Journal of *Imaging*

MDPI

Article

Using Deep Learning to Challenge Safety Standard for Highly Autonomous Machines in Agriculture

Kim Arild Steen *,†, Peter Christiansen †, Henrik Karstoft and Rasmus Nyholm Jørgensen

Department of Engineering, Aarhus University, Finlandsgade 22 8200 Aarhus N, Denmark; pech@eng.au.dk (P.C.); hka@eng.au.dk (H.K.); rnj@eng.au.dk (R.N.J.)

* Correspondence: kim.steen@eng.au.dk; Tel.: +45-3116-8628

† These authors contributed equally to this work.

Academic Editors: Francisco Rovira-Más and Gonzalo Pajares Martinsanz

Received: 18 December 2015; Accepted: 2 February 2016; Published: 15 February 2016

Abstract: In this paper, an algorithm for obstacle detection in agricultural fields is presented. The algorithm is based on an existing deep convolutional neural net, which is fine-tuned for detection of a specific obstacle. In ISO/DIS 18497, which is an emerging standard for safety of highly automated machinery in agriculture, a barrel-shaped obstacle is defined as the obstacle which should be robustly detected to comply with the standard. We show that our fine-tuned deep convolutional net is capable of detecting this obstacle with a precision of 99.9% in row crops and 90.8% in grass mowing, while simultaneously not detecting people and other very distinct obstacles in the image frame. As such, this short note argues that the obstacle defined in the emerging standard is not capable of ensuring safe operations when imaging sensors are part of the safety system.

Keywords: deep learning; obstacle detection; autonomous; ISO

1. Introduction

In order for an autonomous vehicle to operate safely and be accepted for unsupervised operation, it must perform automatic real-time risk detection and avoidance in the field with high reliability [1]. This property is currently being described in an ISO/DIS standard [2], which contains a short description of how to meet requirements for obstacle detection performance. The requirements for tests and the test object are described in Section 5 in the standard. The standard uses a standardized object, shown in Figure 1 which is meant to mimic a human seated (torso and head). This standardized object makes sense for non-imaging sensors such as ultrasonic sensors, LiDARs or Time-of-Flight cameras, which measures the geometrical properties of the objects or distance to the objects. However, for an imaging sensor such as an RGB camera, the definition of this standardized object does not guarantee safety. In this short note, we will present how deep learning methods can be used to design an algorithm that robustly detects the standardized object in various situations, including high levels of occlusion. Based on this, the algorithm is able to comply with the standard, but at the same time, it is not detecting people and animals, as they are not part of the trained model.

Deep convolutional neural networks have demonstrated outstanding performance in various vision tasks such as image classification [3], object classification [4,5], and object detection [4–6]. LeCun et al. formalized the idea of the deep convolutional architecture [7], and Krizhevsky *et al.* introduced a paradigm shift in image classification with the AlexNet [3]. In recent years the AlexNet has been used in various deep learning related publications.

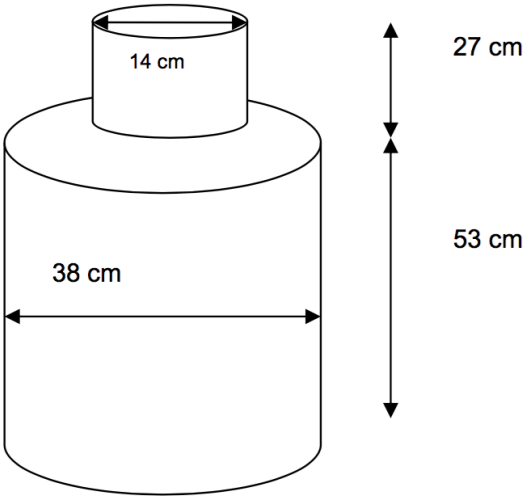

Figure 1. Standardized obstacle.

2. Materials and Methods

In this section, we present the data used for this paper, together with a short description of how the deep learning algorithm was trained and implemented.

2.1. Data Collection

The results in this paper are based on recordings performed in both row-crop fields and grass fields. The recordings were part of a research project aimed at improving safety for semi-autonomous and autonomous agricultural machines [1]. The recordings contain various kinds of obstacles, such as people, animals, well covers and the ISO standardized obstacle. All obstacles are stationary and the data is recorded while driving towards and past them. The sensor kit (seen in Figure 2a) used for the experiments includes a number of imaging devices: a thermal camera, a 3D LiDAR, an HD-webcam and a stereo camera. In this paper, we only focus on the RGB-images. Images from the experiments and recordings are seen in Figure 2.

(**a**) (**b**) (**c**)

Figure 2. Images from experiments and recordings. (**a**) The sensor kit mounted on a mower; (**b**) An example image from the recordings in grass; (**c**) An example image from the recordings in row crops.

2.1.1. Training Data

An iPhone 6 was used to record five short videos of the ISO standardized obstacle. The recordings include various rotations, scales, and intensity of the object. A total of 437 frames from the videos were extracted and bounding boxes of the object were created. Example frames from the videos can be seen in Figure 3.

Figure 3. Examples of training data.

2.2. Training of Deep Convolutional Network

Deep learning is utilized for barrel detection using a sliding window approach similar to [6]. We start by fine-tuning AlexNet (Available at the Caffe model zoo [8]), which is pre-trained on data from the image-net competition [9] for ISO standardized obstacle detection. By fine-tuning the neural network to images of the ISO obstacle, which has a specific shape, texture and color, the algorithm will be very good at detecting the occurance of this specific object in the image. At the same time, the algorithm will also be very good at rejecting other objects in the image (animals, people, etc.), thereby meeting the standard with respect to performance, but not with respect to safety.

To increase the number of training examples (both positive and negative), we randomly sampled sub-windows of the extracted training images. A sub-window was labeled as positive (containing an object) if it had over 50% intersection over union with the labeled bounding boxes. To include additional negatives, non-face sub-windows are collected from *Annotated Facial Landmarks in the Wild* database [10] in a similar approach. A total of 1925 positive and 11,550 negative samples have been used in this paper. These examples were then resized to 114×114 pixels and used to fine-tune a pre-trained AlexNet model. The original AlexNet model outputs a vector of 1000 units, each representing the probability of the 1000 different classes. In our case, we only want to detect if an image patch contains an ISO object or not. Hence, the last layer is changed to output a two-dimensional vector. For fine-tuning, we used 14K iterations and batch size of 100 images, where each batch contained 67 positive and 33 negative examples. During fine-tuning, the learning rate for the convolutional layers was 10 times smaller than the learning rate of the fully connected layers.

After fine-tuning, the fully-connected layers of the AlexNet model can be converted into convolutional layers by reshaping layer parameters [11]. This makes it possible to efficiently run the network on images of any size and obtain a heatmap of the ISO obstacle classifier. Each pixel in a heatmap shows the network response, which is the likelihood of having an ISO obstacle, corresponding to the 114 pixel $\times$114 pixel region in the original image. In order to detect ISO obstacles of different sizes, the input images can be scaled up or down. The chosen training image resolution is half the resolution used in the original AlexNet. Reducing the resolution of the training images allows us to reduce the input image by half, thus reducing processing time, while maintaining the resolution of the resulting heatmaps. An example of a resulting heatmap is illustrated in Figure 4.

Figure 4. Illustration of ISO obstacle and resulting heatmap. (**a**) RGB image from the row crop field; (**b**) Resulting heatmap.

The model was trained using the Caffe framework [12] using a single GPU (4 GB Quadro K2100M). The training time was approximately 1–2 h.

2.3. Detection of ISO Obstacle using Deep Convolutional Network

When the deep convolutional network has been trained, it can be used to detect the ISO obstacle in color images. In order to detect the obstacle at multiple distances, the input image needs to be scaled accordingly. In this paper, we use 13 scales, which are all processed by the same network structure. We use 13 scales to be able to detect the barrel when it is far away (57 pixel $\times$ 57 pixel in the original image) and up close (908 pixel $\times$ 908 pixel). As described in the previous section, the output is a heatmap, where the intensity reflects the likelihood of an ISO obstacle.

Based on the heatmap, one can detect the obstacle and draw a bounding box. Each pixel in the heatmap corresponds to a 114 pixel $\times$ 114 pixel sub-window in the input image. To remove redundant overlapping detection boxes, we use non-maximum suppression [6] with 50% overlap threshold.

3. Results

The results are based on data collected in two different cases, at three different dates. The data has been collected at different times during the days to ensure different lighting conditions. In both cases, we use the model trained on the data presented in Section 2.1.1. Despite this, the results in this paper are of a preliminary nature, as various weather conditions and scenarios are not included in the data.

3.1. Row Crops

Based on the presented algorithm, a total of 7 recordings have been evaluated with respect to detection of the ISO obstacle in row crops. The recordings also contain other kinds of obstacles such as people and animals. The recordings contain a total of 14,153 frames with 20,414 annotated obstacles (8126 of those are the ISO obstacle).

In the ISO standard, the obstacles needs to be detected within a defined safety distance. The safety distance is a product of the expected working speed and machine type. Hence, there is no fixed value for this. In Figure 5, a histogram of the achieved detection distances is shown. It is seen that the algorithm is able to detect the obstacle both at close range (3–6 m) and also at far range (over 15 m). The ISO obstacle is present in front of the machine a total of 14 times and the algorithm is able to detect the obstacle everytime. The detection distance, which is the distance of the first positive detection, for these 14 times, ranges from 10 m to 20 m, with an average of 14.56 m.

Evaluating all frames, at frame level with all annotated objects, we achieve TP (true positive) = 2851, FP (false positive) = 1, TN (true negative) = 7105 and FN (false negative) = 4919. The high number of false negatives is a result of annotations, where the ISO obstacle is located more than 20 m away, which is more than the achieved detection range. Based on this, we achieve a hit rate of 36.7% Equation (1), which is the ratio between positive detections and all annotations of the ISO obstacle, and a precision of 99.9% Equation (2) (As there are less than 10,000 datapoints, we are not able to present the last decimal as a result). In the standard, it is stated that the system must achieve a success rate of 99.99%, however, it is unclear how this should be tested. As stated above, the algorithm is able to detect the ISO obstacle when the ISO obstacle is within 10 m of the machine at a precision of almost 100%.

$$hit\ rate = \frac{TP}{TP + FN} = \frac{2851}{2851 + 4919} = 0.3669 \tag{1}$$

$$precision = \frac{TP}{TP + FP} = \frac{2851}{2851 + 1} = 0.9996 \tag{2}$$

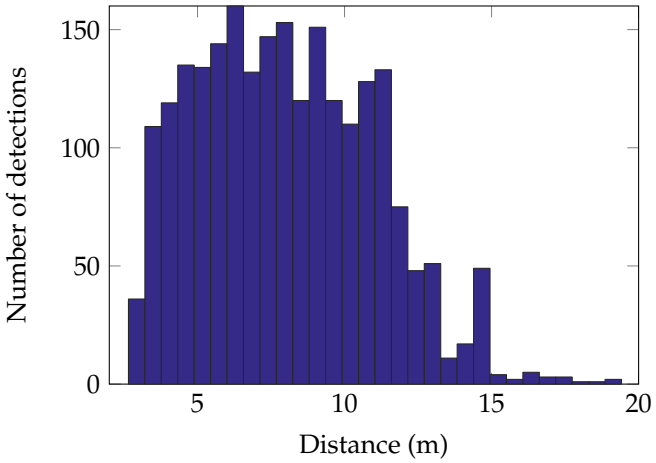

Figure 5. Histogram of detection distances.

In Figure 6, the achieved hit rate for different distance ranges is shown. It is seen that the algorithm is not able to detect the obstacle in all frames (the hit rate is below 1). However, with a precision close to 100%, a single detection is reliable enough to be considered a detection of the ISO obstacle present in the recordings. Hence, the success rate is 99.9%, estimated at frame level, within an average safety distance of 14.56 m.

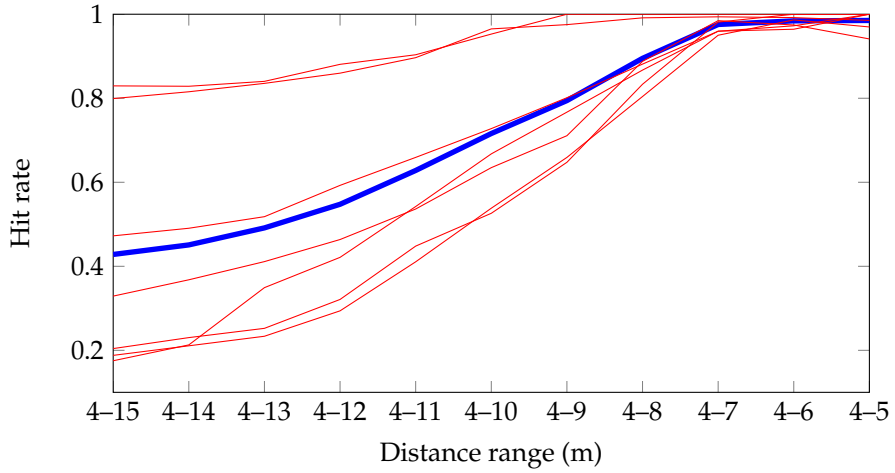

Figure 6. Hit rate, evaluated at frame level, for different distance ranges (e.g. 0.97 is the mean hit rate for all frames in range 4–7 m). Red: hit rate for the 7 recordings, and blue: average hit rate.

3.2. Grass Mowing

We also evaluated the algorithm in a much more difficult case; grass mowing. In grass mowing, the obstacles are often highly occluded (a scenario that is not described in detail in the standard). The recording contains the ISO obstacle and people. The recordings contain a total of 19,390 frames with 936 annotated ISO obstacles.

As with the row crops case, we evaluate the achieved detection distance. In the recording, the ISO obstacle is detected when the ISO obstacle is within approximately 6 m of the machine. The ISO obstacle is present in front of the machine a total of 8 times and the algorithm is able to detect the obstacle everytime.

Evaluating all frames at frame level, we achieve TP = 307, FP = 31, TN = 18,337 and FN = 787. The high number of false negatives is a result of annotations, where the ISO obstacle is located more

than 6 m away, which is more than the achieved detection range. Based on this, we achieve a hit rate of 28.1% Equation (3)

$$hit\ rate = \frac{TP}{TP + FN} = \frac{307}{307 + 787} = 0.281 \tag{3}$$

$$precision = \frac{TP}{TP + FP} = \frac{307}{307 + 31} = 0.908 \tag{4}$$

and a precision of 90.8% Equation (4). Again, the ISO obstacle is successfully detected everytime we are driving towards it, however, the achieved precision is lower than 99.99% which is stated in the standard. As seen in Figure 7, the ISO obstacle is highly occluded. In the standard, it is noted that the obstacle must appear unobscured to ensure high levels of system confidence and if the obstacle is obscured, the manufacturer must understand how this affects performance. In the grass case, we show how the system is affected by this. The precision drops from 99.9% to 90.8% and the achieved detection distance drops from an average of 14.56 m to 6 m. The lower precision is due to a higher number of false positives. Most of these false positives are the tractor in the image. The tractor will always be present in the same position in the image and further developments could remove this in post-processing or by including tractor images as negatives in the training data.

Figure 7. Detections in grass mowing case. Notice that people are not detected.

4. Discussion

The results show that we are able to robustly detect the presence of the ISO obstacle in various conditions including different lighting and heavy occlusion. The results also show that the algorithm is not able to detect the presence of the other obstacles within the image frame—even people. This is not a surprise, as the model has been trained on the ISO obstacle and not on other types of obstacles.

We are using a large deep learning model to detect a simple object and it might seem like we are overdoing it. However, by using deep learning and pre-trained networks, we have been able to design a robust classifier for detecting the ISO obstacle in various scenarios, using only a few minutes of training video. This shows the power of these models and how they can be exploited.

In this paper, we have implemented the algorithm using Caffe and the corresponding MATLAB interface, which means that it does not run in real-time. However, CNN implementations ready for real-time applications exist in literature [13]. Furthermore, deep learning algorithms are also being utilized in detection systems for the car industry, where deep learning models are able to classify between a number of different obstacles. This is powered by high-end embedded boards from NVIDIA,

which has recently launched a 300 USD board [14], enabling real-time deep neural nets on UAVs and UGVs. These observations make it fair to assume that agricultural machines could also benefit from that computing power.

The ISO standard states that the system must have a success rate of 99.99% in various weather conditions, however, it is unclear how this success rate should be measured. We are not able to detect the ISO obstacle in all frames, however, we achieve a precision of 99.9%, and are able to detect the ISO obstacle at an average distance of 14.56 m in row crops. In the grass mowing case, the obstacle was highly occluded which affected the achieved detection distance. Furthermore, the presence of the tractor within the image frame resulted in more false positives. These should be removed to ensure higher precision. We have not been able to test the performance in various weather conditions, hence, the results obtained in this paper is of a preliminary nature.

The most important result in this paper is that we are able to show that an imaging system can be designed to comply with the ISO standard and completely miss important obstacles such as people—both kids and adults. We argue that the standardized obstacle presented in the standard is not fully adequate to ensure safe operations with highly autonomous machines in agriculture. The design of the obstacle is based on a human head and torso (human sitting down), but we show that we can detect this type of obstacle and at the same time completely miss the people in front of the machines. As the ISO standard aims to represent a human sitting down, we suggest that the standardized obstacle should resemble a more life-like person, if a standardized object is required. In our experiments, we have used mannequins for this task. It is important that the life-like obstacles have the same properties as the current ISO obstacles with respect to color, material, and temperature (hot water). The life-like obstacle could be used to test different possible postures, such as standing, lying and sitting. However, other kinds of obstacles could also be present in the fields. Including other types of obstacles, such as animals, in the requirements, could potentially increase safety overall. However, even doing this, there is no guarantee that the methods will be able to detect real obstacles in the fields, as they might not be perfectly described by the trained algorithm. Deep learning methods have achieved very good performance in very difficult image and object recognition tasks. This is accomplished through access to a vast amount of image training data, where objects like people, animals and cars are depicted in thousands of different postures, sizes, colors and situations. The deep learning framework is able to encapsulate this great amount of variability and thereby produce beyond-human performance in object recognition tasks. This is being exploited in the work towards autonomous cars and could also be done in agriculture.

5. Conclusions

In an emerging standard for safety of highly automated machinery in agriculture, a barrel-shaped obstacle is defined as the obstacle which should be robustly detected to comply with the standard. In this paper, we show that by using deep learning techniques, we are able to achieve a high level of precision in two different cases in agricultural operations, with one of the cases concerning highly occluded obstacles. The algorithm detects the ISO specified obstacle in every test run, but it completely misses important obstacles such as people.

Therefore, we argue that the standardized obstacle presented in the standard is not fully adequate to ensure safe operations with highly autonomous machines in agriculture and further work should be conducted to describe an adequate procedure for testing the obstacle detection performance of highly autonomous machines in agriculture.

Acknowledgments: This research is sponsored by the Innovation Fund Denmark as part of the project "SAFE - Safer Autonomous Farming Equipment" (Project No. 16-2014-0) and "Multi-sensor system for ensuring ethical and efficient crop production" (Project No. 155-2013-6). We would also like to thank Ole Green from Agro Intelligence ApS for his contributions to the projects.

Author Contributions: Kim Arild Steen and Peter Christiansen contributed to the data aquisition, algorithm development, algorithm testing and writing of the manuscript. Henrik Karstoft has contributed with internal review of the manuscript, and Rasmus Nyholm Jørgensen has contributed with formalizing the hypothesis of this contribution, field experiment planning, algorithm testing and internal review of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Christiansen, P.; Hansen, M.; Steen, K.; Karstoft, H.; Jørgensen, R. Advanced sensor platform for human detection and protection in autonomous farming. In *Precision Agriculture'15*; Wageningen Academic Publishers: Wageningen, The Netherlands, 2015; pp. 1330–1334.
2. ISO/DIS 18497. Available online: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=62659 (accessed on 17 December 2015).
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
4. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. Available online: http://arxiv.org/pdf/1409.4842v1.pdf (accessed on 18 December 2015).
5. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. Available online: http://arxiv.org/pdf/1409.1556v6.pdf (accessed on 18 December 2015).
6. Farfade, S.S.; Saberian, M.; Li, L.J. Multi-view Face Detection Using Deep Convolutional Neural Networks. Available online: http://arxiv.org/pdf/1502.02766v3.pdf (accessed on 18 December 2015).
7. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
8. AlexNet. Available online: http://caffe.berkeleyvision. org/model_zoo.html (accessed on 10 October 2015).
9. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; *et al*. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2014**, *115*, 1–42.
10. Koestinger, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, Barcelona, Spain, 6–13 Novenber 2011.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Available online: http://arxiv.org/pdf/1411.4038v2.pdf (accessed on 18 December 2015).
12. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. Available online: http://arxiv.org/pdf/1506.02640v4.pdf (accessed on 18 December 2015).
14. Nvidia Jetson TX1. Available online: http://www.nvidia.com/object/jetson-tx1-dev-kit.html (accessed on 1 December 2015).