNN from each possible. One should be able to achieve multiple 3D occupancy by using different combination of frames.

We will form a long short term memory (LSTM) RNN to prevent vanishing or explode the training weight. At each time step, we will feed the possible calculated 3D occupancy from SLAM. Eventually, the network will output the predicted 3D point. The structure is also inspired

by residual neural network (RNN). Instead of calculate the directly mapping from input to output at each hidden layer, it calculates the possible residual change. It has been approved the resNet outperform most network at several different data set.

The result of our algorithm will be evaluated in two different stages. First, we will compare our predicted 3D grid occupancy as well as the ground truth to see how accurately our algorithm will performance. Second, we will evaluate on cases that single SLAM or NN algorithm that fail to prove that our algorithm can be applied widely at different cases.

4. Intermediate and preliminary result

So far, we spend most of our times studying different approaches to solver the 3D reconstruction from traditional SLAM approach to current neural networks approach.

RGB-D SLAM uses frame to frame tracking to minimize both the photometric and the depth error over pixels through global loop closure [2] [3]. However, this method requires a dense observation from Multiview. This is both computation expensive and require a huge amount of data. ORB-SLAM is another approach without utilizing pixel depth information as above. Instead, it first uses ORB extractor [4] to extract points from different frames. Best correspondence is established between two frames which has the most matched key points to form a spanning tree for all frames. Then, initial 3D points are estimated by using the spanning true frames. Last, 3D points are refined by global loop closure through using Multiview. However, this method still utilize dense data [5]. Prior knowledge with semantic priors is also integrated to monocular SLAM [6]. Known objects segmented from the sense is used to enhance the clarity, accuracy, and completeness of the map built by the dense SLAM system.

Convolutional neural networks (CNN) are widely used in image classification, localization, segmentation, and so on. Convolutional neural networks can also be used to infer a 3D representation of a previously unseen object given a single image of this object [7]. Encoder-decoder network takes RGB image and desired viewpoint as input. It output RGB and depth as the output. However, this method might require a large amount of label data for training. 3D Recurrent Reconstruction Neural Network (3D-R2N2) takes in one or more images from arbitrary viewpoints and outputs a reconstruction of the object in the form of 3D occupancy grid [8]. Even though this network does not require any image annotations or object class labels, it can only output low resolution occupancy grid.

We have already get the open source ORB-SLAM running in our laptop. We also build a vanilla LSTM RNN to test the structure. We are planning to combine both together to get some result early this week.

References

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 2100–2106.

[3] T. Whelan, S. Leutenegger, R. F. Salas-moreno, B. Glocker, and A. J. Davison, "ElasticFusion : Dense SLAM Without A Pose Graph," *Robot. Sci. Syst.*, vol. 2015, no. December, 2015.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.

[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct monocular SLAM," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8690 LNCS, no. PART 2, pp. 834–849.

[6] a Dame, V. a Prisacariu, C. Y. Ren, and I. Reid, "Dense Reconstruction Using 3D Object Shape Priors," *Comput. Vis. Pattern Recognit. (CVPR), 2013 IEEE Conf.*, pp. 1288–1295, 2013.

[7] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Single-view to Multi-view: Reconstructing Unseen Views with a Convolutional Network," *arXiv1511.06702 [cs]*, 2015.

[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction," *arXiv*, vol. 1, pp. 1–17, 2016.