Evaluation of 3D vision systems for detection of small objects in agricultural environments

Justin Le Louedec[1][a], Bo Li[2][b] and Grzegorz Cielniak [1][c]
[1]*Lincoln Centre for Autonomous Systems, University of Lincoln, UK*
[2]*University of West England, UK*
{*jlelouedec,gcielniak*}*@lincoln.ac.uk*
bo2.li@uwe.ac.uk

Keywords: Machine Vision for Agriculture, Machine Learning, 3D Sensing, 3D Vision

Abstract: 3D information provides unique information about shape, localisation and relations between objects, not found in standard 2D images. This information would be very beneficial in a large number of applications in agriculture such as fruit picking, yield monitoring, forecasting and phenotyping. In this paper, we conducted a study on the application of modern 3D sensing technology together with the state-of-the-art machine learning algorithms for segmentation and detection of strawberries growing in real farms. We evaluate the performance of two state-of-the-art 3D sensing technologies and showcase the differences between 2D and 3D networks trained on the images and point clouds of strawberry plants and fruit. Our study highlights limitations of the current 3D vision systems for detection of small objects in outdoor applications and sets out foundations for future work on 3D perception for challenging outdoor applications such as agriculture.

1 INTRODUCTION

Thanks to recent advances in 3D sensing technology and rapidly growing data-driven algorithms, the 3D vision has attracted considerable attention in recent years. Compared to 2D images, 3D information provides additional depth cues critical for estimating precise location and assessing shape properties of various objects in the environment. So far, the main focus in the 3D vision community has been centred around benchmark datasets captured in controlled environments with large and rigid objects and fairly stable lighting conditions (e.g. (Dai et al., 2017; Armeni et al., 2017)). There is, however, a strong case for deploying such systems in real-life scenarios and currently it is not clear how well the current state of the art in 3D vision translates into the challenging situations posed by applications such as in agriculture. As part of our research project, we propose a study on the application of modern 3D sensing technology together with the state-of-the-art machine learning algorithms for segmentation and detection of strawberries growing in real farms. The precise information about strawberry fruit location and shape description have a range of applications such as yield monitoring and forecasting, phenotyping or robotic picking. The challenges posed by such a scenario include variable lighting conditions, reflections, occlusions, non-rigid structure of the strawberry plants and relatively small size of the fruit. Since, the current 3D sensing technology has not been deployed widely in such scenarios and most of the modern machine learning algorithms were designed and trained specifically for large and rigid objects, our study aims to assess the usefulness of the sensing and learning methodology for the proposed application.

In particular, our paper provides the following contributions:

- assessment of two competing 3D sensing technologies (i.e. projected stereo IR pattern and Time-of-Flight sensors) for the problem of detection of small objects (i.e. fruit) in outdoor environments;
- assessment of the current state of the art in 3D machine learning and of the required modifications enabling their use for the proposed application;
- comparison of the accuracy for 2D image-based and full 3D approaches;
- validation of the sensing and learning pipeline on

[a] https://orcid.org/0000-0000-0000-0000
[b] https://orcid.org/0000-0000-0000-0000
[c] https://orcid.org/0000-0002-6299-8465

data collected from real strawberry farms.

This paper is organised as follows: we start with a brief overview of the related work in 3D sensing and vision followed by the description of our methodology (Sec. 3) and its experimental evaluation in Section 4. The paper is concluded in Section 5.

2 BACKGROUND

2.1 3D vision in agriculture

In agricultural applications, 3D information can provide important object characteristics such as crop size, shape or location. The most common approach to recognise such objects is based on a combination of 2D images for crop segmentation and detection and 3D information for augmenting the shape and location information. For example, (Lehnert et al., 2018; Lehnert et al., 2017) describe a perception system for harvesting sweet peppers. After scanning an reconstructing the scene using a robotic arm, they use colour information in the point cloud to detect the pepper. They then use the 3D projection of the segmented peduncle to estimate a pose and the optimal grasping/cutting point. (Barnea et al., 2016) also present a perception system for pepper harvesting but use a colour agnostic method to detect fruits using a combined colour and depth information (i.e. provided by RGBD cameras). By using highlights in the image, and 3D plane-reflective symmetries, their system is able to detect pepper fruit based on shape, in heavily occluded situations. (Yoshida et al., 2018) use RGBD cameras and a two resolution voxelisation of the 3D space to detect tomatoes peduncles and the optimal cutting point for harvesting. They first identify regions corresponding to tomatoes whilst using the dense voxelisation on the selected regions and to establish the optimal cutting points on the peduncle. The DeepFruit system (Bargoti and Underwood, 2016) uses Deep Learning networks for the detection of fruits in images taken from the orchards. The system achieves good detection accuracy but by only regressing bounding boxes over instances, it does not consider spatial information and therefore is prone to missed detections when the fruits overlap.

2.2 Deep learning for 3D information

3D information is represented by so-called point clouds - a collection of unconnected and unordered points in 3D space which originate from sensors such as 3D laser range finders or Time-of-Flight cameras (i.e. RGBD cameras) and can be augmented by additional information such as colour, reflectivity, etc. The core of machine learning methods using deep networks is applied to standard images and therefore is based on 2D convolutions. This can be realised effectively in 2D but poses problems in 3D. First of all, the convolution operation requires a discretisation of space into so-called voxels, which is typically associated with some loss of information and large memory requirements. Secondly, 3D convolution operations are computationally expensive which renders them unusable in most of the real-life scenarios. Therefore, there is a big interest in developing methods which can cope with these challenges. Point cloud processing using Deep Learning has been revolutionised by PointNet (Qi et al., 2017a) and subsequently by its improved variant PointNet++ (Qi et al., 2017b). The PointNet architecture can be directly applied to a point cloud, through a prior segmentation/grouping of points in space using clustering algorithms such as K-Nearest Neighbours (KNN) or ball query. Convolutions are not applied to the organised representation of space, but rather to the features extracted from the clustered regions, which can be performed efficiently. PointNet++ is using a multi-scale approach for partitioning point clouds and relies on two main layers used for encoding and decoding information. The first layer (Set Abstraction (SA)) extracts features from points by considering their neighbourhood defined by a radius. The second layer (Feature Propagation (FP)) interpolates the features and learns their decoding into the dimension of the previous SA layer, up until the same size as the input point cloud. For the classification task, the latent representation of the point cloud features after the succession of SA layers is used and passed through a multi-layer perceptron to predict the classification of individual points. For segmentation and other task requiring features associated for each point of the point cloud, each SA layer is associated with an FP layer in charge of decoding the resulting features up to the input size. A multi-layer perceptron is then used to predict per point class.

The basic PointNet architecture has been used to develop further improvements. For example, PointSIFT (Jiang et al., 2018) uses SIFT-like features extracted from the immediate neighbourhood of a point in 8 directions. On the other hand, PointCNN (Li et al., 2018) is applying a convolution-like operator on points grouped and transformed based on predefined properties, ordering points and transforming their features into a latent and canonical space, followed by a standard convolution operation. However, most of the described improvements, whilst increasing the network's discriminatory abilities, suffer from

higher computational demands and do not scale well to real-time applications in realistic scenarios.

2.3 3D datasets

The majority of the existing 3D machine learning algorithms were validated on datasets acquired in indoor environments. The examples of such datasets include ScanNet (Dai et al., 2017) with scans obtained using surface reconstruction and crowd-sourced labelling for annotation. Another example includes Stanford 2D-3D-Semantics Dataset(Armeni et al., 2017) which was created with an RGBD camera. The S3DIS offers a very complete scene description, additional colour information, depth, 2D segmentation, meshes and 3D segmentation with extracted normals. There is a very limited amount of publicly available 3D datasets in outdoor spaces. Notable examples include semantic3D.net (Hackel et al., 2017) which is a very large scale dataset (4 billion points per scene) of various outdoor locations. Another example includes the KITTI (Geiger et al., 2012) dataset, captured aboard a driving car in various streets, providing depth and colour information with annotated bounding boxes and instance segmentation. All the abovementioned datasets feature large, rigid objects which are not typical for agricultural environments.

3 METHODOLOGY

3.1 3D sensing

3D sensing is based on capturing the depth or distance from the camera to each point in the scene. Capturing devices for outdoor use can be divided into three main categories: stereo cameras, Time-of-Flight (ToF) devices and Lidar range finders. Stereo sensors are based on capturing two images from two image sensors apart from each other and matching their features to create a depth map based on epipolar lines between the two sensors. In the case of wrongly matched points or a lack of similarity, surfaces reconstructed in this way, are often distorted or flat with blended edges and objects. This is especially evident with very small objects. Alternative sensing solutions use light wavelengths outside of the visible spectrum (e.g. infrared) which are less prone to changing lighting conditions and more robust matching points. Examples of commercial stereo cameras used in research are ZED cameras and Intel RealSense (Georg Halmetschlager-Funek and Vincze, 2018). Time-of-Flight devices are based on light beams which are being projected into the scene and reflected back to the sensor. The depth is estimated from the time taken for the light to come back. This technology results in more precise depth measurements, but more prone to noise caused by reflective objects. The Microsoft Kinect One (i.e. v2) (Georg Halmetschlager-Funek and Vincze, 2018) and the Pico Zense (Noraky and Sze, 2018) are a good example of recent innovations in this technology. Lidar is a particular example where the beam of light is replaced by a laser pulsed at the scene. We do not consider Lidar technology in our work, however, since its intrinsic properties and resolution are not suited for the detection of small objects in occluded scenarios such as strawberries. (Kazmi et al., 2014) offers a comprehensive study of both sensors applied in different situations.

In this paper, we compare stereo and Time-of-Flight sensing technologies based on their performance in sensing of strawberries in their natural growing conditions. The two selected cameras were the Intel Realsense D435 (IR stereo) and the Pico Zense (ToF). The detailed experimental comparison of these two technologies on 3D data of strawberries collected from their natural growing environment is presented in Section 4.3.

3.2 3D Vision

In our study, we are interested in the feasibility of modern 3D sensing and machine learning for a problem of detecting strawberries in their natural environment. To this end, we select a reliable, robust and popular (see recent applications (Wang et al., 2017; Pham et al., 2019; Yang et al., 2019; Wang et al., 2019)) deep learning architecture PointNet++ (Qi et al., 2017b). The PointNet++ provides segmentation results, i.e. per point classification which, if successful, would enable instance detection of individual strawberries. In this work, however, we assess both sensing technology on segmentation problem only.

3.2.1 Data pre-processing

Before the 3D data can be used by the network, it needs to be pre-processed so that the point clouds provided by both types of sensors are of similar characteristics. The point clouds generated by the sensors are already augmented with registered colour information (RGB). The sensors generate point clouds of different density and number: $\sim$920K points for the stereo camera and $\sim$230K points for the ToF device. Since the sensors are placed in a similar distance to the strawberry plants (i.e.$\sim$60cm) the point clouds can be downsampled to match the spatial resolution of around 3 mm, which is a limiting factor constrained by the depth resolution of both sensors. This results

in point clouds of ∼25K points which are still too large for the PointNet architecture and therefore we partition them into smaller subsets following a procedure employed in (Qi et al., 2017b). The procedure is using a sliding box over the point cloud using K-Nearest Neighbors (KNN) algorithm to guarantee the same size input (8000 points) and allowing us to train the algorithms using mini-batches. During the prediction phase, we use a maximum vote strategy for points belonging to multiple blocks.

3.2.2 PointNet++

The original implementation of PointNet++ requires modifications to make it suitable for the scale and resolution of our problem. PointNet++ offers two approaches for segmentation: the multi-scale approach with one-hot encoding for the classification and a single scale of grouping with an increased number of layers and complexity of the feature space. The initial experimentation on a subset of our data identified minimal differences in accuracy provided by these two methods. The second one, however, requires significantly lower execution times and memory consumption and therefore it was selected for further study. Considering the size of our inputs and the size of strawberries, we augmented the number of layers for the decoding part of the network and adding their counterpart in the decoder. This provides several benefits. Firstly, the point cloud is subdivided progressively into broader versions, with fine-scale features learnt at the start of the training and global features learnt towards the end, which compensates for the lack of an intra-layer multi-scale component. Secondly, we also adapted the radius for the ball-query for points which can be adjusted for the selected density/resolution of points. This second modification is directly linked to the network's output and loss function used. In our implementation, we decided to predict the class of each point (fruit/background) rather than using specific class (ripe or unripe fruits and background). This enabled a better convergence of the learning process and a stronger emphasis on the shape features rather than colour. Following notation introduced in (Qi et al., 2017b), our implementation features a topology of the encoder and decoder as summarised in Table 1 and 2, which summarise the number of points per layer (*#points*), radius for the ball-query (*radius*), feature length (*features*) and configuration of the multi-layer perceptron (*mlps*).

Table 1: Configuration of an encoder used in our PointNet++ implementation.

Layer Type	#points	radius	mlps
SA	4096	0.1	[16, 16, 32]
SA	2048	0.1	[32,32,32]
SA	1024	0.1	[32,32,64]
SA	256	0.2	[64,64,128]
SA	64	0.4	[128,128,256]
SA	16	0.8	[256,256,512]

Table 2: Configuration of a decoder used in our PointNet++ implementation.

Layer Type	features
FP	256,256
FP	256,256
FP	256,128
FP	128,128,128
FP	128,128,64
FP	128,128,64
MLP	[64,128]
MLP	[128,2]

3.2.3 2D vs 3D segmentation

To evaluate the usefulness of the 3D information for segmentation of small objects, we also select a standard 2D image-based architecture for our comparisons. The Convolutional Neural Networks (CNN)s have proven to be very effective for the object/scene segmentation tasks although lacking depth information. For this purpose, we select a state-of-the-art popular network architecture called SegNet (Badrinarayanan et al., 2015), which is very similar to PointNet++. SegNet is a feed-forward network using, similarly to PointNet++, the auto-encoder principle, encoding the feature space down to a specific size (e.g. 512) before decoding it back to the original size of the input image. For each pixel, we can either predict a score for each class (strawberry/background) followed by a softmax function to get the predicted class, or predict a probability for each class. Each of the convolutions is followed by a normalisation and each convolution block by a max-pooling operation. The decoder uses a max-unpool layer as upsampling step and transposed convolutions.

4 EXPERIMENTS

4.1 Data collection

To support the main goal of our application, which is applying a modern 3D vision system for the detection and localisation of strawberry fruit, we collected a dataset from the real environment. To that end, we have deployed our data acquisition system at a mini

Figure 1: The strawberry farm, with a robot roaming in the tabletop rows collecting data (left). The sensor set-up used for data collection (right).

version of a real strawberry farm, located at Riseholme campus of the University of Lincoln. The farm features two polytunnels of 6 tabletop rows, 24 meter long with an industrial variety of strawberries (everbearer Driscoll's Amesti) as depicted in Fig. 1. The data capture setup featuring the Realsense and Pico Zense sensors was mounted on an agricultural robot Thorvald (Grimstad and From, 2017). The robot autonomously navigated the polytunnel rows, stopping every 20 cm to collect a snapshot from both views (see Figure 1). The capturing session took place in October 2019 and resulted in colour images and point clouds representing different growth stage of plants and fruit. The datasets were then manually annotated to indicate location of strawberry fruits resulting in 139 labelled point clouds with around 1900 instances of ripe strawberries for ToF data and 64 point clouds for around 1000 instances for stereo data (see Table 3).

Table 3: The summary of datasets collected.

sensor	stereo	ToF
# point clouds	64	139
resolution	1280x720	1280x720
range	20cm-65m	20cm-70cm
# instances	$\sim$ 1000	$\sim$1900
% strawberry points	$\sim$6%	$\sim$ 6%

4.2 Evaluation Methodology

To evaluate our trained models, we use standard semantic segmentation metrics such as Accuracy and mean Intersection over Union (mean IoU) and also Kappa-Cohen (Cohen, 1960) which is particularly useful when unbalanced number of class instances is used - in our case, background represents the majority of points when compared to strawberries.

The Accuracy is the most used metrics for majority of machine learning systems and measures how accurate the prediction is compared to the ground truth, without taking in consideration the balance of classes and positives/negatives:

$$Acc = \frac{TP+TN}{P+N}. \quad (1)$$

Mean IoU is the overlap of the output predicted by the algorithm with the ground truth and averaged for every class and samples:

$$IoU = \frac{TP}{(TP+FP+FN)}. \quad (2)$$

Kappa-Cohen coefficient is particularly useful for unbalanced data, where one class is more represented than the others. This measure provides a better assessment of the real discriminatory power of the classifier and takes observed and expected accuracies into account:

$$K = \frac{(Acc_{obs} - Acc_{exp})}{1 - Acc_{exp}}. \quad (3)$$

The Observed Accuracy is the number of instances correctly classified through point cloud, and the Expected Accuracy is what any random classifier should be expected to achieve over the point cloud.

We also use precision-recall curves to evaluate the performance of the trained classifiers as in (Everingham et al., 2010). Precision represents a ratio of $\frac{TP}{TP+FP}$ whilst recall is a ratio of $\frac{TP}{TP+FN}$. The precision and recall values are computed over a range of confidence score thresholds of the classifier.

4.3 Results

4.3.1 Quality of Acquisition

We compare the data acquired using Intel Realsense D435 and Pico Zense cameras which both provide RGBD information. We do not use the post-processing filters offered by the Realsense device, as they were proven to be not reliable outdoors with sensitive light and exposition settings. The only built-in enhancement enabled is the spatial filter which is also used in the ToF sensor.

The following qualitative assessment is based on observing the depth maps captured as shown in Figure 2. The stereo camera works using infrared spectrum (third image from the top in Figure 2, the depth obtained is very sensitive to variations in exposure and stereo-matching between both infrared images captured. The stereo sensor also captures visible information as far as possible, reducing the amount of detailed information for small objects and their surfaces. Overall the data captured using stereo camera provides inferior depth information compared to the ToF, and shape information for strawberries is degraded. This is mostly due to the lack of features for

stereo-matching between the two IR cameras. One can also notice the absence of the IR pattern projected by the camera. This pattern is supposed to improve depth with more reliable features to match between images, but it is here completely dispersed by the sun's natural infrared spectrum. The Time-of-Flight

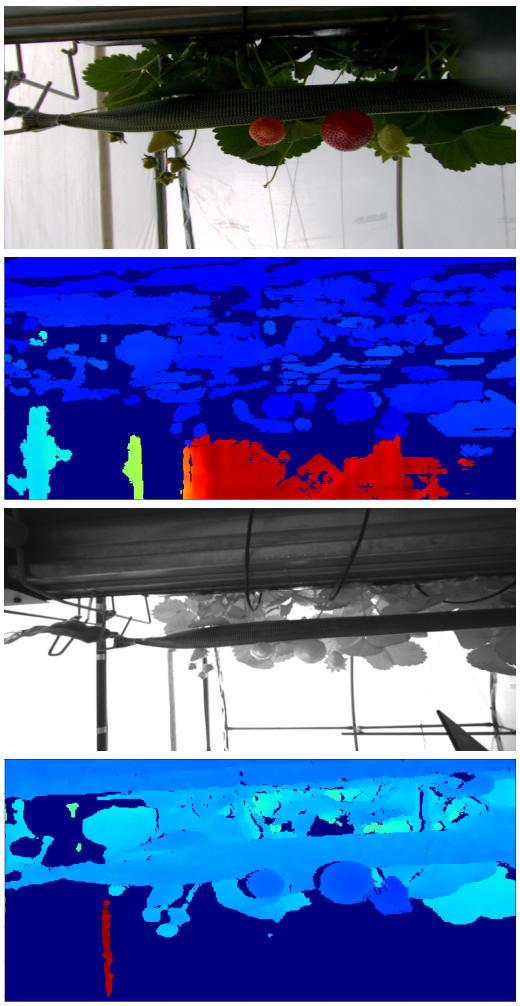

Figure 2: Comparison of depth captured using Stereo and Time-of-Flight technologies

camera, on the other hand, can be programmed to capture information only in a given range depending on the application. In our data collection, the camera was set to near-range setting, improving quality of depth for short ranges (three ranges are usable: Near, Mid, Far), which suits our application, since the interesting information is found up to 70 cm away from the camera. Despite this fact, the quality of ToF depth maps is visually better. This sensor offers a better coverage of the depth information. The only limitation comes from the light spectrum used, which corresponds to some part of the solar infrared spectrum. Flat surfaces, which are the most reflective, appear slightly

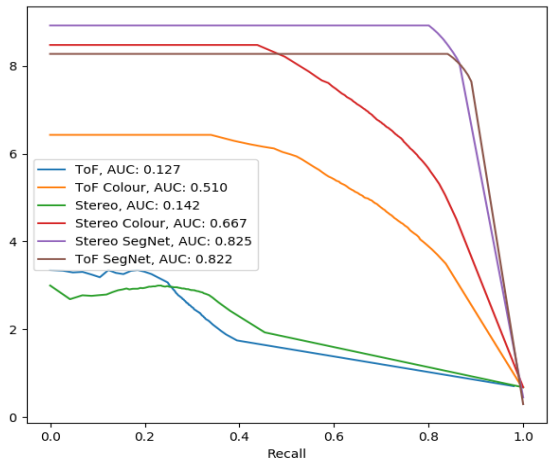

Figure 3: Precision-recall curves for the different networks indicating also area under the curve (AUC).

deformed and in some cases with deformations of the scale matching small strawberries. These qualitative findings suggest that Time-of-Flight technology suits better outdoor environments and is especially beneficial for the detection and shape analysis of small objects such as fruits. Concerning the quality of the RGB images from both cameras, the stereo camera provides clearer and higher quality images overall in an outdoor context.

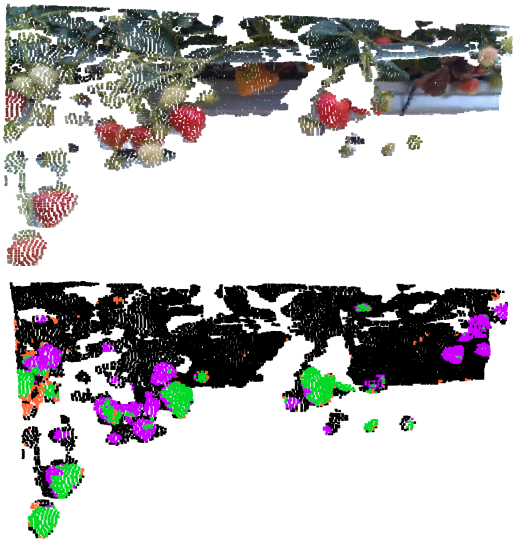

Figure 4: The segmentation results for $PNet_{colour}$ trained on data from the stereo camera: the original point cloud (top), segmentation results (bottom). The colours indicate: TP in green, FP in orange, FN in purple and TN in black.

4.3.2 Network Comparison

To compare both sensors directly, we use the captured data for training the selected machine learning

Table 4: Comparison of different networks trained on data for stereo and ToF sensors.

Model	Camera	Dimension	Accuracy	Kappa Cohen	mean IoU
SegNet	stereo	2D	98.5%	0.71	0.83
SegNet	ToF	2D	99.0%	0.79	0.67
PNet	stereo	3D	91.1%	0.43	0.14
PNet	ToF	3D	90.0%	0.38	0.15
PNet$_{colour}$	stereo	3D	95.4%	0.66	0.48
PNet$_{colour}$	ToF	3D	92.5%	0.54	0.39

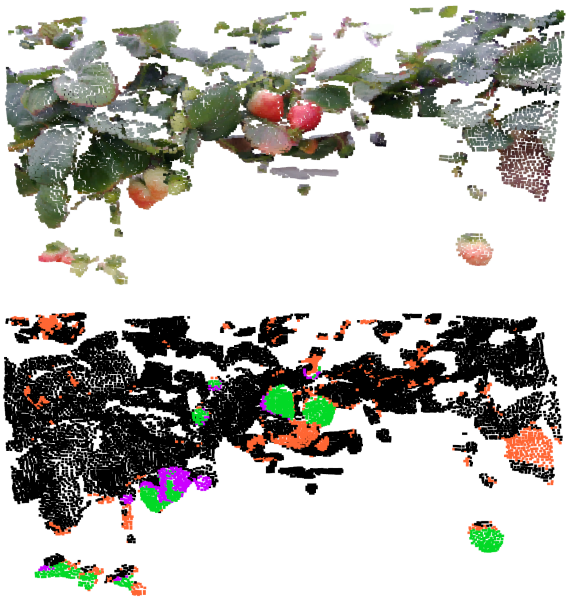

Figure 5: The segmentation results for $PNet_{colour}$ trained on data from the ToF camera: the original point cloud (top), segmentation results (bottom). The colours indicate: TP in green, FP in orange, FN in purple and TN in black.

networks (see Sec. 3) for each sensor separately. In addition, we use the following configurations of the networks and type of input data: SegNet on colour images, PointNet++ on 3D point clouds only (*PNet*) and PointNet++ with additional colour information ($PNet_{colour}$). The results are presented in Table 1 whilst the precision-recall curves in Figure 3. There is a clear difference between networks trained with and without colour information and between networks trained on ToF and stereo datasets with performances significantly improved for both datasets. This can be explained by a greater difference using colour space between strawberries and background than using only shape information. The difference in results for different sensors seems to be amplified when colour information is used and would come from the unreliable readings from large surfaces and shapes with ToF cameras leading to many false positives (FP). The precision-recall curves confirm these findings. The high amount of FP is negatively affecting the networks trained with 3D information only but improved significantly for networks trained in colour. With a very low area under the precision-recall curve, *PNet* is the worst performing classifier than $PNet_{colour}$ on our datasets. A 2D network SegNet performs significantly better than any of the 3D variants for both datasets. To illustrate these findings, we provide example outputs from $PNet_{colour}$ illustrating the quality of segmentation for both sensors. The stereo dataset (Fig.4) is characterised by more omissions of strawberries with a high number of FN, but less false detections. The main red and distinct strawberries are however well segmented. The ToF example (Fig. 5), all the strawberries are mostly segmented out and there are few omissions. However, there is a high number of false positives, especially on the brown/red leaves of the scene with round shape. These examples provide additional support for findings based on the numerical values in Table 4.

The superior performance of detectors based on 2D CNNs for our application can be associated with the structured nature of 2D images, maturity of the developed networks and also low quality of the depth data. Also, through post-processing, sufficient spatial information can be retrieved using the depth map and the 2D segmentation mask, making these algorithms preferable for our application at the time being. The presented 3D approaches, however, offer an advantage in direct localisation of the objects, although their localisation accuracy is a subject of future work. The real-time suitability of the 3D methods is also promising, achieving 5 FPS (each frame a point cloud of $\sim$64k points) compared to 13 FPS (each frame an image of 1280×720 px) for SegNet.

5 CONCLUSIONS

Capturing 3D data and processing it for different tasks such as detection, segmentation or classification is a challenging task especially in the agricultural context presented in this paper. Our study evaluated two 3D sensing technologies for that purpose and com-

pared 3D and 2D variants of state-of-the-art neural networks trained on the data collected from a strawberry growing farm. These results show encouraging performance but also allow us to highlight the limitations of current technologies and algorithms. Time-of-Flight technology, despite its superior quality of point clouds and shape information, struggles with reflective surfaces resulting in a large number of false detections, while stereo technology, lacking detail in acquired depth, fails to detect numerous fruits. Traditional 2D image-based convolutional neural networks still outperform the 3D networks for the task of fruit segmentation and therefore are more suited for this task. This work can be treated as a baseline for future work on 3D information for outdoor applications such as robotic fruit picking and should encourage researchers to pursue more experimentation in such difficult to counteract limitations found in the paper and bridge the gap with state-of-the-art techniques in perception for 2D information.

REFERENCES

Armeni, I., Sax, A., Zamir, A. R., and Savarese, S. (2017). Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv e-prints*.

Bargoti, S. and Underwood, J. (2016). Deep Fruit Detection in Orchards. *arXiv e-prints*.

Barnea, E., Mairon, R., and Ben-Shahar, O. (2016). Colour-agnostic shape-based 3d fruit detection for crop harvesting robots. *Biosystems Engineering*, 146:57 – 70. Special Issue: Advances in Robotic Agriculture for Crops.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*.

Everingham, M., Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision*, 88(2):303–338.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Georg Halmetschlager-Funek, Markus Suchi, M. K. and Vincze, M. (2018). An empirical evaluation of ten depth cameras. *IEEE Robotics and automation magazine*.

Grimstad, L. and From, P. J. (2017). The Thorvald II Agricultural Robotic System. *Robotics*, 6(4).

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., and Pollefeys, M. (2017). SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-1-W1, pages 91–98.

Jiang, M., Wu, Y., Zhao, T., Zhao, Z., and Lu, C. (2018). PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation. *arXiv e-prints*.

Kazmi, W., Foix, S., Alenyà, G., and Andersen, H. J. (2014). Indoor and outdoor depth imaging of leaves with time-of-flight and stereo vision sensors: Analysis and comparison. *ISPRS Journal of Photogrammetry and Remote Sensing*, 88:128–146.

Lehnert, C., English, A., McCool, C., Tow, A., and Perez, T. (2017). Autonomous Sweet Pepper Harvesting for Protected Cropping Systems. *arXiv e-prints*.

Lehnert, C., McCool, C., Sa, I., and Perez, T. (2018). A Sweet Pepper Harvesting Robot for Protected Cropping Environments. *arXiv e-prints*.

Li, Y., Bu, R., Sun, M., and Chen, B. (2018). PointCNN: Convolution On X-Transformed Points. *arXiv preprint arXiv:1801.07791*.

Noraky, J. and Sze, V. (2018). Low Power Depth Estimation of Rigid Objects for Time-of-Flight Imaging. *arXiv e-prints*.

Pham, Q.-H., Thanh Nguyen, D., Hua, B.-S., Roig, G., and Yeung, S.-K. (2019). JSIS3D: Joint Semantic-Instance Segmentation of 3D Point Clouds with Multi-Task Pointwise Networks and Multi-Value Conditional Random Fields. *arXiv e-prints*.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). PointNet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108.

Wang, W., Yu, R., Huang, Q., and Neumann, U. (2017). SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. *arXiv e-prints*.

Wang, X., Liu, S., Shen, X., Shen, C., and Jia, J. (2019). Associatively Segmenting Instances and Semantics in Point Clouds. *arXiv e-prints*.

Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., and Trigoni, N. (2019). Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. *arXiv e-prints*.

Yoshida, T., Fukao, T., , and Hasegawa, T. (2018). Fast Detection of Tomato Peduncle Using Point Cloud with a Harvesting Robot. *Journal of Robotics and Mechatronics*, 30(2):180–186.