A survey of Image Classification and Object Detection based on 3D data

Xiaoke Shen

The Graduate Center, City University of New York

Abstract

Recently, by using the deep neural network based algorithms, the object classification and detection can achieve a new state of the art. For some scenarios, the deep neural network based algorithms can achieve a similar or even have a better performance on 2D image classification/detection and segmantic segmentation than the human expert. However, one main drawback of only using the 2D images is even the boudning box even the pixel level recognization can be well done based on 2D images, the accurate info such as the real location of the object can still not be well collected as the drawback of the 2D image data representation itself such as extension distortion. Meanwhile, as the 3D images, such as 3D cloud point data, can well preserve the accurate location info and sturcture of the objects, they are widely used to resolve the location sensative problems such as Self-Driving Cars and Robot Visions.

In this survey, the both the main algorithms used in the 2D image and 3D image based object classification/detection and semantic segmentation are surveyed. Whether some algoritms used in the 2D can be adjusted in the 3D scenario will be discussed. The 3D only algorithms will also be disussed. Finally, some potential algorithms in the 3D data based object classification/detection will be discussed.

Part I
Introduction

Following the general problem solving approach in both science and engineering area, in order to resolve a problem, we should do the data representation and develop the algorithm/model and then try to resolve those problems. Based on this approach, there is no difference for the 2D images and 3D images based methods. Then, there must be some difference from the performance's perspective. Generally speaking, by introducing the 3D data representation of an object, the performance should be at least the same as the 2D data representation approach as the 3D data representation of the real object will introduce more information compared with the 2D only representation.

Currently, great achievements have been shown in the 2D images area by using the deep neural networks. Actually, the neural network is not a new approach. It was first introduced in 1950s. In 1958, Rosenblatt [1] created the perceptron, an algorithm for pattern recognition. The perceptron algorithm's first implementation, in custom hardware, was one of the first artificial neural networks to be produced. Although the perceptron initially seemed promising, Neural network research stagnated after machine learning research by Minsky and Papert (1969) [2], who discovered two key issues with the computational machines that processed neural networks. The first was that basic perceptrons were incapable of processing the exclusive-or circuit. The second was that computers didn't have enough processing power to effectively handle the work required by large neural networks [3]. The first issue was resolved by introducing more layers of networks and the second issue was resovled by both reducing the complexity of the algorithms and by introducing more powerful computing hardware such as GPU.

By using the deep neural network based on algorithms, especially the convolutional neural networks based algorithms, the computer vision based on 2D images have been making great achievements in image classification, object detection and semantic segmentation since the year 2012. Meanwhile, the 3D image based algorithms have also greatly developed in the past 5 years. In this survey, the main techniques in the deep learning algorithms based on the 2D image data will be introduced. The main algorithms used in the 3D data based vision tasks will also be introduced.

3D image data

For the 2D image data, we can easily collect them in our daily life as the popularity of the smart phone which have at least one camera. The data representation for the 2D image is exactly a two dimensional array with red, green and blue channels. And for a specified row and a specified column, we have the basic unit of the image which is a pixel. For each pixel, the data is commonly represented by a number between 0 to 255. Most people are familar to this 2D image representation method.
Compared with the 2D image, 3D image is not common to the public(at least as of the time this survey is done). However, the 3D images are becoming more and more important and are widely used in reconstructing architectural models of buildings, navigation of self-driving cars, detection face(such as face ID for iphone X), preservation of at-risk historical sites, and recreating virtual environments for the film and video game industries.
Mainly, there are two kinds of hardware available to do the 3D data generation: outdoor and indoor. For the outdoor, one typical hardware is LIDAR(Light Detection and Ranging). The coverage of this equipment can achieve to hundreds and even thousands meters. Googles self-driving car has the LiDAR scanner. For the indoor hardware, in recent years the availability of low-cost sensors such as the Microsoft Kinect have enabled the acquisition of short-range indoor 3D data at the consumer level. Meanwhile, the smart phone such as iPhone X will have the depth camera to catch the 3D image. The robots such as fetch have the layers to collect 3D data. In figure 1, one example of the 3D data collected from the outdoor urban LIDAR scanner is shown. In the figure 2, the depth map generated by the kinect is provided. Also the robot vision of the 3D environment is shown in figure 3.

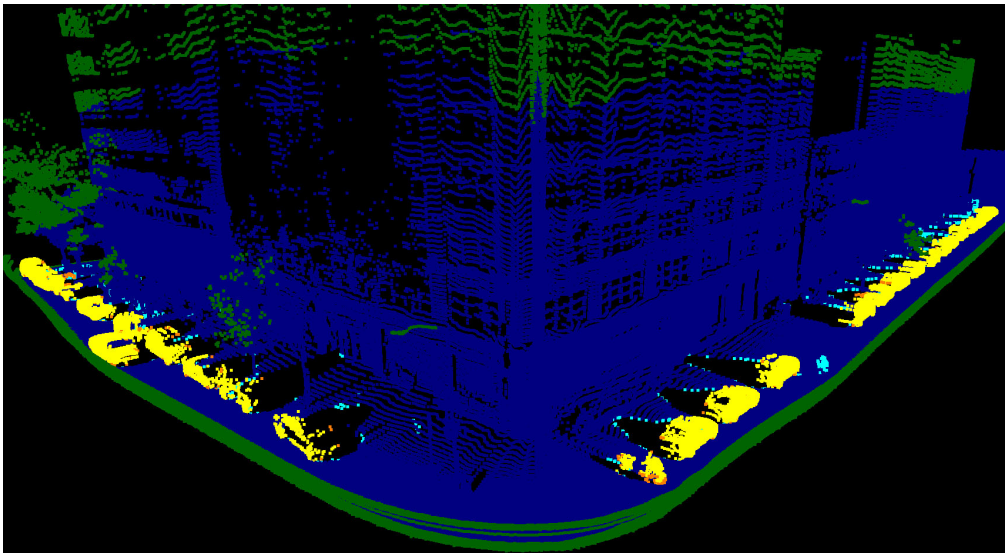

Figure 1: example of the 3D data from the outdoor urban LIDAR scans [4].

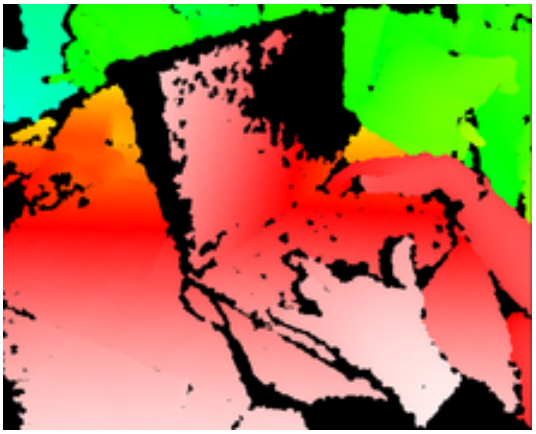

Figure 2: The depth map geenerated by Kinect. The depth map is visualized here using color gradients from white (near) to blue (far) [5]

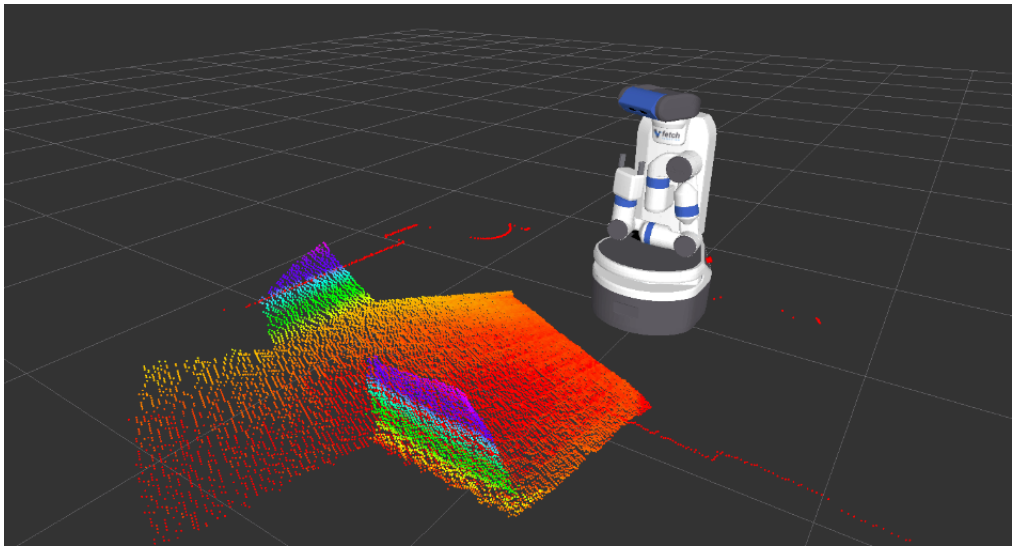

Figure 3: The 3D view from the robot of Fetch's perspective.

Image Classification and Object Detection

In the computer vision research area, researchers focus on resovle the two main problems: what and where. For the image classification problems, the objective is get what info of this image such as whether this object is a cat or dog. For the object detection problem, the task includes two parts: what and where. For what, the interesting objects will be labelled as a category to infer the what exatly this object is. For the where part, a bounding box will be proposed. This kind of research approach is similar in both 2D and 3D based data.

In this report the papers related to the object classification, object detection and object semantic segmentation for the 2D and 3D objects will be discussed. As most of the state of the art algorithms used for those tasks are based on the deep convolutional neural networks, the important papers related to the deep neural networks will also be discussed in this article.
The structure of this article is described bellow: The papers related to the theory part of the deep learning will be discussed in the second session. In the third session, the papers in the 2D object classification will be discussed. In session 4, the 2D object detection papers will be studied. At the same time, an interesting and more challenge related to the 2D image procession or computer vision will be discussed in session 5 which is the 2D object semantic segmentation. In the rest part of this article, the similar tasks in 3D

will be discussed as the final goal of this paper review is finding some possible approaches to improve the current 3D computer vision algorithms based on the state of the are 2D computer vision algorithms.

Part II
Deep Learning Theory

Approximation with Artificial Neural Networks

In order to build the mathematical theory of the artificial neural networks, several papers are published in the 20 century. The universal approximation theorem with proof is given in [6]. The theorem claims [6] that the standard multilayer feed-forward networks with a single hidden layer that contains finite number of hidden neurons, and with arbitrary activation function are universal approximators in $C(R^m)$. The universal approximation theorem is one of the important theoretical support for the artificial neural networks. However, at that time as the huge size labeled data sets are not available, the updated algorithms haven't been invented and also the limited computation power, these ideas can not be verified.
Inn the year of 1991 Kurt Hornik published a paper [7] showed that it is not the specific choice of the activation function, but rather the multilayer feedforward architecture itself which gives neural networks the potential of being universal approximators. This is an important contribution which is the foundation for the current state of the art deep learning architecture such as VGG 16 [8] and resnet [9] The paper of 1986 significantly contributed to the popularisation of BP(Back Propagation) for NNs [10], experimentally demonstrating the emergence of useful internal representations in hidden layers. The Back Propagation algorithm is one of the most critical and fundamental algorithm used in the deep neural network.

Part III
Algorithms developed based on 2D Images

Image Classification

The traditional algorithms used for the image classification are nearest neighbor and the SVM. And the features are the flattened pixel values. In th year 1989, the first important application of using the BP(Back Propagation) to well resolve the real life problem from the literature appeared. From this paper, one important structure of the neural network as show in Figure 4 including the layers with filters were introduced. The similar structure is used in the modern neural network structures such as Alex Net [11], VGG 16 [8] and resnet [9]. The basic idea of the convolutional neural network was also introduced here.

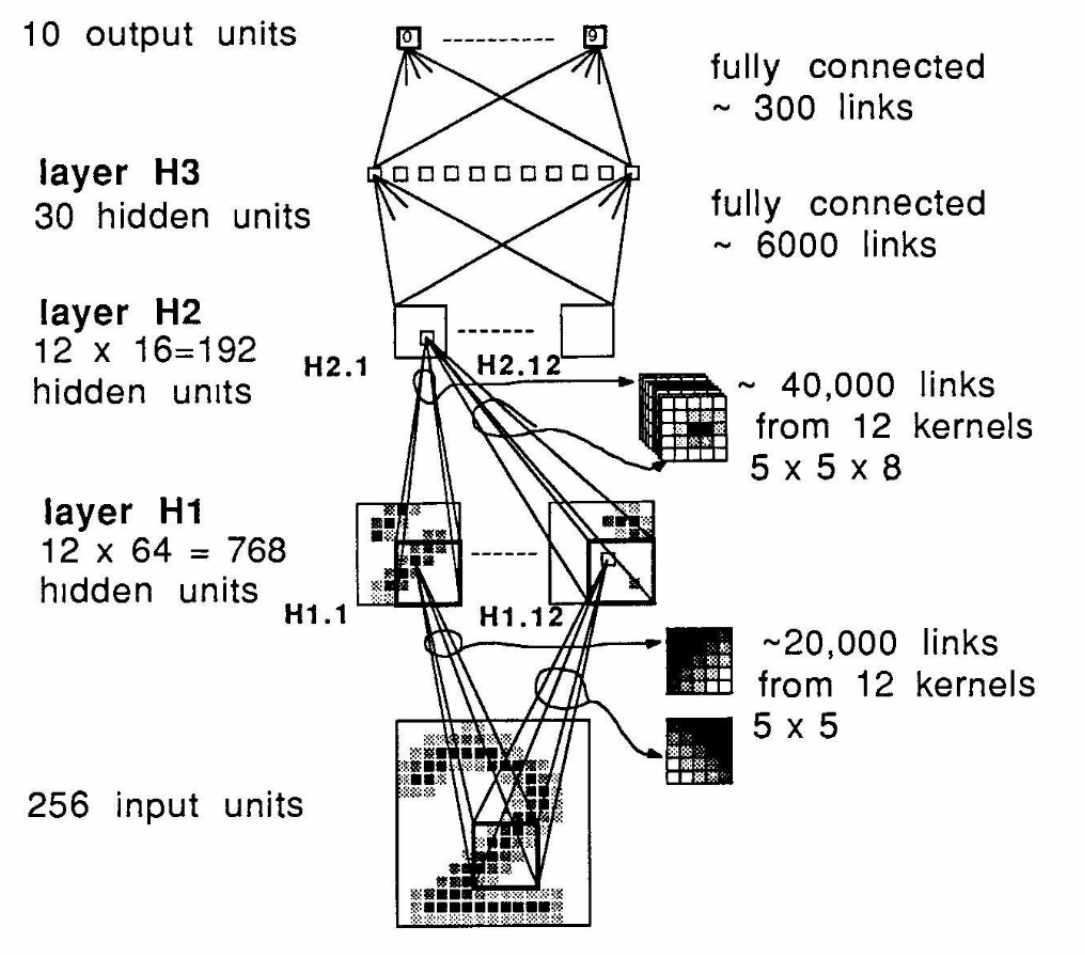

Figure 4: The Neural Network used in [12].

One important reason of the taking over of the deep learning algorithms in the computer vsion area is because of the CNN. Another important reason is because of the availablity of the large labelled dataset. As we know, in the machine learning research area, two kinds of learning approaches can be done: supervised learning and unsupervised learning. For the supervised learning algorithms, the labeled data is required to train the algorithm. So the availability of the labeled data will become very important to the development of the supervised learning based algorithms. The ImageNet dataset [13] provides 1.2 million high-resolution labeled images of 1000 categories. This dataset becomes one of the most important datasets related to the object classification.

In the ILSVRC-2012 competition, the method following discloused in the [11] achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry. The outstanding performance of the deep neural network used in this paper brought the focus of both the academic and industry back to the neural network again. In this paper, the CNN network is used and the CNN network structure is illustrated in the Figure 5. Two GPUs were used to speed up the calculation. Dropout [14] was used here and was proved to be effective to reduce the overfitting problem. The cons of this network is it is a bit complicated and the structure is not so elegant and it will be addressed in the future works by the deep learning researchers. This structure is called as the Alex net to emphasis the unique contribution of the author of this paper.

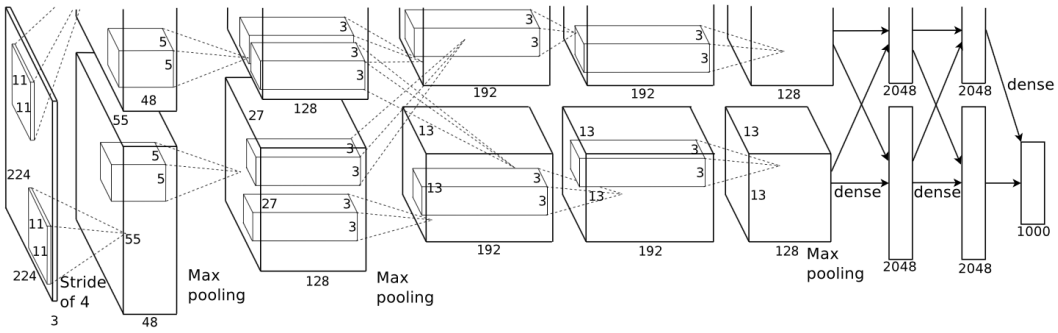

Figure 5: An illustration of the architecture of Alex Net, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network?s input is 150,528-dimensional, and the number of neurons in the network?s remaining layers is given by 253, 440-186, 624-64, 896-64, 896-43, 264-4096-4096-1000 [11].

Two years after the Alex Net, one well orginized VGG network is proposed in paper [11]. The structure of this network is very tidy and elegant. Different to the Alex Net [11] with different size of the convolutional kernels, in the VGG CNN network [8] , only size of 3 by 3 kernels were used for the whole network. At the same time, the trained network weights based on the ImageNet [13] dataset were shared to public. The cons of this model is there are too many parameters and the computation took a long time. Weeks for doing the ImageNet [13] dataset by using a powerful GPU.

This elegant design can also be used in the 3D CNN network to address the 3D object classification problem which will be shown later.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 6: An illustration of the VGG network structure.

In the year of 2015, He proposed a new structure of the neural network called resnet [9] and this new structure increased the layer of the network to more than 100 layers with less parameters than the VGG16 model and better performance. The motivation of this work is trying to find a deeper neural network structure to achieve better performance as the deepest network prior this work is around 20 to 30 layers. The authors had a basic assumption that

if more layers are added, the performance should be at least the same as the less layer networks as the new added layers can be designed as the identify layer then the performance can be at least the same. However, the traditional network's performance decreases when the layers increases to a higher level as shown in Figure 7. In order to address this anti-intuitive problem, the authors of this paper designed a deep residual network by adding a short cut between every other layers and finally it shows that this network can achieve a better performance. The problem is addressed. In this paper, the layer of the neural network can goes to 110 layer and have a better performance than the previous state-of-the-art neural network such as VGG16 [8], which only has 16 layers neural network. However, as mentioned in [9], the parameters used for the 110 layers Resnet is even less than the 16 layers VGG16 network. This seems amazing. However, by reading this article carefully, the main contribution of the reduction of the parameters was using the convolutional network layer instead of the fully connected layer. There is no contribution on the parameters reduction by the Resnet itself.

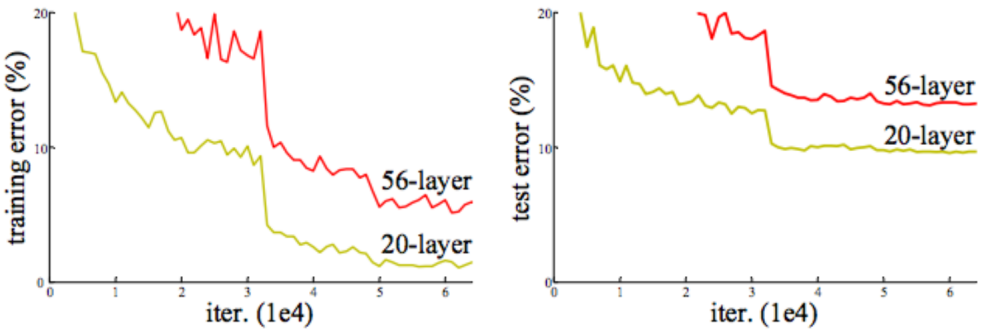

Figure 7: An illustration of deep neural networks fails by using the traditional network structure

Alongside of the changing of the structure of the deep neural networks, another algorithm are proposed to speedup the convergence rate of the training process. This algorithm is called BN(Batch Normalization) [15].

The main contribution of this paper as shown in Figure 8 is it greatly reduced the convergence time for the training process and the BN become one of the standard training step for the deep neural network after the publish.

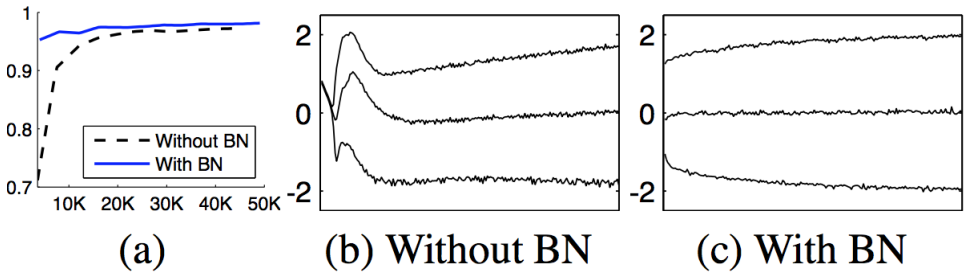

Figure 8: (a) The test accuracy of the MNIST network trained with and without Batch Normalization, vs. the number of training steps. Batch Normalization helps the network train faster and achieve higher accuracy [15]. (b, c) The evolution of input distributions to a typical sigmoid, over the course of training, shown as 15, 50, 85th percentiles. Batch Normalization makes the distribution more stable and reduces the internal covariate shift [15].

Object Detection

In the object detection research area, just like in the image classification where the imagenet dataset is available to train the algorithm, in the object detection, COCO(Common Objects in Context) [16] and VOC dataset can be used to train the algorithms. For the COCO dataset, the bounding box and the mask of the objects are provided. Then the algorithm can use those info to be trained. The first important contribution of using the deep neural network to resolve the object detecgtion problem is the R-CNN [17]. From this paper, the interesting region of an image is proposed by the selective search algorithm, and then the object detection problem is changed to the object detection problem. However, as every proposaled interesting area will be calculated to predict whether that specified region is an object, the speed of this algorithm is very slow. In order to address this problem, a fast R-CNN algortihm is proposed in [18] by improving the feature map generation efficiency. In another paper which is short for faster RCNN [19], instead of using the SS algorithmto generate the region of interesting(ROI), the ROI is proposed by using the deep neural network structure. By the combination, the performance of the algorithm can also be improved.
Another import paper for the object detection is YOLO [20] [21]. For the YOLO algorithm, the speed of detection is faster than the RCNN approach, however, the performance is slightly reduced.
In the year of 2017, the FPN [35] is used to do the object detection. The

basic idea for this network structure is combing different sized network together to boost the performance.

Semantic Segmentation

For the semantic segmentation, a pixel level detection of an object is provied. One important paper in this area is FCN [26]. It upsampled the feature map to make sure a more accurate location info can be preserved, also the data in the previous layers are combined with the deeper layer to preserve more info to help improve the accuracy of the semantic segmentation. After this paper, the FCN becomes the mainstream in the semantic segmentation. DeepLab [27], FCIS [28] and mask-RCNN [29] are using the structure based on the FCN. For the DeepLab [27], the CRF is used to further improve the result by taking benefit of the redundance info of nearby pixels. The CRF approach is firstly introduced in the paper [33].For the FCIS [28], location sensitive feature maps are generated to improve the pixel level prediction. The FCIS [28] is the improved version of the [36] where the region info is used. So far, the FCN and the CRF approach become the standard method in the semantic segmentation area. Besides those papers, [34] [32] are also using the deep neural network approach to improve the performance of the semantic segmentation.

Part IV
Algorithms developed based on 3D Images

Image Classification and Object Detection

Before the Deep Leaning era, some online algorithms were used to do the object detection in the real time. In [30] the object is detected by collect the histogram of the angles of an object. And the cusum technolog is used to do the online detection.

For the 2D images, one import data set is KITTI [24]. The depth info

is provided to give more info about the object.In [22] is a paper to combine the 2D and 3D database info together to reconstruct the 3D model of one object. For the Hand Pose Estimation from the depth images paper [23], the 3D CNN network is built to boost the performance of the algorithm. In [31], the depth info is also used to boost the performance of the network. Multi-View 3D Object Detection Network for Autonomous Driving [25], the 3D CNN is used on the 3D cloud data and the 2D CNN is used on the image data. Finally, the 2D and 3D are combined together to further improve the results.

References

[1] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, 1958.

[2] M. Minsky, "Perceptrons: An introduction to computational geometry," *MIT Press. ISBN 0-262-63022-2*, 1969.

[3] `https://en.wikipedia.org/wiki/Artificial_neural_network`. Accessed: 2017-10-20.

[4] A. Zelener and I. Stamos, "Cnn-based object segmentation in urban lidar with missing points," *International Conference on 3D Vision*, 2016.

[5] `https://en.wikipedia.org/wiki/Kinect`. Accessed: 2017-10-20.

[6] B. C. Csji, "Approximation with artificial neural networks," pp. 11–12, 2001.

[7] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. vol. 4, 1991.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 10 1986.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[14] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[16] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[17] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013.

[18] R. B. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1440–1448, IEEE Computer Society, 2015.

[19] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 91–99, 2015.

[20] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.

[21] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *CoRR*, vol. abs/1612.08242, 2016.

[22] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," *CoRR*, vol. abs/1611.07759, 2016.

[23] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[25] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1605.06211, 2016.

[27] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016.

[28] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," *CoRR*, vol. abs/1611.07709, 2016.

[29] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017.

[30] H. Z. Ioannis Stamos, Olympia Hadjiliadis and T. Flynn, "Online algorithms for classification of urban objects in 3d point clouds," *In 3D Imaging, Modeling, Processing, Visualization and Transmis- sion (3DIMPVT)*, 2012.

[31] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *CoRR*, vol. abs/1301.3572, 2013.

[32] P. H. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," *CoRR*, vol. abs/1506.06204, 2015.

[33] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," *CoRR*, vol. abs/1502.03240, 2015.

[34] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," *CoRR*, vol. abs/1411.5752, 2014.

[35] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016.

[36] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," *CoRR*, vol. abs/1605.06409, 2016.