Towards Autonomous Plant Production using Fully Convolutional Neural Networks

Peter Christiansen[a,*], René Sørensen[a], Søren Skovsen[a], Claes D. Jæger[b], Rasmus Nyholm Jørgensen[a], Henrik Karstoft[a] and Kim Arild Steen[b]

[a] Department of Engineering, Aarhus University, Aarhus, Denmark

[b] AgroIntelli, Aarhus, Denmark

* Corresponding author. Email: pech@eng.au.dk

Abstract

In order for autonomous agricultural vehicles to operate cost efficiently and to be able to operate unsupervised, they must adhere to current EU legislation, and be able to perform automatic real-time path-planning, risk detection and obstacle avoidance. In this context, sensor technologies must be utilized to perceive surroundings and allow autonomous machines to act accordingly.

This paper investigates the perception capabilities of a deep learning approach called "fully convolutional neural network for semantic segmentation" (pixel level classification) in agriculture using an rgb camera sensor. Training a network for semantic segmentation requires the comprehensive task of providing whole scene per-pixel labelling on a large data set. To avoid the task of creating per-pixel labelled data we investigate using a network trained on two already existing databases (ImageNet and Pascal-context) with mostly non-agricultural specific images and classes. A pre-trained network performs pixel wise classification on 59 classes and by remapping the 59 classes to agricultural specific classes (e.g. sky, field, shelterbelts, animal, human and obstacles), we are able to test the network's ability to generalize to agriculture in two case studies: grass mowing and row crop operations.

Based on a small set of 10 per-pixel labelled test images, we show that the network is able to generalize to a grass mowing use cases with an pixel classification accuracy of 95.25%. In the row crop case, the network is less reliable with a classification accuracy of 70.54%. By showing detections of state-of-the-art object detection algorithms (a pedestrian detector and a deep learning object detector), a small qualitative comparison between object detection methods and the semantic segmentation algorithm is made.

Keywords: Deep learning, Autonomous Tractors, Semantic Segmentation, Per-Pixel labelling, Agriculture

1. Introduction

In order for autonomous agricultural vehicles to operate cost efficiently and to be able to operate unsupervised, they must adhere to current EU legislation, and be able to perform automatic real-time path-planning, risk detection and obstacle avoidance. In this context, sensor technologies must be utilized to perceive surroundings and allow autonomous machines to act accordingly.

In the automotive industry, a range of companies (Google, Ford, Uber, Tesla etc.) have demonstrated autonomous vehicles in both prototype and commercial products. Autonomous vehicles in agriculture uses sensor technologies and algorithms found in the automotive industry. However, certain challenges in perception (and path-planning) is specific to agriculture.

For perception, Google Car is highly dependent on both a very detailed static 3D map and under operation expensive laser scanners (Velodyne LiDAR) to get an immediate measurement of its surroundings. A static 3D map is not as feasible in agriculture as the crop is under constant transformation and laser scanners are currently too expensive for farmers. Secondly, depth sensor will hardly detect obstacles that are not protruding the crop surface such as kids, lying humans, hydrants, wells and animals.

Conventional automotive companies evolve from semi- to fully-autonomous by adding affordable features and sensor. Relying on detailed 3D maps is problematic as the car must operate on roads without maps and rely on sensors affordable to consumers and car manufacturers. Mobileye is a leading company in delivering real-time camera-based solution for automotive companies including Audi, Ford and Tesla. However, solutions by Mobileye are not all suited, accessible or trained for agriculture, and algorithms are mostly unpublished. The unpublished results by Mobileye and the limited access for non-automotive industries, makes the actual accuracy performance of Mobileye solutions unclear to researchers, and also if the perception capabilities are comparable to recent deep learning perception algorithms.

Deep learning is an emerging field in Artificial Intelligence/Machine Learning that recently moved the boundary of a computer intelligence and perception. In 2015 deep learning was able to reach human performance in image classification (He, Zhang, and Sun 2015) and speech recognition (Amodei et al. 2015). Especially Convolutional Neural Networks

(CNN) (LeCun et al. 1998) have recently outperformed traditional image recognition methods used for traffic sign recognition(Ciresan, Meier, and Schmidhuber 2012), face detection (Farfade, Saberian, and Li 2015), face recognition (Taigman et al. 2014), image classification (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He, Zhang, and Sun 2015), general object detection (Ren et al. 2015; Redmon et al. 2016; Girshick et al. 2014; He et al. 2014) and semantic segmentation (Long, Shelhamer, and Darrell 8 Mar, 2015; Chen et al. 2015; Torr 2014). Fast and high accuracy CNN-based object detection algorithms have recently been published and open-sourced (Ren et al. 2015; Redmon et al. 2016). Especially YOLO (Redmon et al. 2016) is able to process images at real-time speeds using a high-end GPU on the 20 object types from Pascal VOC (Everingham, Eslami, and Gool 2013). A drawback of object detection algorithms in the context of agriculture, is that not all elements or "stuff" are precisely delimited with a bounding box such as shelterbelt, ground, crop and water. Secondly, object detection algorithms are challenged in agriculture, where obstacles are likely to be heavily or partly occluded by the crop.

Semantic segmentation is an image recognition method that classifies each pixel in the image. In (Long, Shelhamer, and Darrell 8 Mar, 2015) a CNN is converted and modified to a Fully Convolutional Neural Network for Semantic Segmentation (FCS). In (Chen et al. 2015; Torr 2014) Conditional Random Fields are appended to a modified CNN to improve accuracy, though with additional computational cost.

In relation to object detection algorithms, semantic segmentation performs pixel classification and is therefore more suited to detect both obstacles and "stuff" and is presumably less challenged by occlusion.

A drawback of training a semantic segmentation model sufficiently is the requirement for a lot of data with per-pixel level annotations. Transferring the algorithm to agriculture is therefore obstructed by the comprehensive task of making per-pixel level annotations on new data.

This work is a preliminary study demonstrating the perception capabilities of semantic segmentation in agricultural by remapping the 59 predictions from the model in (Long, Shelhamer, and Darrell 8 Mar, 2015) to 11 agriculture specific classes (animal, building, field, ground, obstacle, person, shelterbelt, sky, vehicle, water and unknown). The perception capability is presented by two measures: the classification per-pixel accuracy on 10 test images and a qualitative comparison (by examples) of semantic segmentation and object detection.

The purpose of the publication is to emphasize the powerful perception capability of deep learning semantic segmentation. Traditionally, image recognition algorithms fail to generalize on data not similar to the training data. We show that the comprehensive task of creating training data can in some cases be avoided and that a network trained on general images (PASCAL-Context) is able to generalize to a different context.

2. Materials and Methods

In this section the data used for training and testing FCNN is first presented followed by a description of FCN and the simple remapping approach.

2.1. Data - For training

ImageNet (Berg and Deng 2015) is a image recognition benchmark for image classification, object localization and object detection. The ImageNet benchmark have since 2012 pushed the performance of CNNs and provided a common ground for leading research teams to compete (Google, Microsoft, Baidu). Especially the image classification challenge with an incredible amount of 1.400.000 images with image notation of 1000 different object types.

PASCAL Visual Object Classes (VOC) (Everingham, Eslami, and Gool 2013) is another image recognition benchmark. The benchmark includes a semantic segmentation competition on 20 object classes. In PASCAL-Context (Mottaghi et al. 2014) six in-house annotators have used three months to extend PASCAL VOC with whole scene annotation on 10,103 images, extending the number of object classes from 20 to 407. Figure 1 shows the difference between PASCAL VOC and PASCAL-Context. Note especially annotations of "stuff" such as road, grass and trees is also provided.

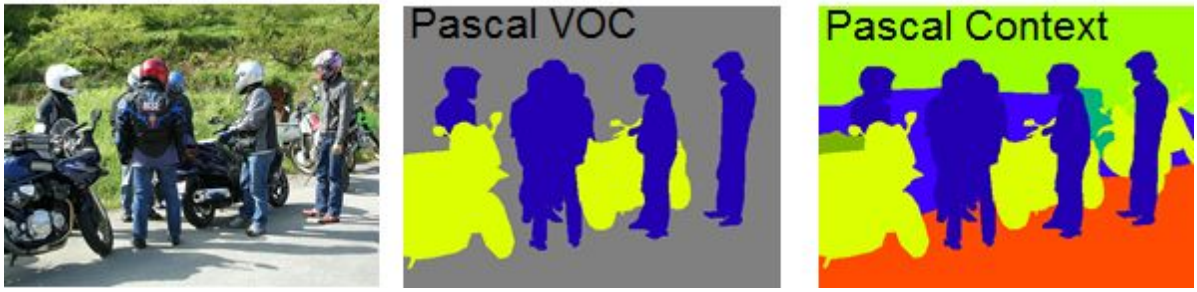

Figure 1. Shows the much denser pixel-annotations provided in PASCAL-Context.

2.2. Data - For testing

A rolling shutter Logitech C920 webcam with a resolution of 1920x1080 and a framerate of 30Hz have been used to record image data. The webcam is placed on a sensor platform including other sensors as described in (Christiansen et al. 2015a; Christiansen et al. 2015b). The sensor platform includes a metal frame with a standard A-frame in the bottom. The A-frame is easily mountable to a tractor and places the camera roughly 2.0m from the ground.

The grass mowing use case is recorded in a 7.5ha grass field near Lem, Denmark in June 2015 under an actual mowing operation. Obstacles are placed in the trajectory of the tractor to simulate collision situations (Christiansen et al. 2015a). The row crop operation use case is recorded in mid-September 2015 in a low row crop maize field in Foulum, Denmark. Unlike the moving use case, objects are placed just outside the tractor trajectory allowing them to remain static. A total of 10 images - 5 from each use case - have been selected and roughly annotated.

2.3. Methods

Semantic segmentation is described in the first section followed by a section describing remapping of model predictions.

2.3.1. Fully Convolutional Neural Network for Semantic Segmentation

A traditional CNN performs image classification, thus it can only take a fixed sized input image and output a single prediction/label. A CNN trained on e.g. faces can only tell if there is a face or not in the image as shown in Figure 2. A CNN can be transformed into a fully convolutional neural network (FCNN) by converting the fully connected layers into convolutions. A FCNN is able to forward larger images through the network and output a grid of prediction, thus providing information on both the object type and object location in the image. Converting a CNN to a FCNN for e.g. face image classification, will provide a coarse heat map as presented in Figure 3, showing the position of a specific sized face in the image. Training the CNN to recognize other objects will provide multiple heat maps one for each object type.

Figure 2. CNN training examples for face image classification. Respectively a negative and positive sample.

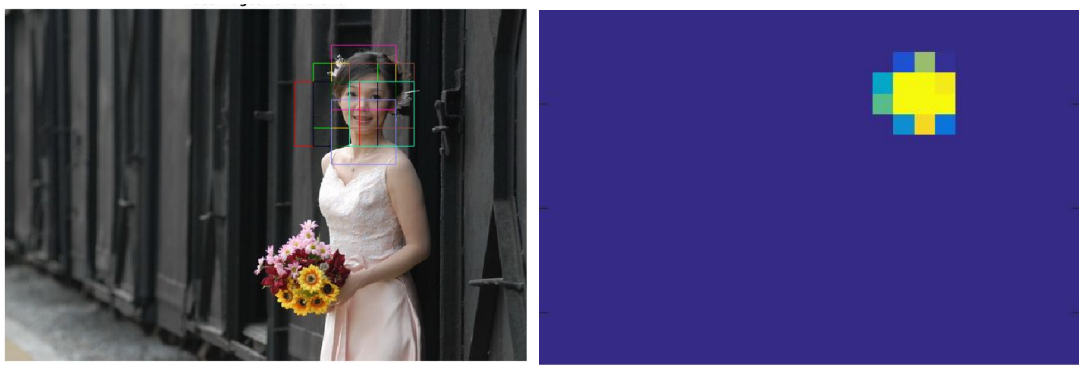

Figure 3. A fully convolutional neural network provides a grid of predictions or a heat map as output.

The above described principle is used in (Long, Shelhamer, and Darrell 8 Mar, 2015) to perform semantic segmentation with FCNN. The network is based on a very deep CNN (Simonyan and Zisserman 2014) (VGG) with 16-layers trained for image classification on ImageNet (Berg and Deng 2015). The CNN network is transformed into a FCNN by discarding the classification layer of VGG and converting the fully connected layers to convolutional layers. A 1x1 convolution is appended with a channel for each object type. The specific structure of VGG allows it to provide classifications for every 32 pixels. By adding a deconvolutional layer, the network will upsample the heatmap to the size of the original image. The network is now able to perform end-to-end training on semantic segmented images. This is defined as a FCN-32 network. To get denser spatial predictions, a 1x1 convolution with a channel for each object is appended to the output of two previous pooling layers with respectively a stride of 16 and 8. The output of earlier pooling layers will provide weaker predictions, but better spatial precision. By fusing the output layers with respectively a stride of 32, 16 and 8, the combination of strong predictions and the refined spatial precision, is found to improve the overall network accuracy. This network is defined as a FCN-8 network. Three FCN-8 models are provided by (Long, Shelhamer, and Darrell 8 Mar, 2015) one trained on 21 classes (PASCAL Voc object including a background class) and two models on PASCAL-Context for both the 33 and 59 most frequent classes. As described in (Long, Shelhamer, and Darrell 8 Mar, 2015), the algorithm is able to roughly process images with 4Hz using a high-end GPU. Using the principles from (Han, Mao, and Dally 2015), the memory footprint and the processing requirements can be reduced for a CNN without damaging accuracy, thus making it suitable for real-time applications.

2.3.2. Prediction mappings

Preferably a network is retrained only on PASCAL-Context classes relevant to agriculture. Alternatively, all object classes from PASCAL-Context are mapped to a few agriculture super-categories and retrained. E.g. dog, cat, cow and horse are all mapped to animal, or road, ground, sand, floor is all mapped to ground. However, as a preliminary study we perform simple mapping of predictions provided by a FCN-8 network to the following 11 agricultural super-categories; *animal, building, field, ground, obstacle, person, shelterbelt, sky, vehicle, water, and unknown*. The result of a prediction and remapping is presented in the Figure 4.

Figure 4. Left: Input image. Right: Result after remapping FCN-8 predictions to agriculture specific classes.

3. Results and Discussion

3.1. Results

Using the ground truth annotations for the 10 test images, the overall classification accuracy is 82.81%. Evaluating the the grass and row crop individually shows - with a classification accuracy of respectively 95.25% and 70.54% - a significant spread between the two use cases. The spread is also clearly demonstrated in Figure 5-9, where the grass and the row crop cases are presented in respectively Figure 5-6 and Figure 7-9. To make a qualitative comparison between semantic segmentation and object detection algorithms the left images in each figure show the input image to the FCN-8 algorithm including detections performed by YOLO (Redmon et al. 2016) and a pedestrian detector (Nam, Dollár, and Han 2014). Both object detection algorithms are close to real-time performance.

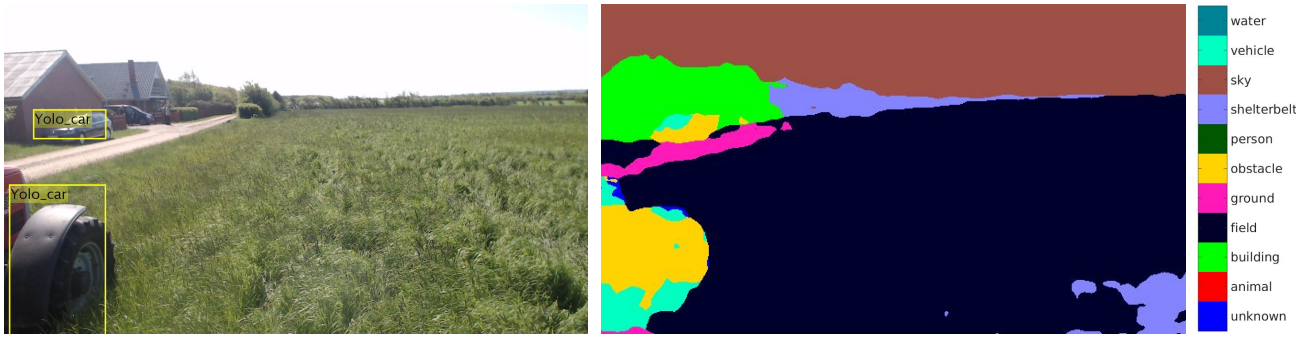

Figure 5. Test image in grass. Semantic segmentation detects field, road, building, shelterbelt, sky, bits of vehicle. However, the tractor is classified as both vehicle, obstacle and bits of unknown. A bit of high grass to the right is classified as shelterbelt. The vehicle right next to house is mostly detected as an obstacle. YOLO is able to detect both tractor and vehicle.

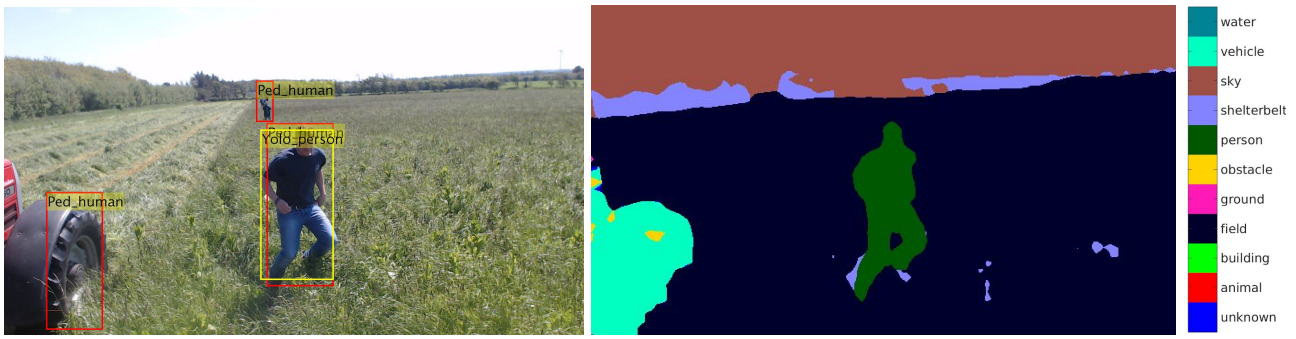

Figure 6. Test image in grass. Semantic segmentation detects field, tractor, sky, shelterbelt and person. However, bits of the field are classified as shelterbelt, a section of the distant shelterbelt is classified as sky and the distant human is classified as field. YOLO detects only the first person. The pedestrian detector detects the close and distant human, but provides also a false positive on the wheel.

Figure 7. Test image in row crop. Semantic segmentation detects ground, shelterbelt, animal, field, building and just a bit of person. However, a large section of the field and shelterbelt is classified as sky and the dark area in the top left corner is classified as building. The top left corner - with very low contrast - is presumably classified as building as many

images in PASCAL-Context are taken inside houses with low contrast walls. Finally, the pedestrian detector and not YOLO detects a human.

Figure 8. Test image in row crop. Large areas of shelterbelt and ground is classified correctly and the green barrels are classified as unknown or obstacles. The shelterbelt to the left is - as a reasonable guess - classified as field. However, large areas of ground are classified as obstacle, building and sky. The distant person is only detected by the pedestrian detector.

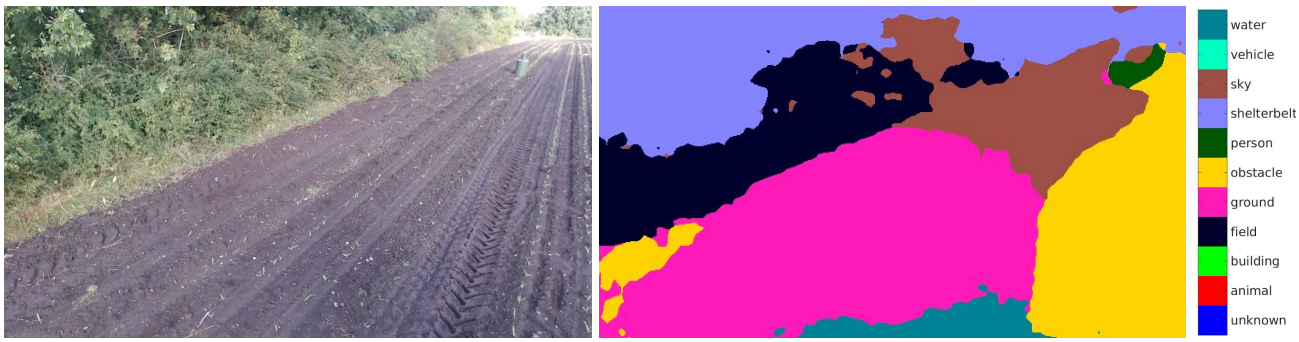

Figure 9. Test image in row crop. Large areas of shelterbelt and ground are classified correctly. The shelterbelt to the left is - as reasonable guess - again classified as field. However, large areas of ground are classified as obstacle, water, sky and person.

3.1.1. Discussion

This preliminary study uses a simple remapping to show the application of deep learning semantic segmentation for autonomous vehicles in agriculture. For an algorithm trained on a completely different data set, a classification accuracy of 95.25% and the presented image examples, show very convincing perceptive capabilities for a grass mowing use case. The row crop use case is less reliable with a classification accuracy of 70.54%. However, the inferior performance in row crops can be explained by the data from PASCAL-Context that do not contain a row crops class. However, we have showed that deep learning semantic segmentation trained on PASCAL-Context is able to generalize to a grass mowing use case, thus allowing us to avoid the comprehensive task of making per-pixel labelling. The preliminary study encourages us to train a new network only on agriculture specific classes from the PASCAL-Context data or alternatively remap all 407 classes to a few agricultural specific classes prior to training. Finally, whole scene annotations of agricultural images would provide even better results.

The image examples show that the object detection algorithms provide fewer misclassifications compared to semantic segmentation. The pedestrian detector is better at detecting people at further distances. However, the YOLO detector is able to detect multiple object types.

Semantic segmentation is able to detect animal, human and vehicle obstacles - as an object detector. However, the benefit of semantic segmentation is both its ability to classify elements that are not precisely delimited with a bounding box and that it provides much denser information of the environment. This information can be used to detect non-traversable areas such as shelterbelts, water, buildings and even unknown obstacles as the barrel. However, a classification of traversable areas such as road, ground or field is also favorable to autonomous farming vehicles when performing navigation and path-planning. To deploy semantic segmentation in a real application, the processing requirements must be evaluated. The current algorithm runs with 4 Hz using a high-end GPU. A smaller CNN network

with lower resolution input images and the principles described in (Han, Mao, and Dally 2015), is able to improve its real-time performance.

Semantic segmentation is as well as most visual camera based solutions not fully reliable and struggles to detect far away elements. A visual camera is also sensitive to weather conditions (rain, fog and snow) and illumination such as direct or dim light (night) (Christiansen et al. 2015b). In the context of agriculture with obstacles below the crop surface, an advantage for a monocular camera is that objects only needs to be visible and not necessarily protruding.

An autonomous vehicle in agriculture should - as in the automotive industry - rely on multiple algorithms and sensor technologies to get more reliable perception of especially visually hidden or camouflaged obstacles and obstacles at far ranges (Christiansen et al. 2015b). However, the low cost of a camera and the power of deep learning perceptive algorithms makes it consumer affordable for unsupervised autonomous vehicles in agriculture. In (Hansen et al. 2016) - also presented at the CIGR conference - the outcome of this work is fused with other sensor technologies and algorithms using occupancy grid maps to detect static obstacles in an agricultural grass field.

4. Conclusions

This preliminary study uses a simple remapping to show the application of deep learning semantic segmentation for autonomous vehicles in agriculture. For an algorithm trained on a completely different data set, a classification accuracy of 95.25% and the presented image examples, show very convincing perceptive capabilities for a grass mowing use case. The row crop use case is less reliable with a classification accuracy of 70.54%. The perception benefits of semantic segmentation compared to object detection has been described and demonstrated using image examples.

Acknowledgements

To Timo Korthals from Bielefeld University for providing a segmentation labelling tool. This research is sponsored by the Innovation Fund Denmark as part of the project “SAFE - Safer Autonomous Farming Equipment” (project no. 16-2014-0).

References

Amodei, Dario, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, et al. 2015. “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1512.02595.

Berg, Alex, and J. Deng. 2015. “Imagenet Large Scale Visual Recognition Challenge 2015.” *Challenge* . http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Large+Scale+Visual+Recognition+Challenge+2010#2.

Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2015. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.” *Iclr*, 1–12. http://arxiv.org/abs/1412.7062.

Christiansen, P., Mikkel Kragh Hansen, Kim Steen, Henrik Karstoft, and Rasmus Nyholm Jørgensen. 2015a. “Platform for Evaluating Sensors and Human Detection in Autonomous Mowing Operations.” *(Only Submitted) Precision Agriculture, Special Issue ECPA*.

Christiansen, P., M. K. Hansen, K. A. Steen, H. Karstoft, and R. N. Jørgensen. 2015b. “Advanced Sensor Platform for Human Detection and Protection in Autonomous Farming.” In *Precision Agriculture ’15*, 291–98. doi:10.3920/978-90-8686-814-8_35.

Ciresan, D., U. Meier, and J. Schmidhuber. 2012. “Multi-Column Deep Neural Networks for Image Classification.” *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3642–49. doi:10.1109/CVPR.2012.6248110.

Everingham, Mark, Sma Eslami, and Luc Van Gool. 2013. “The Pascal Visual Object Classes Challenge–a Retrospective.” *Homepages.Inf.Ed.Ac.Uk*. doi:10.1007/s11263-014-0733-5.

Farfade, Sachin Sudhakar, Mohammad Saberian, and Li-Jia Li. 2015. “Multi-View Face Detection Using Deep Convolutional Neural Networks.” *Cornell University Library*. http://arxiv.org/abs/1502.02766v2.

Girshick, Ross, Jeff Donahue, Trevor Darrell, U. C. Berkeley, and Jitendra Malik. 2014. “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” *Cvpr’14*, 2–9.

doi:10.1109/CVPR.2014.81.

Hansen, Mikkel Kragh, Peter Christiansen, Timo Korthals, Thorsten Jungeblut, Henrik Karstoft, and Rasmus Nyholm Jørgensen. 2016. "Multi-Modal Obstacle Detection and Evaluation of Evidence Grid Mapping in Agriculture." In . Aarhus University.

Han, Song, Huizi Mao, and William J. Dally. 2015. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding." http://adsabs.harvard.edu/abs/2015arXiv151000149H.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." *arXiv Preprint arXiv ...* cs.CV: 1–14. doi:10.1109/TPAMI.2015.2389824.

He, Kaiming, Xiangyu Zhang, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1512.03385.

Krizhevsky, A., I. Sutskever, and Ge Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*, 1097–1105.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–2324.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 8 Mar, 2015. "Fully Convolutional Networks for Semantic Segmentation"

Mottaghi, Roozbeh, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. "The Role of Context for Object Detection and Semantic Segmentation in the Wild." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 891–98. IEEE. doi:10.1109/CVPR.2014.119.

Nam, Woonhyun, Piotr Dollár, and Joon Hee Han. 2014. "Local Decorrelation For Improved Detection." *Advances in Neural Information Processing Systems*, 1–9. http://papers.nips.cc/paper/5419-local-decorrelation-for-improved-pedestrian-detection.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. "You Only Look Once: Unified, Real-Time Object Detection." http://arxiv.org/abs/1506.02640v3.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. 2015. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1506.01497.

Simonyan, Karen, and Andrew Zisserman. 2014. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1409.1556.

Taigman, Yaniv, Marc Aurelio Ranzato, Tel Aviv, and Menlo Park. 2014. "DeepFace : Closing the Gap to Human-Level Performance in Face Verification," June.

Torr, Philip H. S. 2014. "Conditional Random Fields as Recurrent Neural Networks." *arXiv Preprint*. http://arxiv.org/abs/1502.03240v1.