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.

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.
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, and soon projects like Googles Tango will bring depth sensors to tablets and other mobile devices.

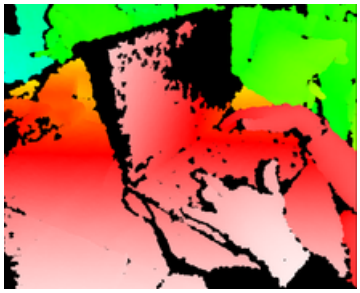

Figure 1: The depth map geenerated by Kinect. The depth map is visualized here using color gradients from white (near) to blue (far) [4]

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.

Deep Learning Theory

Approximation with Artificial Neural Networks [5]

In order to build the mathematical theory of the artificial neural networks, several papers are published in the 20 century. In this paper one main contribution is the universal approximation theorem with proof. The universal approximation theorem claims [5] 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.

Approximation Capabilities of Multilayer Feedforward Networks [6]

The unique value of the Kurt Hornik (1991) [6] paper is it 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 [7] and resnet [8]

Learning representations by back-propagating errors [9]

The paper of 1986 significantly contributed to the popularisation of BP(Back Propagation) for NNs [9], 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. This paper will be read in the future.

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift [10]

The main contribution of this paper as shown in Figure 2 is it greatly reduced the convergence time for the training process and the BN(Batch Normalization) become one of the standard training step for the deep neural network after the publish of this paper.

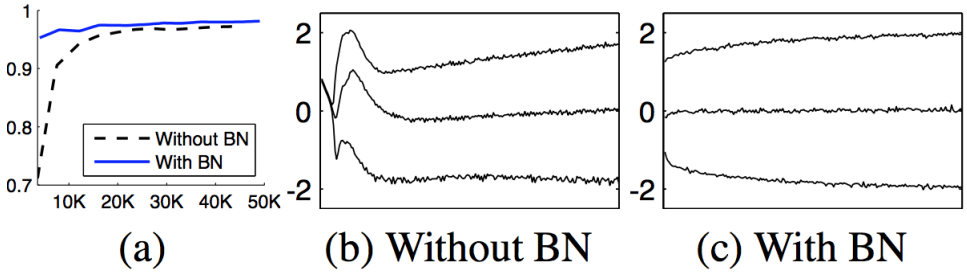

Figure 2: (a) The test accuracy of the MNIST network trained with and without Batch Normalization, vs. the number of training steps. Batch Normalization helps the network train faster and achieve higher accuracy [10]. (b, c) The evolution of input distributions to a typical sigmoid, over the course of training, shown as 15, 50, 85th percentiles. Batch Normalization makes the distribution more stable and reduces the internal covariate shift [10].

Object Classification for 2D Images

Backpropagation applied to handwritten zip code recognition [11]

The first important application of using the BP(Back Propagation) to well resolve the real life problem from the literature. From this paper, one important structure of the neural network as show in Figure 3 including the layers with filters were introduced. The similar structure is used in the modern

neural network structures such as Alex Net [12], VGG 16 [7] and resnet [8]. The basic idea of the convolutional neural network was also introduced here.

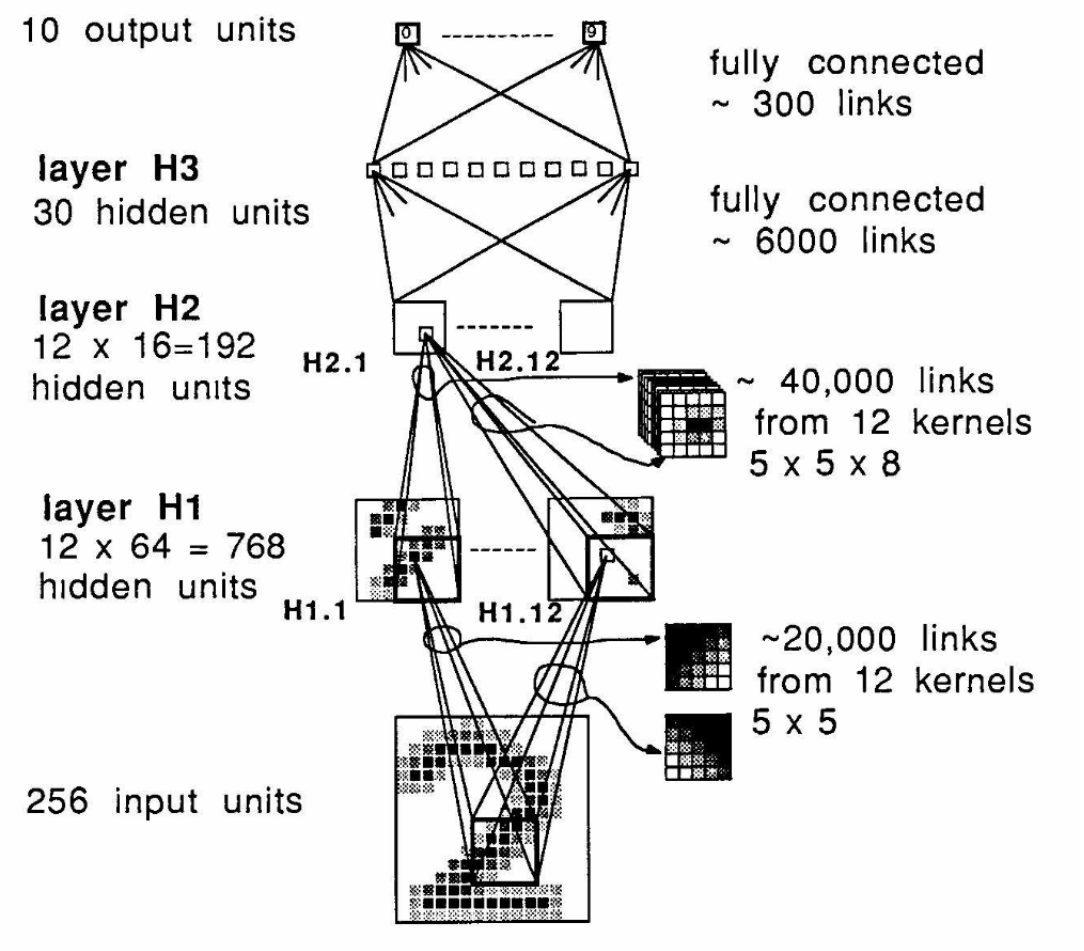

Figure 3: The Neural Network used in [11].

ImageNet: A Large-Scale Hierarchical Image Database [13]

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.

ImageNet Classification with Deep Convolutional Neural Networks [12]

To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" [14] that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.
The CNN network structure is illustrated in the Figure 4

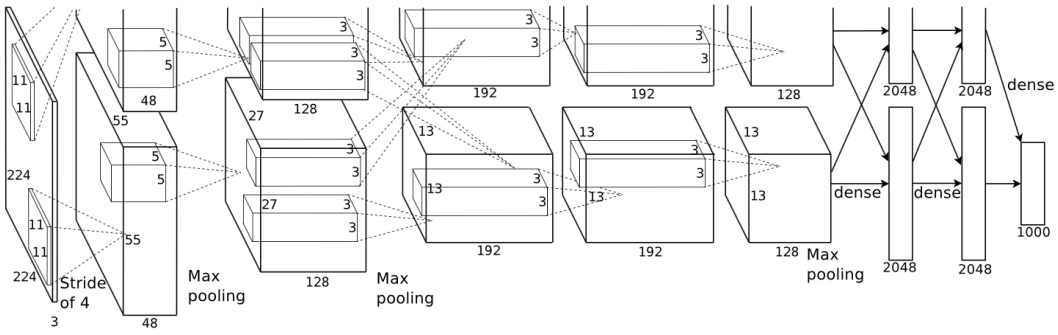

Figure 4: An illustration of the architecture of Alex Net, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network?s 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 [12].

Very Deep Convolutional Networks for Large-Scale Image Recognition [7]

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 5: An illustration of the VGG network structure.

Deep Residual Learning for Image Recognition [8]

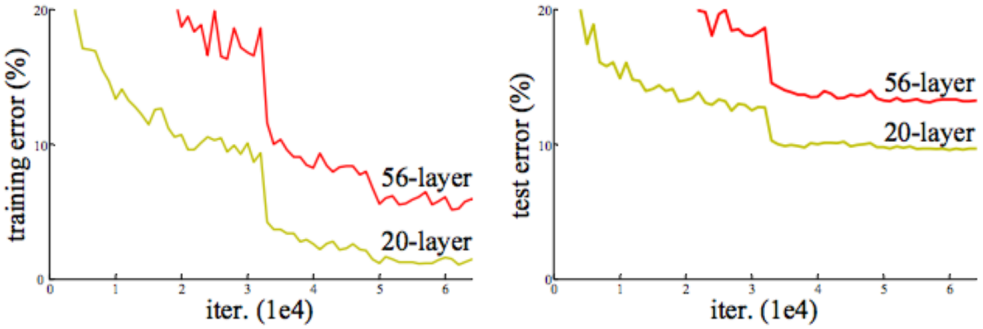

Figure 6: An illustration of deep neural networks.

Object Detection for 2D Images

Microsoft COCO: Common Objects in Context [15]

Rich feature hierarchies for accurate object detection and semantic segmentation [16]

Fast R-CNN [17]

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [18]

You Only Look Once: Unified, Real-Time Object Detection [19]

YOLO9000: Better, Faster, Stronger [20]

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$_n eural_n etwork$. *Accessed* : $2017 - 10 - 20$.

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

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

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

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

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

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

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

[11] 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.

[12] 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.

[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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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