First Experiments of Deep Learning on LiDAR Point Clouds for Classification of Urban Objects

Y. Zegaoui[1,4] M. Chaumont[1,2] G. Subsol[1] P. Borianne[3] M. Derras[4]

[1] LIRMM, Univ. Montpellier, CNRS, France
[2] Univ. Nîmes, France
[3] AMAP, Univ. Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France
[4] Berger-Levrault, Toulouse, France
Corresponding author: *younes.zegaoui@lirmm.fr*

1 Presentation of the context

Large urban agglomerations nowadays are facing some major issues such as economic restrictions, environmental challenges, global and systemic approaches in city management [8]. One of them is to precisely monitor urban objects which can be natural (trees), artificial (traffic lights or poles), static or moving (cars). This is essential to analyze their mutual interaction (for example branches of a tree which are close to an electric pole) and to prevent risks associated with them (for example, dead parts of a tree which may fall on a street). Thus being able to localize urban objects and provide informations about their status is essential. One way of achieving this task is to mount LiDAR scanners on vehicles and analyze acquisitions performed on a regular time basis. However this requires to automatically process the 3D point clouds given by the LiDAR.
Automatic object detection has seen an increase of popularity in the recent years due to the rise of Deep-Learning methods [3] being able to achieve human-like performances for object classification in 2D images. This increase of performance is starting to have a huge impact on research in 3D shape recognition (see papers presented on the Web site[1]).
Given a complete scene obtained with a LiDAR scanner, i.e. a cloud of 3D points, the ultimate goal is then to localize and identify 3D urban objects. In this paper, we focus on the classification of small part of the 3D scene reduced to a single object. This requires to use a new Deep Learning method dedicated to 3D points. In the following, we briefly review existing methods and we present some preliminary results obtained with the PointNet network.

2 Deep Learning on 3D point sets

Existing Deep Learning methods for 3D point sets can be roughly divided in three categories.
Multi-view CNN methods (see for example [7]) create a set of 2D images from various camera angles for each 3D object which are then processed by a classic CNN. These methods tend to give the best results in term of accuracy but make the assumption that the object has been isolated and that the surface of the scanned object is complete. Those conditions are satisfied on CAD models in which multi-view networks perform very well but this is not usually the case in LiDAR acquisition where there are many occlusions especially in urban environment.
Voxel-based methods (see for example [4]), as the name suggests, require a pre-processing step where the point cloud is structured as a 3D image composed of a given number of voxels. This 3D image can be considered as a stack of 2D images and thus it is possible to adapt a 2D CNN so that it can take the voxel representation as an input. Nevertheless, as point cloud data are sparse, we need to work with huge 3D images in order to keep a good precision. Applying 3D convolutions can then saturate the GPU memory quite fast. Furthermore, these methods do not give better results compared to the others.
PointNet method [1], instead of using a transformation to get structured data, directly send the point cloud to the network. Published results are very good and the authors demonstrated that the method is less sensitive to occlusion than the previous ones which is an important property in the case of LiDAR acquisition. We then select PointNet framework for our experiments.

3 Description of PointNet network

Two major concerns when processing 3D points is to deal with the order of input points and the coordinate frame which are not the same for all the objects. PointNet tackle those issues by adding transformation modules and by using a symmetric function.

[1] *http://modelnet.cs.princeton.edu/*

The transformation module works as a joint alignment network. It computes a 3×3 matrix (which corresponds to an affine transformation) and multiply it with the input point cloud. The matrix coefficients are calculated by a mini-network which architecture is similar to the main network one. The idea is to align all input clouds to a canonical space before feature extraction.

After the transformation, the point cloud enters a series of MLP layers which extract features from each point. These features are then aggregated into global features by a max pooling layer. The combination of those two functions, the MLPs followed by max pooling, forms a symmetric function which is insensitive to the point order inside the input cloud.

4 Datasets

In order to create a significant training dataset for urban objects, we aggregated 3D point sets from 3 public databases[2] [2, 5, 6]. This dataset A of 720 3D point sets is composed of representative classes of urban objects (see Figure 1) as "tree" (154 samples), "car" (88), "traffic sign/light" (141), "pole" (22). We also add classes "person" (164), "building" (94) and "noise/scan artifact" (57) to enlarge distinctiveness possibility. The number of points of an object is less than 100 points for the class "person" and can go up to 3,000 for the classes "tree" or "building".

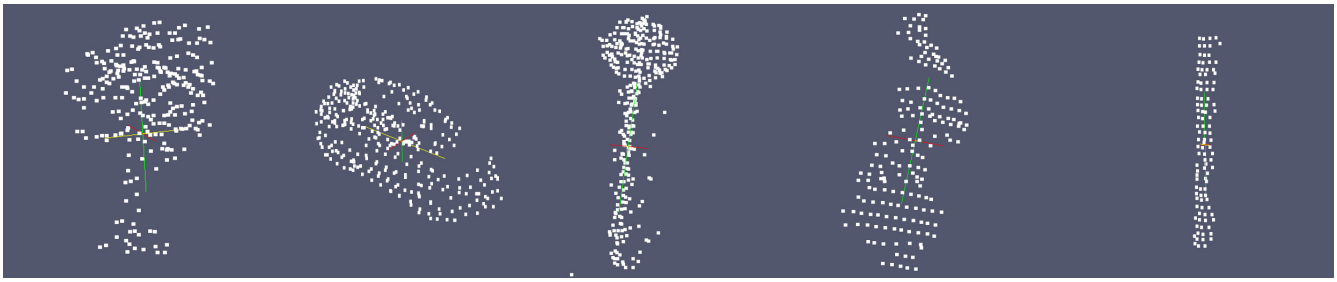

Figure 1: Dataset A was created by aggregating public databases. Left to right, a tree, a car, a traffic sign, a person, a pole.

We also performed a LiDAR acquisition with a Leica Pegasus backpack[3] carried by a walking person. The acquisition itinerary, about 200 meters long, was executed back and forth. In the point cloud (see Figure 2, left), we can recognize a tramway station, some trees with and without leaves and a small residential area. The resulting 3D point cloud contains over 27 million points and we were able to manually isolate 160 urban objects. The resolution of these objects is several times higher than objects in dataset A: more than 1,000 points for a person and more than 20,000 points for most of the trees. The resulting dataset B is composed of 75 trees, 39 cars, 8 traffic sign/light, 23 poles and 15 persons. We did not isolate building facades or noise artifacts.

Figure 2: Left: part of the LiDAR acquisition of an urban scene. We can recognize a tramway railway and trees. Right up: a tree was manually isolated in the point cloud (160,000 points). Right down: a downsampled version (267 points) by voxel-grid filter is used for classification.

[2] *www-personal.acfr.usyd.edu.au/a.quadros/objects4.html cmm.ensmp.fr/šerna/rueMadameDataset.html sites.google.com/site/kevinlai726/datasets*

[3] The Leica Pegasus is a dual Velodyne VLP-16 scanner coupled to 5 RGB cameras with a 600 point-per-second acquisition speed for a relative accuracy of about 3 cm for outdoor and indoor. We would like to thank Leica Geosystems for providing us the LiDAR acquisition.

5 Classification experiments

In order to perform experiments, the 3D point sets coming from datasets A and B are downsampled by applying a voxel-grid filter so that they contain less than 512 points. We then use the Tenserflow implementation of PointNet[4] with data augmentation by applying random rotations along the same axis and Gaussian noise jittering ($\sigma = 0.02$), Adam optimizer and batch normalization.

In a first experiment, the dataset A is separated in three subsets: 64% for the training set, 16% for the validation set and 20% for the test set. The overall accuracy of the classification (i.e. the number of objects correctly predicted over the total number of objects of the test set) is 0.718. Accuracy by class is very good for the urban objects "tree" (0.778), "car" (0.880) and "traffic sign/light" (0.793) but is very low for the class "pole" (0.111). Notice that the accuracy for the classes "building" (0.833) and "person" (0.739) is also high, at the contrary of the class "noise" (0.182).

In a second experiment, dataset A is the training set (separated in training and validation sets with a ratio of 80% and 20%) and we perform classification on dataset B. Contrary to the first experiment, the two datasets are completely different. Nevertheless, the overall accuracy appears high (0.743) but we do not test any building or noise samples. In Table 1, we display the confusion matrix. We can notice that classification results are satisfactory for urban objects "tree", "car". An exception is the confusion between the "traffic sign/light" class and the "pole" class, which can be explained by the similarity of those shapes, especially at low resolutions.

		Ground truth				
		tree (75)	***car (39)***	***traffic sign/light (8)***	***pole (23)***	***person (15)***
Classification	***"tree"***	69	2	0	8	0
	"car"	1	33	0	0	0
	"traffic sign/light"	4	0	4	12	2
	"pole"	0	0	3	1	0
	"person"	1	0	1	0	12
	"building"	0	0	0	2	0
	"noise"	0	4	0	0	1
	F measure	0.896	0.904	0.267	0.074	0.828

Table 1: Confusion matrix for experiment 2: we can see by column, the classifications for a given class of objects. In green and orange, we emphasize good results on urban objects whereas we can see in red some significant errors.

6 Future work

We plan to expand our databases by performing new LiDAR acquisitions and we will make them available. We are also currently working on detection, that is localizing automatically in the complete scene the different urban objects.

References

[1] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, July 2017.

[2] K. Lai and D. Fox. Object Recognition in 3D Point Clouds Using Web Data and Domain Adaptation. *International Journal of Robotics Research*, 29(8):1019–1037, July 2010.

[3] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[4] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, Sept 2015.

[5] A. Quadros, J. Underwood, and B. Douillard. An Occlusion-aware Feature for Range Images. In *Robotics and Automation, 2012. ICRA'12. IEEE International Conference on*. IEEE, May 14-18 2012.

[6] A. Serna, B. Marcotegui, F. Goulette, and J.-E. Deschaud. Paris-rue-Madame Database - A 3D Mobile Laser Scanner Dataset for Benchmarking Urban Detection, Segmentation and Classification Methods. In *ICPRAM*, 2014.

[7] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, Dec 2015.

[8] I. Zubizarreta, A. Seravalli, and S. Arrizabalaga. Smart City Concept: What It Is and What It Should Be. *Journal of Urban Planning and Development*, 142(1):04015005, 2016.

[4] *http://github.com/charlesq34/pointnet*