3D Object Detection in the Wild

Matt Clifford
Supervised by Dr R. Santos-Rodriguez

November 28, 2018

1 Introduction

With recent advancements in mixed reality technology from Google glass and Microsoft HoloLens, there is an interest in the use of mixed reality to help create 'smart spaces', where mixed reality users are informed and guided through environments unknown to them.

Fracture Reality [21], has specific interest in 3D object detection for future projects. They specialise in creating bespoke mixed reality software for both private and government sectors. They mostly work with mixed reality visualisations of maps of an environment, for example to aid the control centres in airports [22]. Although they have many projects that would benefit from 3D object detection, an ongoing project investigating how the affect of mixed reality is in tackling circulation issues. The use of 3D object detection would identify and locate objects such as stairs, doors, elevators and escalators. This aids the user in identifying if these objects are correct pathway for them, addressing bottle neck problems. A variant use of this would be to help the visually or navigationally impaired, with mixed reality help from detected objects of importance.

The objects of interest for the object detection are on a case to case basis, due to specific needs of each project. Fracture Reality are able to create some object specific use case data, but in the region of hundreds of examples, and due to the specificity of each task, finding existing object datasets might not be possible. This leaves retraining a detector for every new object is infeasible. The need for a system that can detect objects given as little training examples as possible is therefore needed. This can be achieved by utilising general knowledge learnt from similar or relevant tasks and quickly applying it to the specific task of interest [4].

2 Literature review

In 2012, Alex Krizhevsky et al. revolutionised computer vision with a convolution neural network (CNN), inspired by [16]. It performed image classification on the ImageNet dataset [26]. The CNN, named AlexNet, consists of 5 convolution layers followed by 3 fully-connected layers. It

won ImageNet's ILSVRC-2010 and ILSVRC-2012 image classification contests [14]. In [14], they claim 'All of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available', and since AlexNet's success there have been consistent advancements in CNN state of the art. In 2014, [28] propose a CNN named VGG16, consisting of 16 convolution layers followed by 3 fully-connected layers. VGG16 achieves a top-5 error rate of 7.4% on ILSVRC-2014 compared to AlexNet's 17.0%. [28] achieves this performance boost by using smaller convolution filters, 3x3 compared to AlexNet's 11x11 which decreases the number of weights to train at each convolution layer, alongside a deeper convolution architecture, which can extract deeper relevant semantic meaning from the images. Further performance improvements were made by the use of 'inception models' [32][33], which stack the outputs of several convolutions of the same input, followed by filter concatenation. As well as 'residual learning' [13], which connects the outputs of multiple convolution layer. [31] formulates a combination of 'inception' and 'residual' models. Further improvements to 2D image classification include [17] [40].

To identify the location of objects in images, [9] uses a regional CNN (RCNN) model. Regional proposals of the image, which are run though a modified AlexNet. The positive classification results from the regional proposals are then adjusted using a linear regression model to obtain better object bounding boxes. Computational speed ups are proposed in Fast-RNN [8], which pools the regional proposals. Further computational speed ups are proposed in Faster-RNN [24], which combines the selective search regional proposals into the CNN. [23] proposes a grid search grid method rather than the more expensive regional proposals approach, known as 'you only look once' (YOLO).

[27] proposes a method of using the YOLO object detection approach in 3D space named Complex-YOLO, using only point cloud data from LIDAR depth sensors. Complex-YOLO uses a Euler-Region Proposal Network which estimates the orientation of objects by adding an imaginary and real part for each proposal box box. This results in 5 times speed up in object detection from the previous state of the art, with on par or better accuracy evaluated on data from KITTI benchmark suit [7].

The KITTI benchmark suit is an autonomous driving dataset with 200,000 3D object annotations captured in cluttered scenarios, with up to 15 cars and 30 pedestrians in each image. The data is obtained from a stereo camera and LIDAR sensor mounted on top of a car that is driven in the real world. Although the KITTI benchmark suit is a rich 3D object dataset, it is not as directly applicable to mixed reality application due to the sensing quality differences between LIDAR and portable depth camera. As well as KITTI only focusing on 8 autonomous driving classes such as pedestrians, cars and bicycles. [29] is a large-scale 3D object dataset with 32040 object poses and 45 different objects. The point cloud data is triangulated from 11 different views, making highly detailed scenes. The scenes are controlled and do not represent what would be captured from mixed reality depth sensors due to the triangulated different views. SUNRGBD benchmark suite [30] is a 3D object dataset consisting of 10,335 images with 64,595 3D object bounding boxes. The data is collected on various portable RGBD cameras such as the Kinect device, with indoor scenes focusing on objects such as doors, tables and chairs. A similar

quality popular dataset is the Pascal Visual Object Classes (VOC)2012 [6], which consists of 11,530 annotated object images, indoors and outdoors with 20 classes such as chairs, cars, dogs. However, PASCAL VOC 2012 only consists of 2D data. [35] extends the PASCAL VOC 2012 dataset with proposed 3D CAD style projections of the 2D objects. Although rich, this 3D data of the object is dissimilar to that of a depth sensor, leaving SUNRGBD the most suitable starting dataset for this project.

[5] proposes a 'cut and paste' style approach to synthesising 2D object detection datasets. First an object mask is predicted for the object, which is then applied to the image to 'cut and paste' the object into background scenes. Occlusions, truncations and blends are then applied to the object, helping it fit more naturally into the scene. [11] shows that training scene detectors on synthetic data produces comparable results on real life tests, with object detectors trained on the SUNRGBD state of the art dataset.

Current state of the art 3D object mostly use LIDAR derived point clouds from the KITTI benchmark suite [37][36][18][39]. Others use controlled, detailed objects [3][1][34]. Using this type of model directly is unsuitable for mixed reality inputs. [20] and [15] use 2D object detectors to aid the regional proposal of the 3D object, by searching only the 3D space in the point cloud occupied from the projected frustum obtained from the 2D object detector. This vastly reduces the search space and determines the object class for the 3D detector, resulting in improved speed and detection compared to using point cloud data alone, especially if the object suffers from occlusions or has a sparse representation. An alternative approach could combine object mask detectors[2][12] with the 3D projection of the mask to help further refine the 3D object search space. [25] uses latent support surfaces for 3D object detection on the SUNRGBD dataset. Another notable 3D object detector that use the SUNRGBD dataset is [10].

Transfer learning is a well studied area of deep learning [4][19][38], where a network trained for a specific task is re-purposed for a similar task. This is often achieved by truncating the last few layers of a pre-trained network where the network is specific to the trained task, and keeping the starting layers that have more general representations. Since the start of the network is already trained for general tasks relevant to both the original training task and the new desired task, only the last few layers need to be re-trained for the new task. This can be done using considerably less training data than the original network was trained with. Transfer learning could be used to help solve the problem of using as little 3D data as possible to train a 3D object detector.

Common representation – do a little research,, Auto encoders to possibly solve this

Unsupervised learning for creating data cheaply

3 Project plan

- collect dataset

- CNN as baseline for classification

- faster-RCNN as baseline for 2D object detection

- Make baseline 3D object detector using point cloud data

- Transfer 2D model to 3D, elaborate on this

- Assess results of transfer learning 3D detector with a few 3D new object examples

- Auto-encoder to find common representation between 2D and 3D?

- Generating objects in scenes using cut and paste in 3D?

4 Progress

Extracted relevant annotations and labels form SUNRGBD dataset Using 2D annotations, only take a subset of objects and crop full image to just the object area. Train CNN on these cropped images to have a baseline 2D classifier.

- show some examples of the dataset with annotations

- show examples of cropped inputs

- discuss how this will help with RCNN

- how RCNN will be used for the frustum method

References

[1] Vassileios Balntas, Andreas Doumanoglou, Caner Sahin, Juil Sock, Rigas Kouskouridas, and Tae-Kyun Kim. Pose Guided RGBD Feature Learning for 3D Object Pose Estimation. *ICCV*, 2017.

[2] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Unsupervised Learning of Important Objects from First-Person Videos. *ICCV*, 2017.

[3] Anders Glent Buch, Lilita Kiforenko, and Dirk Kraft. Rotational Subgroup Voting and Pose Clustering for Robust 3D Object Recognition. *ICCV*, 2017.

[4] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *PMLR*, 2013.

[5] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. *ICCV*, 2017.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, jun 2012.

[8] Ross Girshick. Fast R-CNN. *ICCV*, apr 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, nov 2014.

[10] Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurélien Plyer, and David Filliat. SnapNet-R: Consistent 3D Multi-View Semantic Labeling for Robotics. *ICCV*, 2017.

[11] Ankur Handa, Viorica PatrAucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding Real World Indoor Scenes With Synthetic Data. *CVPR*, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *ICCV*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CVPR*, dec 2015.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.

[15] Jean Lahoud and Bernard Ghanem. 2D-Driven 3D Object Detection in RGB-D Images. *ICCV*, 2017.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. *CoRR*, dec 2017.

[18] Wenjie Luo and Raquel Urtasun. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net. *CVPR*, 2018.

[19] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, oct 2010.

[20] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum PointNets for 3D Object Detection from RGB-D Data. *CVPR*, 2018.

[21] Fracture Reality. Company website. fracturereality.io.

[22] Fracture Reality. Hololens — airport command and control centre. https://tinyurl.com/yau4cj88.

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *CVPR*, jun 2016.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*, jun 2015.

[25] Zhile Ren and Erik B Sudderth. 3D Object Detection with Latent Support Surfaces. *CVPR*, 2018.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[27] Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-YOLO: An Euler-Region-Proposal for Real-time 3D Object Detection on Point Clouds. *CoRR*, 2018.

[28] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, sep 2014.

[29] Thomas Solund, Anders Glent Buch, Norbert Kruger, and Henrik Aanas. A Large-Scale 3D Object Recognition Dataset. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 73–82. IEEE, oct 2016.

[30] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. *CVPR*, 2015.

[31] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI*, feb 2016.

[32] Christian Szegedy, Wei Liu, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CVPR*, 2015.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CVPR*, dec 2015.

[34] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-Time Seamless Single Shot 6D Object Pose Prediction. *CVPR*, 2018.

[35] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82. IEEE, mar 2014.

[36] Bin Xu and Zhenzhong Chen. Multi-Level Fusion based 3D Object Detection from Monocular Images. *CVPR*, 2018.

[37] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. *CVPR*, 2018.

[38] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *NIPS*, nov 2014.

[39] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. *CVPR*, 2018.

[40] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. *CVPR*, jul 2018.