International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-6, August 2020

Review on 3D Mapping and Segmentation

Akash Kuamr Ghanate, Aashish M, Santhosh M Patil, Sowmyarani C N, Ramakanth Kumar P

Abstract: ***The deployment of a robot in a remote environment is a field of research that has huge applications. The robotic system must have the capability of sensing its surroundings and being aware of what it is around. We concluded two key tasks for this purpose, which are 3D mapping and segmentation. This paper shows a comprehensive review of the different 3D mapping and segmentation methods. Mapping techniques include those using RGB images, RGBD images and LIDAR. Segmentation techniques include PointNet, PointNet++, 3D semantic and instance segmentation and joint instance segmentation. We also describe two end-to-end approaches for mapping and segmentation. These methods are reviewed elaborately, comparisons are drawn between them, challenges are presented and future directions in addressing these challenges are pointed out.***

Keywords: 3D Mapping, JSNet, Segmentation, SLAM, Sfm, PointNet

I. INTRODUCTION

The robot in a remote environment has to be aware of its surroundings. It has to build a detailed 3D map of the environment and perform semantic 3D point cloud computing. We believe that this is important for the robot to get a high-level understanding of the surrounding objects and to make context-aware decisions. This has numerous applications in the field of remote healthcare, disaster relief, personal assistants and infrastructure mapping.

To achieve robot interaction in a remote environment, the robot must be capable of sensing its surroundings, so that it can relay the information to another location. There is increasing interest in adding high-level knowledge to many robotics applications in recent years to make robots more capable, even ready to react to unexpected events. To this end, this paper deals with methods of building a 3D map based on the sensor data and performing semantic segmentation of the acquired point cloud.

Revised Manuscript Received on July 05, 2020.
* Correspondence Author
Akash Kumar Ghanate*, B.E., Department of Computer Science and Engineering, R.V College of Engineering, Bangalore, India. E-mail: akashkg.cs16@rvce.edu.in
Aashish M, B.E., Department of Computer Science and Engineering, R.V College of Engineering, Bangalore, India. E-mail: aashishmukund@gmail.com
Santhosh M Patil, B.E., Department of Computer Science and Engineering, R.V College of Engineering, Bangalore, India. E-mail: santhoshpatil937@gmail.com
Dr. Sowmyarani C N, Associate Professor, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore, India. E-mail: sowmyaranicn@rvce.edu.in
Dr. Ramakanth Kumar P, Professor & HoD, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore, India. E-mail: ramakanthkp@rvce.edu.in

3D mapping refers to creating a 3D environment model that depicts the shape and presence of real-world objects in the form of a point cloud. We deal with 3D mapping methods for 3 kinds of inputs, which are RGB images, RGB-D images and the input from LIDAR.

Methods using RGB mainly involve SfM (Structure from Motion) and SLAM (Simultaneous Localisation and Mapping) and another method uses a semantically guided hierarchical SfM approach for 3D reconstruction. As for RGB-D input, we discuss Kinect Fusion as illustrated in [6] and another method which estimates camera pose directly from the SDF using the information it encodes in each voxel as described in [9]. The method with the LIDAR input constructs a 3D model by processing high-density LIDAR data points. We do not cover techniques of 3D mapping for dynamic environments or those environments which involve non-rigid or deformable objects.

With the growth of Neural Networks, the segmentation and object detection in 2D images has made remarkable progress. This advancement in the identification and segmentation of 2D artefacts has encouraged the extension of research to the 3D environment. Older methods of predicting bounding boxes for 3D objects were performed with a single input RGB-D frame with handcrafted feature architecture and then extended the technique to work on learnt features. Further path requires the use of RGB frame data to increase the accuracy of classification of detected objects. But the proposed model does the combined learning between RGB and geometry for explicit spatial mapping. Frustum PointNet[2] uses an alternative approach, where identification is achieved by a 3D image and then projected back onto 3D, using which the final bounding boxes are optimized. Their SGPN[3] approach is based on PointNet++ variation on semantic segmentation. They propose segmentation of instances as a clustering problem through the implementation of a similarity matrix prediction similar to the concept inspired by panoptic segmentation on a semantically segmented point cloud. Although deep learning has been effectively utilized for RGB images, the feature learning capabilities of 3D point clouds with irregular data structures still pose a lot of challenges. PointNet[11] has recently become one of the first methods to specifically apply neural networks to point clouds. This uses mutual multi-layer perceptron and max-pooling to learn from unordered point sets profound features. PointNet, however, is having trouble catching local features. PointNet++[12] dealt with this downside with a hierarchical neural network. The max-pooling operation is a crucial structure for both PointNet and its extended version PointNet++ to extract features from point cloud. But it only retains the best activation of feature maps on a local or global area, which can cause some useful detailed information to be lost for semantic segmentation tasks. While in 3D-SIS[13],

they specifically map all multi-view RGB inputs with 3D geometry to conclude end-to-end segmentation of the 3D instance together. To the best of our understanding, this is the first paper which reviews 3D mapping and segmentation,

the two key initial tasks for a robot in a remote environment.

The rest of the paper is organised as follows. Chapter 2 summarises the various methods on 3D Mapping and Segmentation, Chapter 3 summaries the comparisons of different methods, their evaluations and the challenges involved. Chapter 4 speaks of the future scope of 3D Mapping, Segmentation and the end to end methods, Chapter 5 is the conclusion for the review conducted and Chapter 6 has all the references for this review paper.

II. METHOD OVERVIEW

A. 3D Mapping Methods

1) Using RGB images

There are several techniques for 3D mapping using RGB images using SfM (Structure from Motion). SfM is a photogrammetric imaging technique for estimating 3D models from 2D image sequences. Traditional methods which use SfM are limited by their computational efficiency. They also have the drawback that the 3D map cannot be constructed in real-time and it is difficult to obtain the real-scale tool. [8] uses a combination of SLAM (Simultaneous Localisation and Mapping) and SfM (Structure from Motion) to eliminate this drawback. The key idea is to use SfM to generate a local photo map with no real-scale followed by SLAM for estimating the 3D locations among the local maps. SLAM generates a map that is globally consistent by calculating the real-scale. This kind of approach allows learning on the fly, online mapping as well. This is illustrated in Fig. 1.

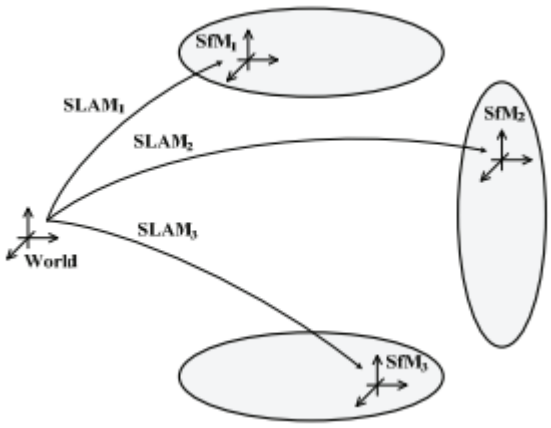

Fig 1.Local map generation using SfM and globally consistent mapping using SLAM in 3D photorealistic mapping with SfM and SLAM method

The SfM consists of four procedures. They are two view triangulation, RANSAC refinement, image stitching and texture mapping. Each of the local maps undergoes these procedures. The local maps generated have their coordinates which do not contain the global information. When the local maps are being built, 3D SLAM procedures are integrated in such a way that the translation and rotation of the robot used is embedded into the two-view triangulation procedure. This is done for re-scaling each local map into a real one.

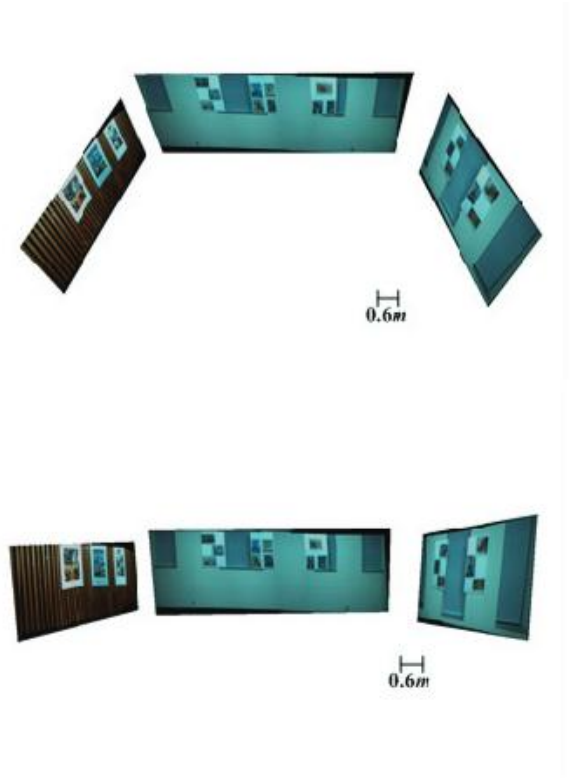

Fig 2. Automatically generated 3D photorealistic map using SfM and SLAM

The most recent method at the time of writing this paper is [7] which uses an approach which is a semantically guided one. This is the hierarchical SfM approach for indoor 3D reconstruction. This method integrates into one single pipeline- clustering of images, segmentation of objects, and reconstruction of the 3D point model. The approach performs SfM in an annotated hierarchical manner with which the cluttered images are classified independently followed by reconstruction along a hierarchical scene tree. This improves the efficiency of computation and also balances the error distribution.

Fig 3. Annotated hierarchical SfM approach workflow

The 3 steps involved in this approach are:

- Extraction of Semantic Information and Classification of Images

  The Fisher vector encoding based on Bag of Visual Words (BOVW) and a popular classification algorithm, Support Vector Machine (SVM) is combined to acknowledge and classify the images to characterise the indoor scenes features for classification more robustly.

The BOVW algorithm groups or clusters features which are similar, as a visible word and then counts the number of times every work occurs within the image. This makes the feature vector needed to improve the semantic level. The result of this stage is a well-categorized image set , which represents the diverse indoor objects.

- Object-Oriented Partial Scene Reconstruction
  With the well-classified images of the scene, the next step reconstructs the object models separately from the classified images by exploiting the SfM algorithm. This stage uses a framework consisting of object recognition, joint semantic annotation and reconstruction.
- Point Cloud Registration and Optimization
  After obtaining the separate object models in the previous step, the separate point cloud models of the acquired indoor objects are merged using the RGPA algorithm into one complete indoor model.

Fig 4. Reconstructed model of a meeting room using hierarchical SfM method

2) Using RGB-D images

[10] made it possible to reconstruct surfaces by integrating groups of aligned range images. The volumetric representation consists of a cumulative weighted signed distance function (SDF). Each image was scan-converted to a distance function and then combined with the data already acquired using an additive scheme. This paved the way for real-time 3D reconstruction using a stream of RGBD images.

Many methods use SDFs. One such method is [6] which uses SDFs as a non-parametric representation to fuse partial depth scans. KinectFusion uses a low-cost depth camera for real-time mapping of arbitrary indoor scenes. The incoming RGBD stream is used to perform real-time dense SLAM which produces a consistent 3D scene in an incremental manner. Simultaneously, the camera's pose is tracked using all of the depth data in each frame.

The 4 components which make up this system are:

- Surface measurement: The raw depth measurements are used to generate a dense vertex map and normal map pyramid. This is a pre-processing stage.
- Surface reconstruction update: The surface measurement is fused into the scene model maintained with a truncated signed distance function (TSDF) representation, given the pose estimated by making use of the depth data from a new frame. This is a global scene fusion process.
- Surface prediction: The loop between localisation and mapping is closed by tracking the live depth frame against the globally fused model. This is done by a rendering technique called raycasting. The SDF is ray-casted into the estimated frame in order to provide a dense surface prediction.
- Sensor pose estimation: This is the localisation part where the 6DOF pose of the camera is tracked using a multi-scale ICP algorithm.

These 4 components are illustrated in Fig. 5

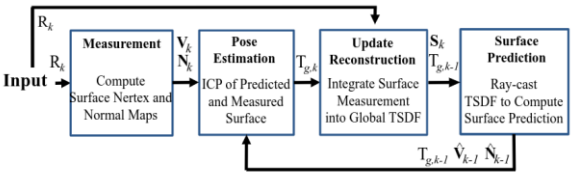

Fig 5. Workflow of KinectFusion

Another method makes use of the fact that the SDF already encodes the distance of each voxel to the surface. As a result, this method illustrated in [9] also uses dense depth images as input and SDF for geometric representation but does not use ICP to achieve real-time performance. The camera pose is optimised directly on the SDF by minimizing the error of depth images on the SDF. This allows the pose optimization to be carried out quickly. The camera poses are iteratively estimated and the RDG-G data is integrated in the voxel grid to get a detailed reconstruction of an indoor environment.

3) Using LIDAR:

LIDAR comprises high frequency which is very precise especially for short-distance measurement. LIDAR has an advantage over Radio Detection and Ranging (RADAR) and Sound Navigation and Ranging (SONAR) in speed, density and accuracy of data. LIDAR can also be used to prepare Digital Elevation Models (DEM) with a precision of 0.1 m.

LIDAR is fixed on a servo motor which enables it to move in all the 3-dimensions, therefore, LIDAR calculates the distance from a stationary point. The servo motor of the LIDAR is programmed to move so that the elevation angle from the other motor will shift from 0-180 degree with a 5 degree difference. After receiving all the values of LIDAR, the code is run to process and change the values to 3D points X, Y and Z [1].

The values received from LIDAR are

R= distance
Θ = angle of elevation
Φ = angle of azimuthal

the code converts the value into 3D points x, y, z using the formula:

$X = R^* \sin(\Theta)^* \cos(\Phi)$
$Y = R^* \sin(\Phi)^* \cos(\Theta)$
$Z = R^* \cos(\Theta)$

Fig 6. Shows the real box picture[1].

3D Mapping is used in the areas of farming, optimisation of wind turbines and rescue. 3D- Mapping is very reliable and cost-effective using this method[1]. LIDAR is installed on a servo motor which makes a 3D map of the front hemisphere of the box. High-density LIDAR data points that plot high-resolution mapping of the 3D hemisphere are processed. The sharpness of the plots can be reduced further by interpolating the data points.

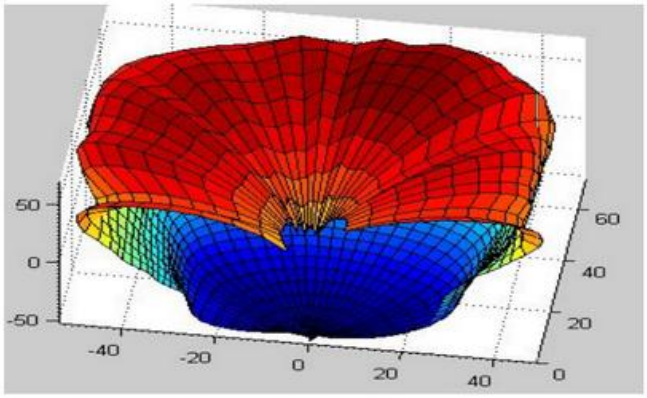

Fig 7. shows the front-top view after filtration [1].

B. 3D Segmentation Methods

1) PointNet

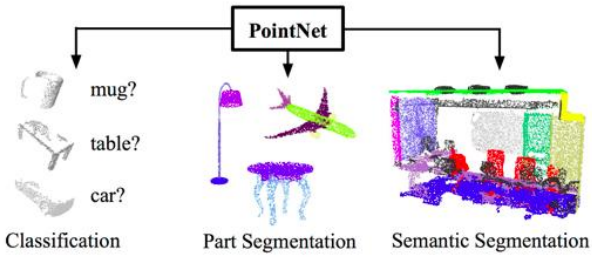

Fig 8. PointNet Applications

PointNet is a deep net architecture suitable for 3D consumption of unordered point sets without rendering or voxelization. It is a cohesive architecture that learns features from global as well as local points, providing a quick, powerful and active approach to various 3D detection tasks. PointNet takes point clouds directly as input and assigns class labels for the whole point cloud or per point segment or section labels for each input point. They train the network to perform 3D instance and semantic segmentation and semantic scene parsing tasks. They provide a detailed empirical and theoretical study of our method's stability and effectiveness. And illustrate the 3D simulated functions of the selected neurons on the net and establish intuitive explanations for their output.

The key approach of the model is the use of a single symmetric function and max pooling, to deal with unordered input collection. Effectively the network learns a series of optimization functions that pick points of the point cloud that are important or insightful and encode the reason for the selection. The network's final completely connected layers aggregate these learned optimal values for the entire shape into the global descriptor or are used to predict per point labels. Our input format is simple to add to stiff or affine transformations, as each point is independently transformed. Further, they added a data-dependent spatial transformer network that tries to canonize the data before it is processed by the PointNet to further boost the performance.

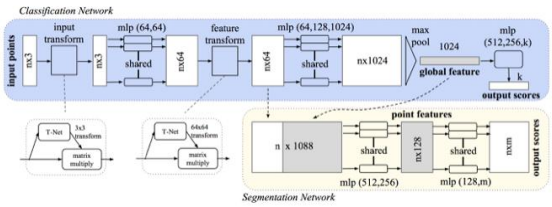

Fig 9. PointNet architecture.

The network that performs classification takes input as n points, applies input and performs feature transformations, and then aggregates point features by max pooling. The performance is the score of classification for groups m. The segmentation network is an extension of the ranking network. It concatenates global and local characteristics per point scores and outputs. Mlp stands for multilayer perceptron, its layer sizes are the numbers in the frame. Batchnorm is employed with ReLU for all layers. Dropout layers in the classification system are used for the final mlp.

2) PointNet++

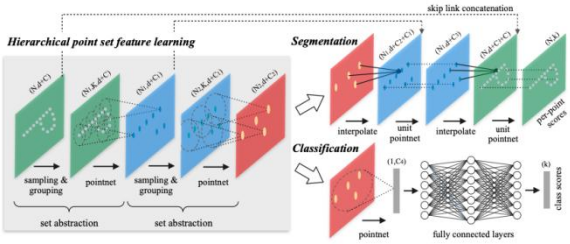

Fig 10. PointNet++ architecture.

pointNet++ is a novel deep learning network which processes in a hierarchical fashion a set of points sampled in a metric space (2D points are used for this example in Euclidean space). PointNet++ overall concept is clear. We initially partition the set of points by the distance metric of the underlying space into overlapping local regions. As with CNNs, we extract local characteristics by capturing of fine geometric structures from small vicinities; these local characteristics are then grouped into larger units and processed to create higher-level characteristics.

This cycle is repeated until we get all point set apps.

3) 3D-Semantic and Instance segmentation(3D-SIS)

The provided architecture is the first attempt on using both 2D features from RGB images and 3D features from point cloud for end-to-end learning to perform 3D segmentation and detecting the object bounding boxes. 3D-SIS is a fully convolutional model, allowing to effectively deduce prediction of huge 3D areas in a single shot. In contrast to other methods, they specifically map all multi-view RGB inputs with 3D features to conclude end-to-end segmentation of the 3D instance together.

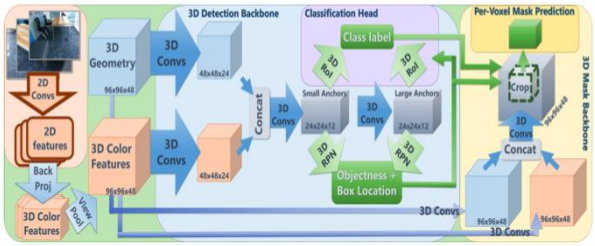

Fig 11. 3D-SIS network architecture.

Model takes the input as the 3D Map and the corresponding RGB frames. Set of 2D convolutions are being applied on the RGB frames to extract the 2D features i.e, they use ENet architecture for 2D semantic segmentation and then they are back-projected into the voxel grid. On the other hand, 3D convolutions operate on the scanned 3D point cloud, where the features are jointly learned from both geometry and RGB data. These generated features are used to identify the class labels and their associated bounding boxes are generated by processing through a 3D- Region Proposal Network and prediction of done class labels is done using a 3D-Region of interest network with a set of pooling layers for each object. Further for each identified object and their corresponding characteristics from both 2D colour and 3D geometry are fed into a per-voxel instance mask prediction network where the training is performed in an end-to-end fashion.

4) Joint Instance and semantic segmentation of 3D point clouds(JSNet)

JSNet consists of a more efficient Point Cloud Feature Fusion(PCFF) module to produce more discriminative features and enhance point prediction accuracy. They propose a novel model module that is joint instance and semantic segmentation(JISS) to facilitate mutual segmentation of instances and semantics. This module further increases the accuracy during the training phase with reasonable GPU memory usage. They have achieved good results on S3DIS dataset[14] along with the major improvements on the segmentation of the 3D instances. Additionally, ShapeNet dataset experiments suggest that JSNet can achieve adequate performance for the task of component segmentation.

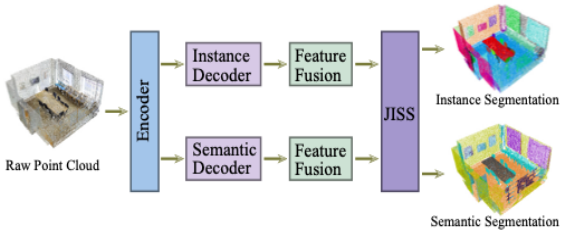

Fig 12. JSNet network

The entire network consists of four main components including a common encoder, two parallel decoders, one point cloud feature fusion module for each decoder, the last element being a joint segmentation module. One aims to extract semantic features for each point for the two parallel divisions, while the other is a segmentation job. For example, we can directly use PointNet++ or PointConv as our backbone network by duplicating a decoder explicitly for the function encoder and two decoders as the two decoders have the same structure. however, for semantic segmentation, the PointNet++ can lose most of the detailed information thanks to max-pooling operation and even the PointConv uses expensive GPU memory during the training process. They are combining the PointNet++ and PointConv in this work to create a more efficient backbone network with reasonable memory costs. The backbone encoder is built by concatenating a PointNet++ set abstraction module and three PointConv encoding layers of apps. Likewise, the

decoders are composed of PointConv's three profound decoding layers followed by a PointNet++ function propagation module.

C. End to End Methods of Mapping and Segmentation

1) 3D Semantic Mapping with Convolutional Neural Networks (CNN)

To analyse an environment thoroughly and perform tasks as simple as fetching an object needs knowledge of both what the object is and where it is located. It would be useful to be able to fetch semantic information from a map by simply offering a database of written tasks about the semantics of a map that was earlier created. The geometric information from a SLAM (simultaneous localisation and Mapping) system ElasticFusion is combined with semantic segmentation using CNN[5].

SLAM system is used to build the 3D map from the corresponding 2D frames. It helps to combine the CNN's predictions into a detailed segmented map as seen in Fig 13.

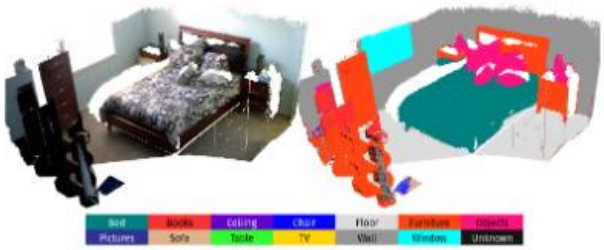

Fig 13. a detailed reconstruction of a video in the left [5] and semantically annotated map on the right.

The map's structure also offers valuable knowledge that can be used to control the final predictions efficiently. The system accuracy is tested on the NYU v2 data and demonstrated that using the information from an unlabeled video, therefore the segmentation efficiency is bosted using only one frame.

This suggests that not only does the SLAM provide an instantly usable semantic 3D map, it also suggests that most of the 2D individual frame semantic segmentation methods may be improved in efficiency if used with SLAM[5]. Through improvising on the dataset to complete room reconstruction, it was discovered that the device was especially well equipped for lengthier scans with a relatively larger range of viewpoints.

SemanticFusion is composed of three steps shown in fig 14:

1.Real-time SLAM system ElasticFusion,
2.Convolutional Neural Network
3.Bayesian update scheme

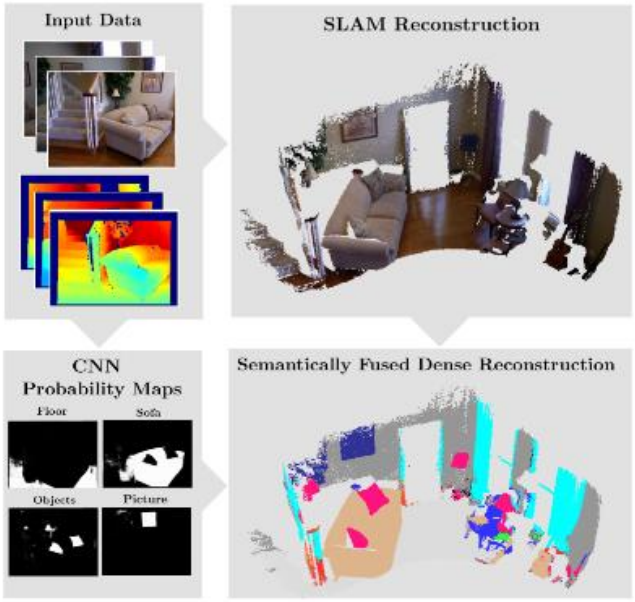

Fig 14. Map is constructed from images using SLAM. By Bayesian updates, these maps are merged into a detailed semantic map[5].

The SLAM method provides a widely compatible map of the merged surface elements, then CNN obtains a 2D image (RGBD) and produces a set of probabilities for each pixel class. Ultimately, a Bayesian update scheme for all surfel measure the class probability distribution and uses SLAM to refine those probabilities based on predictions by CNN's.

2) 3D Reconstruction and Class Segmentation

In this method, the simultaneous segmentation of images and 3D reconstruction is developed as a combined volumetric inference process over multiple labels, using class-specific smoothness assumptions to improve the efficiency of reconstruction. The method uses a parametric representation for the all smoothness priors, that results in a condensed representation for the priors and allows the underlying parameters to be modified simultaneously from training data.

As a volumetric approach operating on a standard polygon mesh grid, this method shares the limitations with many other volumetric methods regarding spatial resolution.

The method proposes to study the likelihoods of appearance and class-specific geometry priors in an initial step for surface orientations of the training data[6]. The priors are used to identify pairwise and individual potentials of segmentation framework, complementing that of a calculated evidence acquired from depth maps. Optimizing the label assignment in its volume, picture-based probabilities, machine stereo depth maps, and priors communicate with each other, resulting in improved detailed reconstruction and labelling.

III. COMPARISONS AND EVALUATIONS

The 3D photorealistic method which uses SfM and SLAM has the main advantage that it can run on the fly, that is, online. The execution time of the method was calculated to be 3.92 sec when it was run on 11 sets of data using an Intel i7 870 CPU. This execution time is just enough for real-time implementation. The main challenge is to speed-up this process to the microsecond level.

The annotated hierarchical SfM method is computationally efficient compared to traditional incremental SfM methods which involve exhaustive pairwise image matching. The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) building is used as the dataset. Three bag-of-words based SVM classification methods were run on this dataset and the results are shown in Table 1.

Table 1: Results of the three classification methods

Encoding Method	Number of Words	Classification Accuracy	Mean Average Retrieval Precision	Classification Time (s)
BOVW	1000	0.942857	98.21%	62.45
BOVW	2000	0.957143	98.69%	220.97
VLAD	25,600	0.957143	99.25%	21.02
FV	20,480	0.985714	99.46%	57.19

The proposed method was also compared to the state of the art VisualSfm (VSFM) method. The results are shown in Table 2.

Table 2. Comparison of hierarchical SfM to VSFM

Dataset/Method	Meeting Room Dataset			Lobby Dataset		
	Error (pixel)	Time (s)	No. of Views Recovered	Error (pixel)	Time (s)	No. of Views Recovered
VSFM (Wu, 2013)	2.641	18735	287	2.293	13987	235
The proposed method	2.454	2025	304	2.040	1526	243

The main challenge of the hierarchical SfM is that it cannot be used for real-time purposes. Also,this method makes it difficult to reclaim partially occluded models. With RGB-D images as input, KinectFusion method was run with different voxel resolutions as the sensor reconstructs in a volume of a 3-metre cube. The time taken is illustrated Fig 15.

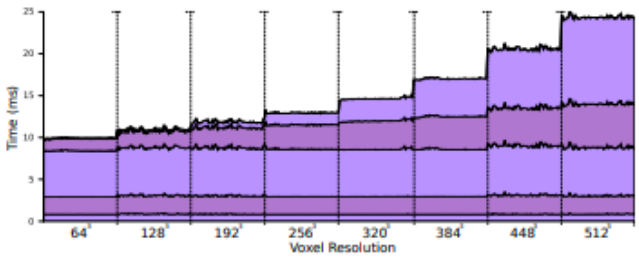

Fig 15. Real-time cumulative timing results of KinectFusion over a range of resolutions

The challenge with KinectFusion is the reconstruction of large scale models such as the interior of a whole building. In such large models, another important challenge is to efficiently perform automatic relocalisation when the tracking has failed.

The 3D reconstruction using signed-distance functions method does not use the ICP algorithm. An evaluation of benchmark data has shown that this method is more accurate and robust than the ICP algorithm used by KinectFusion. It also provides a similar accuracy at a much higher speed when compared to bundle adjustment methods such as RGB-D SLAM.

Table 3. Semantic Segmentation results on ShapeNet dataset

Method	**mIoU(mean intersection over union)**
PointNet[11]	83.7
PointNet++[12]	84.9
JSNet	85.8

The experimental results show that fusing the characteristics of different layers(JSNet) could increase the precision of segmentation due to the richer features after fusion. As for the only instance fusion of semantic segmentation and only segmentation of semantic awareness instances, the results indicate that better instance predictions could allocate more accurate category labels to semantic branches, which could boost semantic efficiency.

Table 4. Segmentation results on ScanNet Dataset

Method	**Avg(Mean average Precision)**
Mask R-CNN[15]	5.8
SGPN[3]	14.3
R-PoinNet	30.6
3D-SIS[13]	36.2

By comparing the above results 3D-SIS outperforms the previous(Mask R-CNN) or current state of art methods(PointNet) on ScanNetV2 3D semantic instance benchmark. With an IoU threshold of 0.25 over 23 groups have been tested by mean average accuracy. Therefore, joint colour-geometry function helps us to achieve more accurate performance in segmentation instances.

IV. FUTURE SCOPE

There is a lot of scope for improvement for 3D mapping methods. The joint solution using SfM and SLAM can focus on genuine real-time implementation using either parallel computing or system-on-chip technique. The method can also be extended to apply to an environment with non-plane geometric objects. The hierarchical SfM method can look at combining geometric and semantic priors to determine the dense point cloud and recover partially occluded models. Extension of the dataset size and employing improved, more robust feature extraction methods can lead to better model quality. KinectFusion can be extended for large scale models by using a sub-mapping framework. For 3D reconstruction using SDFs, colour information can be included for camera tracking and methods with more efficient representation of 3D geometry can be explored. 3D Mapping with LIDAR to be mounted on a drone for military purposes, It can be used in self-driving vehicles to create a detailed map of the surroundings. Global System for Mobile Communications can use the method for the places inaccessible to humans, for example in case of an earthquake.

As for semantic segmentation, the models provided above are just a starting point for obtaining 3D semantic segmentation from 3D point clouds of high quality, which is a common issue for RGB-D reconstructed models. The problem of semantic segmentation in the 3D environment is distant from being solved, and the semantic instance of 3D segmentation is in its infancy as well.There are also specific questions about the representation of the scene to realize 3D CNN models, and how to deal with mixed sparse-dense representations of data. We also look into the enormous possibility for integrating multi-modal characteristics in 3D reconstruction for generative assignments, such as scene completion and texturing. For the end to end methods, the improvement can be achieved with longer trajectories, which will result in better labelling. As with the volumetric approach operating on a standard polygon mesh grid, the system faces the limitations of spatial resolution, adaptive representation of this data may be a potential solution, for a finer segmentation there should be an increase in the number of object categories.

V. CONCLUSIONS

The paper establishes the two key tasks for a robot in a remote environment to sense its surroundings and be aware of its environment, which are mapping in 3D and segmentation. We summarize the important methods for mapping and segmentation separately along with two end-to-end methods which do the joint task of mapping and segmentation. The results of each method are displayed and comparisons are drawn. Each method's evaluation is summarised and challenges involved are mentioned. The future directions to overcome these challenges are also suggested.

REFERENCES

1. M. H. Riaz, S. A. Bukhari, F. Mukhtar, T. Kamal, H. Sarwar and M. U. Tahir, "3d mapping using light detection and ranging," 2017 International Multi-topic Conference (INMIC), Lahore, 2017, pp. 1-4. DOI:10.1109/INMIC.2017.82894680
2. CharlesRQi,WeiLiu,ChenxiaWu,HaoSu,andLeonidasJ Guibas. Frustum pointnets for 3d object detection from rgb-d data. *arXiv preprint arXiv:1711.08488*, 2017
3. Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neu- mann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2569–2578, 2018.
4. J. McCormac, A. Handa, A. Davison and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks," 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 4628-4635.
5. C. Häne, C. Zach, A. Cohen, R. Angst and M. Pollefeys, "Joint 3D Scene Reconstruction and Class Segmentation," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 97-104.
6. R.A.Newcombe et al.,"KinectFusion: Real-time dense surface mapping and tracking," 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, 2011, pp. 127-136.
7. Ding, Y.; Zheng, X.; Zhou, Y.; Xiong, H.; Gong, J. Low-Cost and Efficient Indoor 3DNReconstruction through Annotated Hierarchical Structure-from-Motion. *Remote Sens.* 2019, 11, 58.
8. Choi, H., Jun, C., Li Yuen, S., Cho, H., & Doh, N. L. (2013). Joint Solution for the Online 3D Photorealistic Mapping Using SfM and SLAM. International Journal of Advanced Robotic Systems.
9. Bylow, E, Sturm, J, Kerl, C. (2013) Real-time camera tracking and 3D reconstruction using signed distance functions. In: Robotics: Science and systems conference (RSS).
10. B. Curless and M. Levoy. A volumetric method for building complex models from range images. In ACM Transactions on Graphics (SIGGRAPH), 1996.
11. R. Q. Charles, H. Su, M. Kaichun and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 77-85.
12. Y. Lian, T. Feng and J. Zhou, "A Dense Pointnet++ Architecture for 3D Point Cloud Semantic Segmentation," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 5061-5064.
13. J. Hou, A. Dai and M. Nießner, "3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4416-4425.
14. Zhirong Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1912-1920.
15. K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

AUTHORS PROFILE

Akash Kumar Ghanate, B.E. Computer Science and Engineering, R.V College of Engineering, Bangalore- 560059

Santhosh M Patil, B.E. Computer Science and Engineering, R.V College of Engineering, Bangalore- 560059

Aashish M, B.E. Computer Science and Engineering, R.V College of Engineering, Bangalore- 560059.

Dr. Sowmyarani C N, Associate Professor, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore-560059

Dr.Ramakanth Kumar P, Professor & HoD, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore-560059