SEMANTIC SEGMENTATION OF INDOOR POINT CLOUDS USING CONVOLUTIONAL NEURAL NETWORK

K. Babacan, L. Chen, G. Sohn

Dept. of Earth & Space Science & Engineering, York University Toronto, M3J1P3 Canada –
(babacan, leihan, gsohn)@yorku.ca

KEY WORDS: Indoor Modelling, Semantic Segmentation, Mobile Laser, Point Cloud, Deep Learning, Convolutional Neural Network

ABSTRACT:

As Building Information Modelling (BIM) thrives, geometry becomes no longer sufficient; an ever increasing variety of semantic information is needed to express an indoor model adequately. On the other hand, for the existing buildings, automatically generating semantically enriched BIM from point cloud data is in its infancy. The previous research to enhance the semantic content rely on frameworks in which some specific rules and/or features that are hand coded by specialists. These methods immanently lack generalization and easily break in different circumstances. On this account, a generalized framework is urgently needed to automatically and accurately generate semantic information. Therefore we propose to employ deep learning techniques for the semantic segmentation of point clouds into meaningful parts. More specifically, we build a volumetric data representation in order to efficiently generate the high number of training samples needed to initiate a convolutional neural network architecture. The feedforward propagation is used in such a way to perform the classification in voxel level for achieving semantic segmentation. The method is tested both for a mobile laser scanner point cloud, and a larger scale synthetically generated data. We also demonstrate a case study, in which our method can be effectively used to leverage the extraction of planar surfaces in challenging cluttered indoor environments.

1. INTRODUCTION

Semantic information is increasingly becoming an indispensable ingredient of BIM. Applications such as energy flow monitoring, emergency management, retrofit planning, visualisation (Volk et. al., 2014), crucially depend on the availability of the class information of the entities in the model. For new constructions, this information is essentially input in the design phase. In contrast, only after an existing building is geometrically modelled, semantic enrichment of that model takes place. In other words, the labels are given to extracted surfaces; the unstructured point cloud is not perceived to possess categorical information.

However, the case of holding this semantic information prior to geometric modelling could greatly contribute to the conventional modelling process. For instance, directly acquiring a recognition of which points belong to the category of wall, could bypass the need for calculating the surface normal, and making the assumption that horizontal normal form a good basis to comprise that particular class.

In regard to this interest of employing more meaningful features in modelling, the concept of semantic segmentation has arisen in different research domains, mainly in computer vision, and robotics (Thoma, 2016). As an important notion towards complete scene understanding, semantic segmentation is applied to numerous application such as autonomous driving, augmented reality, and computational photography (Garcia-Garcia et al., 2017).

The research in indoor modelling for semantic segmentation is usually performed on RGB-D sensor depth images for small indoor scenes. On the other hand, the importance of 3D point clouds for devising better performing classifiers has been demonstrated (Koppula et al., 2011). Thereafter the robotics community effectively employs SLAM based techniques to jointly extract localization and more meaningful maps of the indoor environments. Despite considerable advances of drift error reduction by pose estimation graphs, there are still feasibility issues for large scale indoor semantic segmentation (Fuentes-Pacheco et al., 2015).

Keeping in mind the large scale building indoor models, beside the inherent limitations of these commonly employed data acquisition techniques in terms of scale, the generally accompanying methodological framework of probabilistic graphical models such as Conditional Random Fields (CRF) also suffer from a similar problem. In order to mitigate the computational burden encountered in optimization, a necessary clustering like super pixel grouping (Fulkerson et al., 2009), or line extraction (Jung et al., 2016) is drawn on the data for large scale classification applications.

In the last five years we have witnessed the revival of neural networks, as deeper architectures become effectively possible (Krizhevsky et al., 2012), (Szegedy et al., 2015). Especially Convolutional Neural Network (CNN) has been the leading network type of many successful practical applications (Karpathy et al., 2014), (Farfade et al., 2015). In CNN, powerful hierarchical representation is generated through self-learned features in a supervised manner directly from the data. Compared to conventional features engineered by a specialist, these self-learned features provide a powerful geometric discriminator among data categories. Initially practiced on image classification as a whole, a number of other computer vision tasks such as object detection, and recognition also benefit from modified networks and algorithms, among semantic segmentation. However, huge labelled training data necessary to match the

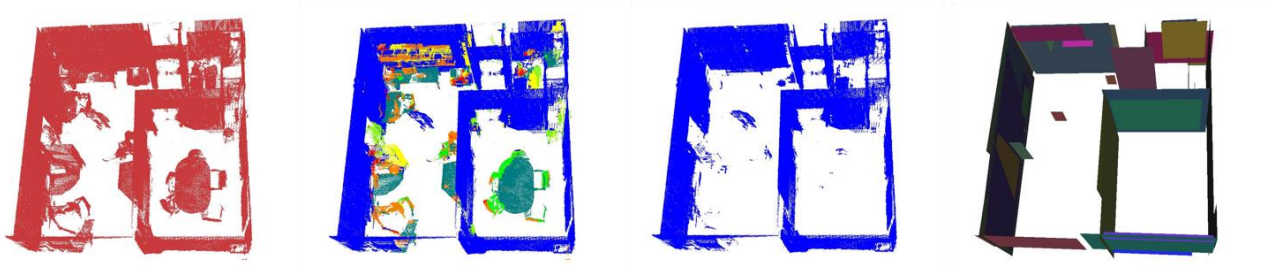

Figure 1. The input and outputs of the method; a) raw point cloud of the cluttered indoor environment; b) direct application of previously trained CNN classifier to the test site; c) selection of the wall points determined by the CNN; d) the result of the planar extraction, applied only to the corresponding wall points

depth of the network, and the number of parameters to be optimized, becomes more deficient with per pixel labelled training data required for semantic segmentation.

In this study, a method to directly classify large scale 3D indoor point clouds by using Convolutional Neural Networks has been developed. Our main contributions can be expressed as follows;

- Large-scale indoor point cloud classification: We have acquired and prepared suitable datasets both real and synthetic, and 3D input–output relationships compatible with a simple, fast CNN architecture and tuned our algorithm to train and run over a high number of points, and a floor size indoor space.
- Effective clutter removal based on semantics: We have tackled the problem of clutter in indoor modelling by reframing it as a simple semantic filtering.
- Demonstration of enhanced planar extraction: Finally we demonstrate how geometry reconstruction can benefit from semantic segmentation with a case study of planar extraction enhancement in particular. (Fig.1)

2. RELATED WORK

The literature related to our research can be divided into five interrelated categories. We start with geometric indoor modelling from point clouds in which semantics follow the geometry extraction separately. Then the important work on semantic segmentation originated in image processing, extending to depth images in small scale indoor scenes are addressed. Consequently, the introduction of Convolutional Neural Networks to indoor scene segmentation, and the advances achieved in image processing are discussed. Besides, the implementation of CNNs on 3D data in general with a focus on point clouds are briefly touched. Finally, in order to provide a background for our case study of planar extraction in indoor modelling, relevant papers are mentioned.

Semantic Indoor Modelling

In indoor modelling from point clouds, it is common practice to first extract planar primitives, and subsequently classify them into horizontal structural elements of ceiling and floor, and vertical walls. Various methods have recently been developed based on projection plane histogram analysis (Okorn et al., 2010), plane sweep (Budroni and Boehm, 2010), surface normal (Sanchez and Zakhor, 2012), stacking (Xiong et al., 2013), and diffusion embedding (Mura et al., 2013). Common to all these methods is the sequential approach to label the structural elements according to previously segmented planar surfaces.

Semantic Segmentation

In contrast to the sequential approach, semantic segmentation is the segmentation of the data as a natural result of the classification procedure of a basic unit, i.e. usually being pixel or superpixel level. An early example can be found in the work of Huang et al. (2002), for land cover classification using Support Vector Machines. In order to provide the general framework, and impose the consistency of the segments CRF has been a standard technique for the last decade. A pioneering work of Silberman and Fergus (2011) applied CRF to achieve dense labelling in small indoor scenes captured by a low-cost depth sensor.

Convolutional Neural Networks on Semantic Segmentation

Recently, convolutional neural networks have become the state of the art for semantic segmentation tasks. For indoor scenes, depth sensors are continued to be employed, and a CNN version of full scene labelling is introduced by Couprie et al. (2013). With the advancement of CNN research, different efficient network architectures are proposed. Among them; Deeplab which combines CNNs with fully connected CRF (Chen et al., 2016), Fully Convolutional Networks (FCN) which employs 1*1 convolutions and some skip connections and upsampling (Long and Darrell, 2015), Deconvolutional Neural Networks (Noh et al., 2015), CRF-Recurrent Neural Networks (Zheng et al., 2015), and SegNet (Badrinarayanan et al., 2015) could be named as significant developments. For more details about these architectures the reader is referred to the review paper by Garcia-Garcia et al. (2017).

Convolutional Neural Networks on 3D Data

After a brief period of pause following the very early attempts of the implementation of convolutional neural networks directly on 3D data, a rapid attention has been shown in a variety of research communities. VoxNet (Maturana and Scherer, 2015) based on volumetric representation as the name implies, is one of the first effective implementation of CNN on object detection, in which the whole of a bounding box is classified as the segment based. On the other hand, Multiview 2D CNNs achieve slightly better results due to their exploitation of pre-trained models on very large image datasets. Recently in a similar approach to ours Huang and You, (2016) classify the urban Lidar points without

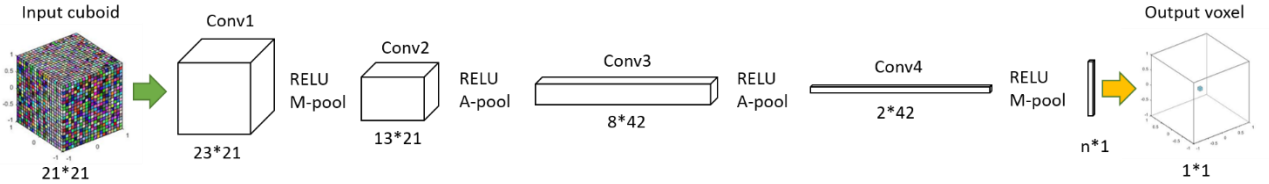

Figure 2. Our Convolutional Neural Network has four main blocs, each having their respective convolution-pooling and activation layers. Input cube is fed into the network and the result of the soft-max layer is assigned to the centre voxel. n is the variable for the number of categories.

any object detection or pre-segmentation. A detailed survey can be found in the survey of Ioannidou et al., (2017).

It is worth here to mention another very recent approach that attempts to break the dilemma of CNNs applicability to 3D data either as volumetric 3D or Multiview 2D, by proposing a new deep learning architectures based on Autoencoders that could directly operate on 3D point cloud (Qi et al., 2017).

Planar Extraction in Indoor Modelling

Primitive extraction is a critical part of many indoor geometry modelling methods for extracting planar surfaces in man-made indoor spaces. Planar surface fitting in laser data is an already extensively studied field in remote sensing community. A comprehensive research by Nurunnabi et al. (2014) compares some of the existing algorithms. A similar comparison for mobile indoor mapping can be found in (Nguyen et al., 2007). For indoor reconstruction implementation, variants of Hough Transform (Okorn et al., 2010), (Oesau et al., 2014) and RANSAC (RAndom SAmple Consensus) (Dumitru et al., 2013), (Ochmann et al., 2014) are employed beside plane sweeping by Budroni and Boehm (2010), or EM (Expectation-Maximization) by Thrun et al. (2004). Sanchez and Zakhor (2012) utilized RANSAC in a region growing method to detect planar primitives. Recently Boulch et al., (2014) incorporated regularized edge, and corner, while Monszpart et al. (2015), has imposed regularity constraints on extracted planes.

3. METHOD

The input to our method is the raw point cloud, and the output is the densely labelled point cloud, being that a label is assigned for each point. In order to be able to employ a fast and a simple CNN architecture, the point cloud is densely voxelized, and an occupancy representation is formed in the first place. For labelling the training data, manual classification of the point cloud is transferred into the voxel domain by means of a majority voting of the corresponding points in that voxel. Subsequently voxels are agglomerated into cubes (Fig.2). To provide the variability of the training data, each cube is created by shifting along every prime direction with the smallest stride size, being a voxel. Once the cubes are generated, they are fed into the CNN architecture which is designed to handle 3D voxel data. The result of the CNN classifier is assigned to the centre voxel, and the classification is carried on to ensure a continuous dense classification of all voxels.

The details of the preparation of the data and the CNN architecture are explained in this section.

3.1 Data Preparation

A deep Convolutional Neural Network classification paradigm is extremely data dependent, therefore data is at utmost importance. A major drawback of CNN for practical applications is the requirement for large amounts of data. There have been continuous research community efforts in computer vision, culminating in large-scale image databases such as ImageNet, presented (Deng et al, 2009) freely to the service of researches. Though not in the same scale, a similar strive provides RGB-D datasets for indoor scene reconstruction purposes. However, when it comes to per-pixel labelled datasets, the options are still scarce. Moreover, large scale indoor modelling for point clouds lack a direct point cloud database that could be utilized for training deep networks. Therefore, there is an obvious need for such a database.

CNN demands a lattice structure as an input while the raw point cloud is unordered which cannot directly be processed in CNN architecture. Therefore, to abide in 3D the raw point cloud should be transferred into a volumetric representation. Depending whether the task is classification or segmentation, a number of alternative data preparation paths that can be taken are summarized in Table 1. Initially the bounding box of raw point cloud is calculated. Now we describe how we divide this bounding box into two scales: voxel and cuboid.

A 3D bounding box is firstly divided into voxels which have a certain pre-defined size. There are different ways to represent a voxel, among which the simplest method is the binary occupancy value which evaluates whether there are points existing within the voxel. In the case of present points, the intensity value of this voxel is set to one, otherwise it is zero. A finer way is to count how many points fall into the voxel and assign this number of density as the voxel's intensity value. There are also some more advanced representations which takes into account probability distributions (Maturana and Scherer, 2015). In our case, considering computational simplicity and information preservation, we select this density representation which is both computational efficient and preserves more information of raw point cloud than just occupancy value.

Table 1

INPUT (CUBOID AND VOXEL)		**OUTPUT (CUBOID OR VOXEL)**
Cuboid Label (~image)	**Voxel Representation (~pixel)**	**Classification / Segmentation**
Has label (from Cen. Vox.)	Binary	Center Voxel (from Cuboid Label) –segm.
Has label (from Maj. Vox.)	Density	Entire Cuboid (from Cuboid Label) – class.
Has no label		Entire Cuboid (from Voxel Label) -

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-4-W4-101-2017 | © Authors 2017. CC BY 4.0 License.

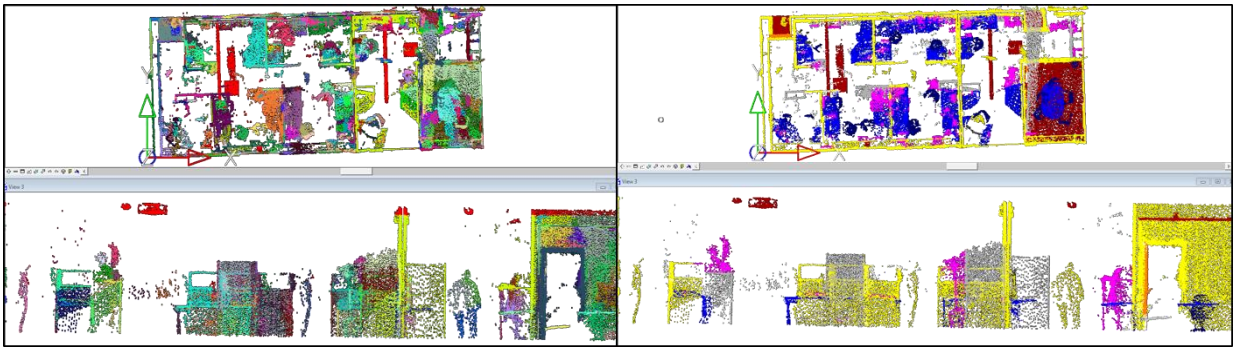

Figure 3. Clustering algorithms like connected component analysis and region growing algorithm are not effective in cluttered environments (left-hand side). Hence a laborious manual labelling for training dataset generation has been directly applied on 3D point cloud (right-hand side).

Next, the generated voxels are encapsulated as cuboids which will be the input of the neural network. Specifically, given a certain voxel, its corresponding cuboid is an aggregation of the voxels in three dimensions in which the centre voxel is the given voxel. In our setting, each cuboid consists of 21*21*21 voxels. For the voxel locating in the boundary of bounding box, we use zero-padding to generate its corresponding cuboid to confirm all cuboids have the same size. The concept of voxel and cuboid is analogous to the relation of pixel and image in 2D image plane, which expresses the raw point cloud in a 3D raster format as an input for CNN architecture.

For training data, we should also assign the label of every voxel. Considering the difficulty to manually assign label to every voxel, we manually assign the label to raw 3D point cloud and then the label of voxels are determined with the label of its points' label using majority voting. For the label of cuboid, we directly assign it as its centre voxel label. The empty voxels will be ignored both in the training and test data.

For test dataset, we also generate the cuboid for each non-empty voxels and assign the classification result to the centre voxel of the cuboid. The points which fall into this voxel will all be assigned the label of this centre voxel. In this setting, we can generate point-based semantic segmentation result, without resorting to a more sophisticated CNN architecture such as FCN or Deconvolutional Neural Networks

3.2 Model

A CNN is essentially a discriminative classifier which models the desired output *y* in this form;

$$y = f(x; \theta, w) = \varphi(x; \theta)^{T} \omega \quad (1)$$

φ is the feature set learned through optimizing the parameters θ. ω maps the feature set to the output. In the case of a deeper network these mappings are generated with some non-linear activation functions, and renders the classifier highly applicable to non-linear classification problems.

$$f(x) = f^{(n)}(f^{(n-1)}....(f^{(3)}(f^{(2)}(f^{(1)}(x))) \quad (2)$$

Network Architecture

$f^{(n)}$ is generally composed of convolution and pooling layers which is depicted in Fig. 2.

The architecture could be represented as;

$C(w_{c1}; h_{c1}; d_{c1}; f_{c1}) - P(w_{p1}; h_{p1}) –$
$C(w_{c2}; h_{c2}; d_{c2}; f_{c2}) - P(w_{p2}; h_{p2}) –$
$C(w_{c3}; h_{c3}; d_{c3}; f_{c3}) - P(w_{p3}; h_{p3}) –$
$C(w_{c4}; h_{c4}; d_{c4}; f_{c4}) - P(w_{p4}; h_{p4}) –$
$FC(n) - LR(n); \quad (3)$

where;
w, *h*, *d*, denotes width, height, depth of the layers.
f denotes the number feature maps.
FC(n) is the fully-connected layer with input size n,
LR(n) is the softmax layer with input size n.

Optimization

In the optimization process, we employed the cross-correlation entropy cost function with the weight decay value of 0.001. Instead of using the relatively slow Stochastic Gradient Descent, we approximate the minima with ADAM method (Kingma and Ba, 2014). A batch size of 256 with 45 Epoch training is set during the whole training process.

4. EXPERIMENTAL RESULTS

We conduct our experiments on two different kind of datasets according to their generation sources. These two datasets are dense mobile laser scanner data, and a large scale synthetic point cloud data populated from an architectural CAD model. The overall segmentation results are in parallel with the complexity and the challenge expected from diversifying the sets. The overall classification accuracies are indicated at the right-bottom of the confusion matrices. CNN is implemented with MatConvNet library (Vedaldi and Lenc, 2015). All experiments are conducted with 5 cm voxel resolution.

4.1 Indoor Point Cloud Segmentation

Part of a mobile laser point cloud acquired by a TIMMS platform equipped with sideways Faro Scanners are used for real data evaluation (Fig.4). The cluttered indoor environment consists of two semi-detached rooms, of which another room resides within one of them. The point cloud consists of 4.5 million points in a total area of 90 m^2.

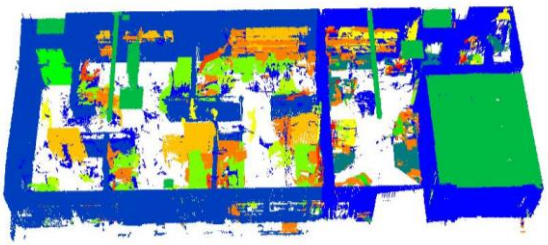

Figure 4. The real dataset is generated by a mobile laser scanner. The left hand side partition of the environment is set as the training site, where the right-hand side partition has both similar computer laboratory office characteristics, meanwhile a fairly different meeting room space is also present to test our algorithm (green part on the lower right).

For training data preparation, an attempt to benefit from pre-segmentation by using algorithms like connected component analysis or region growing turns out to be ineffective due to the highly cluttered environment. Hence we resort to the full manual classification of the points which proves to be laborious in large scale (Fig.3). As our method envisages the possibility for an online testing as a following study, the point cloud is not treated with a pre-processing of noise reduction. Nevertheless, there is %2 of the points which cannot be possibly recognized to be classified by the human operator. In order to see the potential of our method, we select the general computer lab. / office area as training site, and reserve the meeting room to the test site (Fig.4).

Table 2

class	***wall***	***desk***	***chair***	***human***	***shelf***	***object***	***monitor***
prec.	89.09	64.46	70.03	28.58	35.44	39.36	27.98
recall	95.21	75.90	47.11	41.50	15.76	34.85	12.62

Table 3

confusion	*wall*	*desk*	*chair*	*human*	*shelf*	*object*	*monitor*	*recall*
wall	1339110	8523	126	12714	28887	15105	2068	0.95
desk	4593	72624	5122	49	276	11778	1246	0.76
chair	1026	9045	22692	1627	685	12150	942	0.47
human	6657	145	478	8602	496	4348	0	0.42
shelf	104133	2030	10	2518	21983	7643	1192	0.16
object	28655	18165	3934	4584	9608	37335	4858	0.35
monitor	18919	2137	41	1	94	6532	4004	0.13
precision	0.89	0.64	0.70	0.29	0.35	0.39	0.28	Acc./ 0.81

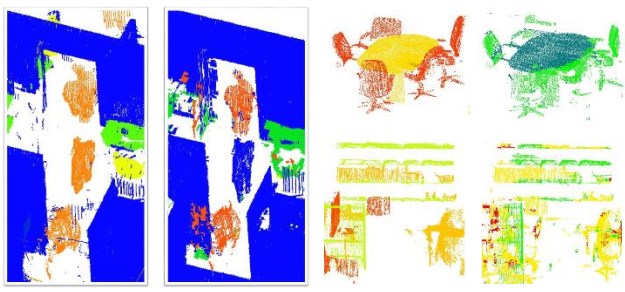

Figure 5. Particular details from segmentation results. Left hand side of the image pairs are the ground truth, while the right hand sides are the CNN predictions. a) of the 3 human in the scene, two of them correctly labelled while the one in the middle is mistaken as a wall; b) the table in the meeting room is generally segmented with surrounding chairs, despite been trained only from desk examples; c) shelves are fairly detected while monitors are largely missed.

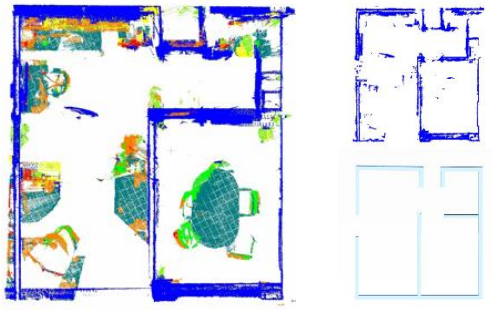

Figure 6. a) full semantic segmentation results from the top view; b) labelled wall points largely delineate the floor plan; c) CAD model digitized from the point cloud for quick reference.

Our method is evaluated with 7 classes, and the results are denoted in Table 2 and Table 3. The overall accuracy is 0.81. In particular the wall detection is very satisfactory, and promises a substantial leverage to our previous object–driven space partitioning framework (Babacan et al., 2016), whereas, accuracies for small objects, such as monitor, and shelf are not very gratifying. Object scale in classification appears as a problem, in addition to the class ambiguities themselves. For instance, the category shelf consist of both standalone bookshelves in any part of the room, and the longer wall-attached ones. Likewise, the category object includes anything of any size that could be counted as an object that is not covered in other categories, i.e. an artificial tree or a computer case. The other objects categories like desk, and chair appear to have fairly well exploitable geometric structures. We provide more exposure about the detailed results of some particular scenes in Fig.5.

The full segmentation is depicted in Fig.6a. By virtue of the very high accuracy wall detection results, the segmented wall points become very representative of the indoor model even in their raw format (Fig.6b). For reference we digitize a CAD model of the test site in AutoCAD Revit. Despite some individual erroneously classified small clutters, and a large cabinet occluding the small wall, the segmented wall points closely follow the indoor model.

4.2 Synthetic Data Segmentation

In order to gain different insight, we further analyse another indoor dataset, this time being a synthetic data devoid of any clutter, but increased to a floor-scale environment. A synthetic point cloud is populated from a CAD model of a basement consisting over 40 rooms, and corridors (Fig.7). The model is relatively simple in terms of the diversity of its categories; it only consists of the structural elements of the building as wall, floor, door, and beam. We generate 10 million points for the whole model, but only run the training with %20 of the points due to computation limitations.

As it could be seen from table 4 and table 5, CNN is very successful in dominant horizontal and vertical architectural structures such as walls, and floor, and also promising door detection results only from the door frame, as opposed to our previous complete door detection framework (Fig.8). Beams are also recognized in majority. Apart from a thin slice of a wall misclassified as beam, the results are satisfactory in general, and inviting for an object detection framework, as there is no other misclassification at the object scale. The overall accuracy is 0.89.

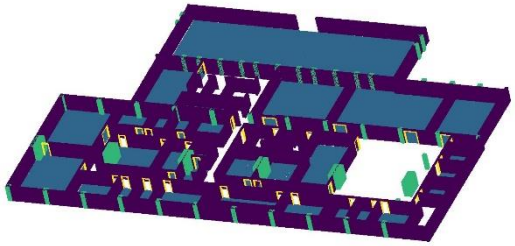

Figure 7. The synthetic point cloud generated from a CAD model.

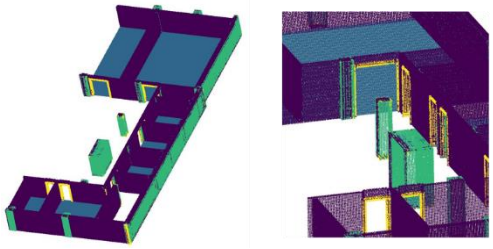

Figure 8. a) the result of the semantic segmentation, among a great majority of correctly labelled points, a façade wall is misclassified as a long beam; b) a close-up to the results, the door frames are recognized well.

Table 4

class	***wall***	***floor***	***beam***	***door***
prec.	94.68	93.97	42.12	76.58
recall	85.61	99.74	67.74	55.61

Table 5

confusion	*wall*	*floor*	*beam*	*door*	*recall*
wall	779685	30391	95341	5317	0.86
floor	1429	566507	40	0	0.99
beam	26652	5371	71726	2142	0.68
door	15703	595	3172	24389	0.56
precision	0.95	0.94	0.42	0.77	acc./ 0.89

4.3 Semantic Planar Segmentation

We finally display the power of CNN based semantic segmentation in indoor modelling by applying the segmentation results to drive a planar extraction on the point cloud. Our motivation is the fact that planar extraction could become extremely challenging in indoor environments, especially in the presence of high clutter and occlusion. We previously have circumscribed this problem by means of favourable slice selection. However, this solution necessarily limits the information content into a narrow 3D, even convert the problem into 2D line extraction. Leaving aside the pros and cons of this approach, we present an alternative that could effectively be employed directly in 3D.

Once equipped with semantic information, the approach is simple and straightforward; comes down to selecting the relevant categories for plane candidates, and applying the extraction algorithm individually to each individual category. Here we exhibit results for RANSAC algorithm (Schnabel et al., 2007) applied to the mobile laser scanner point cloud dataset.

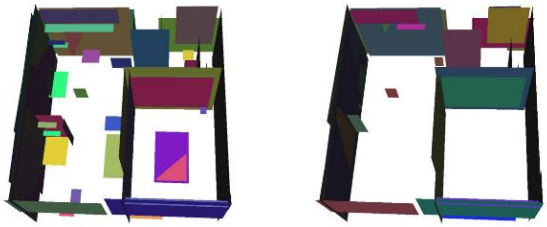

Figure 9. RANSAC planar extraction results; a) applied to the raw point cloud; b) applied to wall points detected by CNN. It can clearly be seen that the number of irrelevant planes declines, and a cleaner planar extraction is achieved.

As can be seen on Fig.9 the semantic selection of the wall points can effectively reduce the number of irrelevant planes to be deployed in geometric modelling, hence increase the percentage of wall corresponding planes in the overall extraction. This framework is also potentially beneficial in the overall space partitioning process, as demonstrated in our previous work.

5. CONCLUSION AND OUTLOOK

In this paper we propose a viable method to extract semantic information for indoor modelling. A convolutional neural network is designed for 3D data to obtain semantic segmentation of indoor point clouds. Experimental results demonstrate that the methodology can adapt to different kind of datasets, both real and synthetic at various densities with categorical assortment. A simple example of how this semantic information can be deployed to mitigate the challenge of geometry modelling is also given.

Yet, there is great room to improve the results. First, the inherent class ambiguity issue should be tackled. This problem is closely related to dataset size and variations, hence indoor modelling research community needs to give emphasis on producing large datasets with diverse categories. Present datasets are mainly for small scenes and / or object oriented; the relation between the real environments deprived of prior segmentation information should be established. Finally, we advocate that a geometry modelling could benefit immensely by considering semantics, which we strive to further in future study.

6. ACKNOWLEDGEMENTS

We wish to thank Applanix Company for providing the TIMMS mobile scanning, and pre-processing the point cloud. We are also grateful to Dr. Jungwon Kang for sharing the CAD model that enabled us to generate the synthetic data.

REFERENCES

Babacan, K., Jung, J., Wichmann, A., Jahromi, B. A., Shahbazi, M., Sohn, G., Kada, M. (2016). Towards object driven floor plan extraction from laser point cloud. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 41.

Badrinarayanan, V., Kendall, A., Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561.

Boulch, A., de La Gorce, M., Marlet, R. (2014, August). Piecewise-Planar 3D Reconstruction with Edge and Corner Regularization. In Computer Graphics Forum (Vol. 33, No. 5, pp. 55-64).

Budroni, A., Boehm, J., 2010. Automated 3D Reconstruction of Interiors from Point Clouds Automated 3D Reconstruction of. Int. J. Archit. Comput. 08, 55–74. doi:10.1260/1478-0771.8.1.55

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915.

Couprie, C., Farabet, C., Najman, L., LeCun, Y. (2013). Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.

Dumitru, R., Borrmann, D., Nuchter, a, 2013. Interior Reconstruction Using the 3D Hough Transform. 3D-Arch 2013 XL-5/W1, 25–26. doi:10.5194/isprsarchives-XL-5-W1-65-2013

Farfade, S. S., Saberian, M. J., Li, L. J. (2015, June). Multi-view face detection using deep convolutional neural networks. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 643-650). ACM.

Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J. M. (2015). Visual simultaneous localization and mapping: a survey. Artificial Intelligence Review, 43(1), 55-81.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009, September). Class segmentation and object localization with superpixel neighborhoods. In Computer Vision, 2009 IEEE 12th International Conference on (pp. 670-677). IEEE.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J. (2017). A Review on Deep Learning Techniques Applied to Semantic Segmentation. arXiv preprint arXiv:1704.06857.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. Int. J. Rob. Res. 31, 647–663. doi:10.1177/0278364911434148

Huang, C., Davis, L. S., Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. International Journal of remote sensing, 23(4), 725-749.

Huang, J., You, S. (2016, December). Point cloud labeling using 3d convolutional neural network. In Pattern Recognition (ICPR), 2016 23rd International Conference on (pp. 2670-2675). IEEE.

Ioannidou, A., Chatzilari, E., Nikolopoulos, S., Kompatsiaris, I. (2017). Deep Learning Advances in Computer Vision with 3D Data: A Survey. ACM Computing Surveys (CSUR), 50(2), 20.

Jung, J., Chen, L., Sohn, G., Luo, C., Won, J. U. (2016). Multi-Range Conditional Random Field for Classifying Railway Electrification System Objects Using Mobile Laser Scanning Data. *Remote Sensing*, *8*(12).

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

Kingma, D., Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koppula, H. S., Anand, A., Joachims, T., Saxena, A. (2011). Semantic labeling of 3d point clouds for indoor scenes. In Advances in neural information processing systems (pp. 244-252).

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).

Maturana, D., Scherer, S. (2015, September). Voxnet: A 3d convolutional neural network for real-time object recognition. In Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on (pp. 922-928). IEEE.

Monszpart, A., Mellado, N., Brostow, G. J., Mitra, N. J. (2015). RAPter: rebuilding man-made scenes with regular arrangements of planes. ACM Transactions on Graphics (TOG), 34(4), 103.

Mura, C., Mattausch, O., Villanueva, A.J., Gobbetti, E., Pajarola, R., 2013. Robust reconstruction of interior building structures with multiple rooms under clutter and occlusions. Proc. - 13th Int. Conf. Comput. Des. Comput. Graph. CAD/Graphics 2013 52–59. doi:10.1109/CADGraphics.2013.14

Nguyen, V., Gächter, S., Martinelli, A., Tomatis, N., Siegwart, R., 2007. A comparison of line extraction algorithms using 2D range data for indoor mobile robotics. Auton. Robots 23, 97–111. doi:10.1007/s10514-007-9034-y

Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1520-1528).

Nurunnabi, A., Belton, D., West, G., 2014. Robust statistical approaches for local planar surface fitting in 3D laser scanning data. ISPRS J. Photogramm. Remote Sens. 96, 106–122. doi:10.1016/j.isprsjprs.2014.07.004

Ochmann, S., Vock, R., Wessel, R., Tamke, M., Klein, R., 2014. Automatic Generation of Structural Building Descriptions from 3D Point Cloud Scans. Int. Conf. Comput. Graph. Theory Appl.

Oesau, S., Lafarge, F., Alliez, P., 2014. Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. ISPRS J. Photogramm. Remote Sens. 90, 68–82. doi:10.1016/j.isprsjprs.2014.02.004

Okorn, B., Xiong, X., Akinci, B., Huber, D., 2010. Toward Automated Modeling of Floor Plans. Proc. Symp. 3D Data Process. Vis. Transm. 2.

Qi, C. R., Su, H., Mo, K., Guibas, L. J. (2016). Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*.

Sanchez, V., Zakhor, A., 2012. Planar 3D modeling of building interiors from point cloud data. Proc. - Int. Conf. Image Process. ICIP 1777–1780. doi:10.1109/ICIP.2012.6467225

Schnabel, R., Wessel, R., Wahl, R., Klein, R., 2007. Shape Recognition in 3D Point Clouds Shape Recognition in 3D Point-Clouds. Symp. A Q. J. Mod. Foreign Lit. 2, 40–51. doi:10.1306/D4268F04-2B26-11D7-8648000102C1865D

Silberman, N., Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on* (pp. 601-608). IEEE.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Thoma, Martin. "A survey of semantic segmentation." arXiv preprint arXiv:1602.06541 (2016).

Thrun, S., Martin, C., Liu, Y., Hahnel, D., 2004. A real-time expectation maximization algorithms for acquiring multi-planar maps of indoor environments with mobile robots. IEEE Trans. Robot. 20, 433–443.

Vedaldi, A., Lenc, K. (2015, October). Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 689-692). ACM. Chicago

Volk, R., Stengel, J., Schultmann, F. (2014). Building Information Modeling (BIM) for existing buildings—Literature review and future needs. Automation in construction, 38, 109-127.

Xiao, J., Furukawa, Y., 2012. Reconstructing the World's Museums. Eccv 668–681. doi:10.1007/s11263-014-0711-y

Xiong, X., Adan, A., Akinci, B., Huber, D. (2013). Automatic creation of semantically rich 3D building models from laser scanner data. Automation in Construction, 31, 325-337.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D. Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1529-1537).