designs

Article

Multi-Camera-Based Person Recognition System for Autonomous Tractors

Taek-Hoon Jung, Benjamin Cates, In-Kyo Choi, Sang-Heon Lee and Jong-Min Choi *

LSMtron Ltd., Dongan-gu, Anyang-si 14118, Gyeonggi-do, Korea; taekhoon.jung@lsmtron.com (T.-H.J.); Benjamin.cates@lsmtron.com (B.C.); thenevers@lsmtron.com (I.-K.C.); papalsh@lsmtron.com (S.-H.L.)

* Correspondence: jongmin@lsmtron.com

Received: 13 November 2020; Accepted: 7 December 2020; Published: 9 December 2020

Abstract: Recently, the development of autonomous tractors is being carried out as an alternative to solving the labor shortage problem of agricultural workers due to an aging population and low birth rate. As the level of autonomous driving technology advances, tractor manufacturers should develop technology with the safety of their customers as a top priority. In this paper, we suggest a person recognition system for the entire environment of the tractor using a four-channel camera mounted on the tractor and the NVIDIA Jetson Xavier platform. The four-channel frame synchronization and preprocessing were performed, and the methods of recognizing people in the agricultural environment were combined using the YOLO-v3 algorithm. Among the many objects provided by COCO dataset for learning the YOLO-v3 algorithm, only person objects were extracted and the network was learned. A total of 8602 image frames were collected at the LSMtron driving test field to measure the recognition performance of actual autonomous tractors. In the collected images, various postures of agricultural workers (ex. Parts of the body are obscured by crops, squatting, etc.) that may appear in the agricultural environment were required to be expressed. The person object labeling was performed manually for the collected test datasets. For this test dataset, a comparison of the person recognition performance of the standard YOLO-v3 (80 classes detect) and Our YOLO-v3 (only person detect) was performed. As a result, our system showed 88.43% precision and 86.19% recall. This was 0.71% higher precision and 2.3 fps faster than the standard YOLO-v3. This recognition performance was judged to be sufficient considering the working conditions of autonomous tractors.

Keywords: autonomous tractor; deep learning; object detection; YOLO-v3

1. Introduction

Smart farming and precision agriculture involve the integration of advanced technologies into existing farming practices in order to increase production efficiency and the quality of agricultural products [1]. As an added benefit, the quality of life for farm workers is improved by reducing the demands of labor and tedious tasks.

Replacing human labor with automation is a growing trend across multiple industries, and agriculture is no exception. Most aspects of farming are exceptionally labor intensive and repetitive tasks.

Current and impending agricultural technologies that are expected to become the pillars of the smart farm primarily fall into three categories: autonomous robots, drones or UAVs (Unmanned Aerial Vehicle), sensors, and the Internet of Things (IoT).

The tractor is the heart of the farm, used for many different tasks depending on the type of farm and the attached implement. Autonomous tractors will become more capable and self-sufficient over time, especially with the inclusion of additional cameras and artificial intelligence vision systems, GPS for navigation, IoT connectivity to enable remote monitoring and remote operation, and RaDAR and LiDAR for object detection and avoidance. All of these technological advancements will significantly diminish the need for humans to actively control these machines. As such, autonomous tractors are being researched as a solution to the problems associated with an aging population and labor reduction worldwide.

In the agricultural machinery industry, the technical level of autonomous tractors is divided into four stages. The first stage is non-automated agricultural machinery and the core functions are mechanical operation and repetitive operation. For the second stage, automated agricultural machines still have to be operated by the driver, but the role of monitoring as part of the functions of agricultural machines is also included. The third stage is a restrictive autonomous agricultural machine with core functions, including automatic operation, the generation of a work route, and the following of that work route. In the final stage, stage four, the agricultural machine is responsible for operating as a fully autonomous agricultural machine, including a function for generating a work route, following that work route, as well as the ability to move to a work site or field by itself.

In order to advance to level three or higher, a recognition sensor must collect data on the tractor's surrounding environment, analyze it, and combine recognition and judgment techniques with respect to obstacles to determine dangerous factors while driving. These cognitive and judgment techniques are especially important because they directly affect the safety of drivers and farmers.

Recognition sensors include cameras, LiDAR, RaDAR, and Ultrasonic, each with their own strengths and weaknesses. Therefore, sensors shall be selected and used appropriately with respect to the characteristics of the environment. The driving environment of automobiles has a standardized infrastructure, such as lanes and guard rails on paved roads in free space, but the driving environment of tractors differs in that they are exposed to an atypical environment in which the size and shape of the field change over time, such as crops or unpaved cultivated land. In unformatted agricultural environments, the recognition of obstacles to RaDAR and LiDAR sensors has the disadvantage of being difficult to distinguish between obstacle objects and crops. On the other hand, camera sensors are able to extract unique characteristics about the obstacle like shape and color, and thus recognize obstacles even in unformatted environments that change over time.

Past object recognition studies have been conducted by designing and detecting features of objects such as SIFT (Scale Invariant Feature Transform) [2], SURF (Speeded-Up Robust Feature) [3], Haar [4], and HOG (Histogram of Oriented Gradients) [5]. For example, these methods were utilized to detect books; they appear in trapezoidal shapes in images, create angles at vertices, and have a degree of thickness. Since then, DPM (Deformable Part-based Model) [6] increased object recognition performance by dividing objects into several parts to organize characteristic information and connecting the flexible structure of each part with machine learning methods such as SVM (Support Vector Machine) [7]. These feature-based object recognition techniques tended to rely more on domain knowledge than on the data itself, and it was difficult to tune passively if these features had multiple parameters.

Object recognition methods using deep learning became mainstream as the synthetic product neural network showed overwhelmingly superior performance of existing feature-based object recognition in ImageNet Challenge 2012 [8]. CNN first appeared in cursive number recognition [9], which applied convolution computation. Although the existing neural network structure lacked the ability to express local information around pixels.

The problem of recognizing what an object is in the image has been solved to some extent by CNN, but finding out where the object exists in the image is another problem. In recent years, research has emerged on how to detect the location of object. It aims to find objects of interest in an image and draws bounding boxes around them while also categorizing its class.

The application of deep learning technology for object recognition is becoming increasingly popular in the field of computer vision. The location and the classification of target objects can be determined by using object detection. Recently, object detection has been applied in many fields such as agriculture [10–12], military and civil areas [13,14], intelligent surveillance [15,16], autonomous driving [17–19], and intelligent transportation [20,21].

Due to the rapid development of graphics processing units (GPU) hardware performance, deep learning has made significant progress [22] and object detection algorithms based on these deep learning techniques are widely used in our daily life, such as person detection [23], face detection [24], and vehicle detection [25,26].

The convolutional neural network (CNN) is a kind of network with many layers used to extract features based on invariance of regional statistics in images [27]. By training on the dataset, CNN can learn the features of the objects that need to be detected autonomously and the performance of the model can be gradually improved. As a result of continuously improving computer hardware, the structure of CNN can become much deeper and complex.

The state-of-the-art object detection algorithms based on CNN can be divided into two categories. The first category is referred to as two stage object detection algorithms, such as R-CNN [28], Fast R-CNN [29], Faster R-CNN [30], Mask R-CNN [31], etc. The object detection process is divided into two phases by these algorithms. They use a Region Proposal Network (RPN) to generate candidate anchor boxes and then the location and class of the candidate objects can be predicted and identified by the detection network. These algorithms can achieve high accuracy for object detection. However, they are not end-to-end object detection algorithms.

The second category is one stage object detection algorithms, such as OverFeat [32], Single Shot Detector (SSD) [33], YOLO series [34–38], etc. In these algorithms, there is no need to use RPN to generate the candidate objects. Instead, the target location and class information are predicted directly through the network. They are end-to-end object detection algorithms. End-to-end algorithm refers to training a possibly complex learning system represented by a single network that represents the complete target system, bypassing the intermediate layers usually present in traditional pipeline designs. Therefore, the speed of the one stage object detection algorithm is faster [39].

YOLO networks excel because they are capable of high accuracy detection while likewise having the ability to run in real time. The YOLO algorithm "only looks once" at the image, so it requires just one forward propagation pass through the neural network to make forecasts. After non-maximum suppression which ensures the object detection algorithm just detects each object once, the algorithm then yields recognized objects along with the bounding boxes.

YOLO networks are capable of high accuracy detection and can perform in real-time. The YOLO algorithm "only looks once" at the image, so it requires just one forward propagation pass through the neural network to make forecasts. Non-maximum suppression ensures the object detection algorithm detects each object only once, and the algorithm outputs classified objects and their bounding boxes. YOLO trains on full pictures and improves detection performance in a straightforward manner. YOLO-v3 was chosen of the YOLO series variants because it uses fewer parameters, resulting in a faster algorithm processing speed, and improves recognition accuracy.

There are studies on vision-based obstacle recognition in the agricultural environment [40–42]. However, these studies only focus on detecting obstacles to the front. Unlike automobiles, agricultural tractors work with an implement attached to the rear. It is necessary to ensure safety in all directions from the tractor because forward and reverse gear shifts occur frequently. Therefore, autonomous tractors must be able to recognize and judge objects in all direction.

In this paper, a total of four cameras were attached to the tractor in order to recognize people appearing in the surrounding environment, and object recognition based on deep learning was performed through four-channel image data. The configuration of these camera systems has enabled all detection areas around the tractor in one image frame, helping to speed up algorithm processing time by recognizing the person for a single frame without recognizing the person individually for each direction. the YOLO-v3 network was adopted for person recognition, and some structural changes were made to apply the proposed system. Only person objects from the COCO dataset were extracted and used for network learning. The person recognition performance of the standard YOLO-v3 and Our YOLO-v3 was compared using the test datasets acquired at LSMtron driving test field.

2. Materials and Methods

2.1. Autonomous Tractor Vision System

The system focuses on recognizing when a person appears in the surrounding environment of the tractor using cameras and determining which direction the person is from the tractor, for the purpose of ensuring the safety of autonomous tractor drivers and farmers. In the case of tractors, safety must be ensured not only towards the front but in all directions because the rear of the tractor is frequently equipped with a rotavator, plow, etc. The proposed system uses a wide-angle camera with a 120° viewing angle instead of a narrow-angle camera [43]. Unlike narrow-angle cameras, wide-angle cameras have the disadvantage of intensifying radiation distortion as they move away from the principal point. However, the advantage is that image data can be acquired from a wide viewing angle at once.

We virtually identified the detection range of 120° wide-angle camera using the NX 3d modeling tool, and based on this, mounted the cameras on the tractor to minimize dead zones.

A total of four cameras are located on the front, rear, left and right sides of the tractor, and data can be obtained for the entire 360° area surrounding the tractor, as shown in Figure 1.

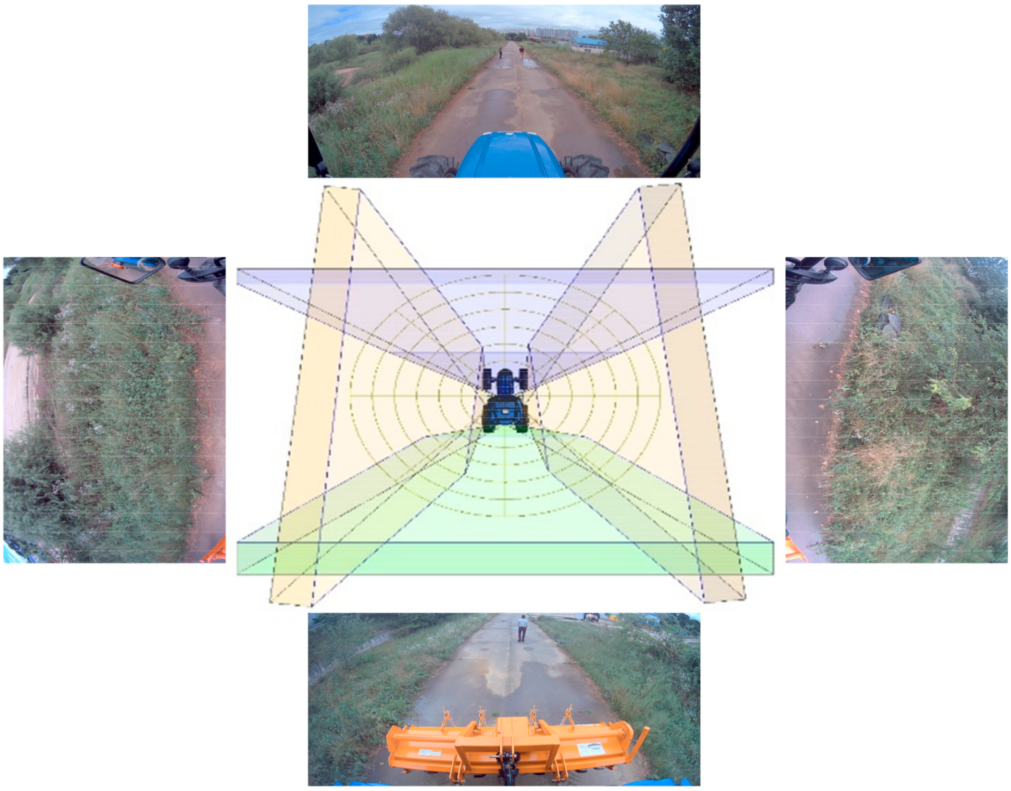


Figure 1. Vision camera detection range.

2.2. Person Recognition System

Figure 2 shows the structure of a system that uses many cameras to recognize a person. 1928 × 1208 resolution images from each camera are input to the camera interface board at a rate of 30 frames per second. The interface board is designed to control the image acquisition function and to output images without time delay between each camera module.

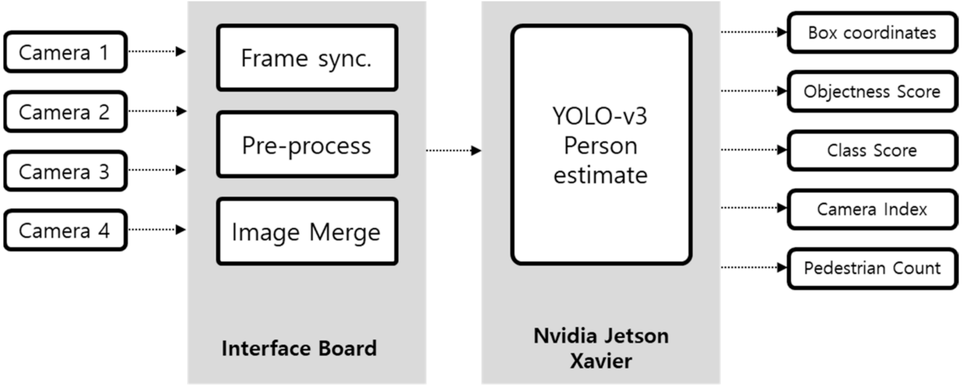


Figure 2. Multi-camera-based person recognition system architecture.

Image resolution is highly correlated with object recognition algorithm processing speed. For an autonomous driving system in which real-time performance is sufficient, the resolution of an image can be reduced to increase the processing speed. Each input image was down sampled to 960 × 544 resolution, as shown in Figure 3. Furthermore, a preprocessing step for matching a four-channel video image to one frame is performed so that a person recognition algorithm can be processed for a single image frame.

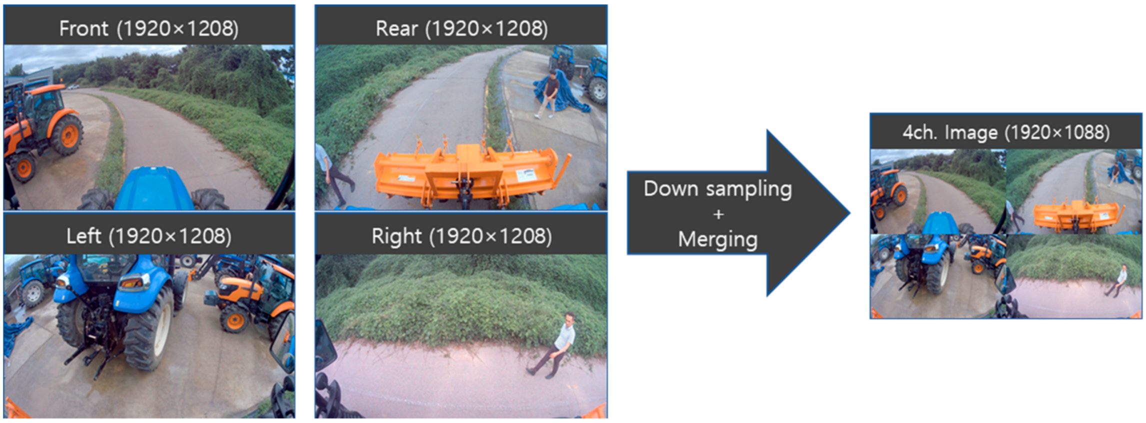


Figure 3. Input image down sampling and merging.

During the preprocessing step, a full HD resolution split image of 1920 × 1088 is generated, which is used as the input to the YOLO-v3 based person recognition algorithm applied in this paper. Object recognition technology using deep learning consists of a neural network structure with many layers. The larger the structure, the more parameters it has, the more computational resources are required. In order to solve such a resource problem, hardware capable of performing high-speed GPU-based parallel processing is essential. This paper used the Jetson AGX Xavier [44], NVIDIA's representative artificial intelligence platform, as an experimental system. Table 1 lists the specifications for the Jetson AGX Xavier.

Table 1. NVIDIA Jetson AGX Xavier technical specifications.

Processing	Description
GPU	512-core Volta GPU with Tensor Cores
CPU	8-core ARM v8.2 64-bit CPU, 8MB L2 + 4MB L3
Memory	32GB 256-bit LPDDR4x | 137GB/s
Storage	32GB eMMC 5.1 + SSD 256GB
DL Accelerator	(2x) NDVLA Engines
Vision Accelerator	70-way VLIW Vision Processor
Encoder/Decoder	(2x) 4Kp60 | HEVC/(2x) 4Kp60 | 12-bit support

2.3. Person Recognition Algorithm with YOLO-v3

The YOLO-v3 algorithm uses a single neural network to obtain a solution by considering the probability of a bounding box, which is the coordinate information of a detected object in the image, and its box as a regression problem. This method is much faster than the object recognition algorithm in the existing R-CNN series and learns not only the appearance characteristics of the object but also the overall context.

The YOLO-v3 network architecture divides the input image into S × S grid cells to determine whether an object is centered in that cell, and the predicted feature map has N prediction boxes per grid, as shown in Figure 4. Each prediction box contains bounding box coordinates including spatial information *tx*, *ty*, *tw* and *th*, *objectness scores*, and *class scores*.

$$tx = \frac{absolute\_x}{image\_width} \tag{1}$$

$$ty = \frac{absolute\_y}{image\_height} \tag{2}$$

$$tw = \frac{absolute\_width}{image\_width} \tag{3}$$

$$th = \frac{absolute\_height}{image\_height} \tag{4}$$

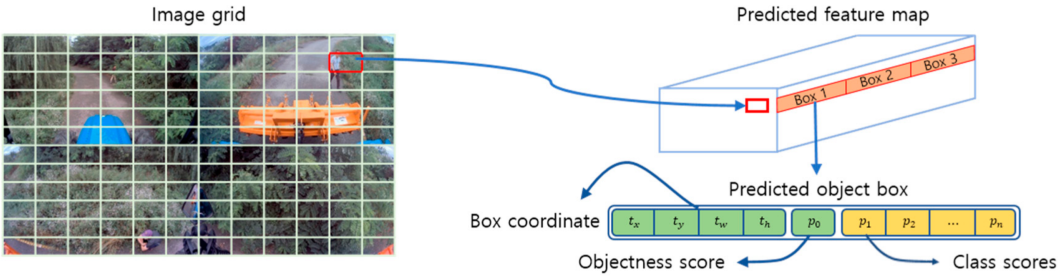


Figure 4. Attributes of YOLO-v3 prediction feature map.

The location information of the bounding box *tx*, *ty* is expressed in Equations (1) and (2). Size information *tw*, *th* is expressed in Equations (3) and (4). These parameters are expressed as floating point values in the range (0.0, 1.0] as the relative coordinate values of the object for the resolution of the input image, where *absolute_x*, *absolute_y* is the pixel coordinates of the upper-left corner point, not the center coordinates of the object.

$$objectness\ score = P_r(object) \cdot IoU \tag{5}$$

$P_r(object)$ is a measure of the probability that the bounding box contains an object. Intersection over union (*IoU*) is used to evaluate the object detection algorithm. It is the overlap between the ground truth and the predicted bounding box, i.e., it calculates how similar the predicted box is to the ground truth. The *objectness score* is defined as $P_r(object) \cdot IoU$. If no object exists in the cell, the *objectness score* is zero. Otherwise, the *objectness score* is equal to the intersection over union between the predicted box and the ground truth.

$$conditional\ class\ probability = P_r(class_i \mid object) \tag{6}$$

$$class\ scores = P_r(class_i) \cdot IoU = objectness\ score\ \times conditional\ class\ probability \tag{7}$$

Finally, the confidence prediction represents the *IoU* between the predicted box and any ground truth box. Each grid cell also predicts conditional class probabilities, $P_r(class_i \mid object)$ Equation (6). These probabilities are conditioned on the grid cell containing an object. We only predict one set of class probabilities per grid cell, regardless of the number of boxes.

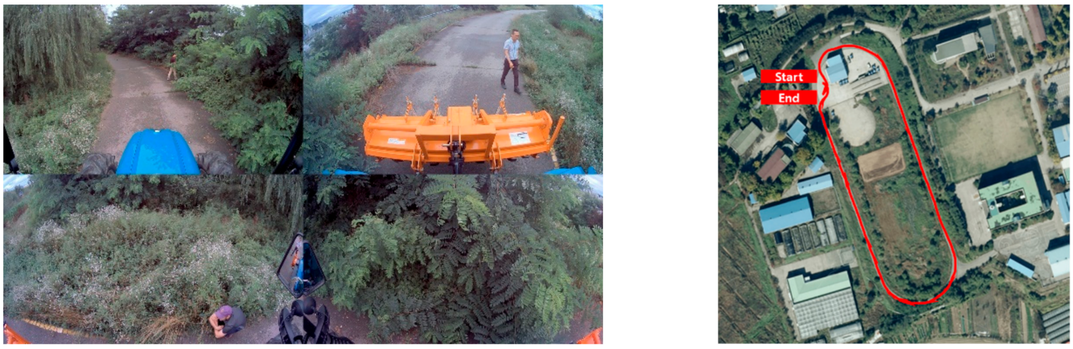
At test time we multiply the conditional class probabilities and the individual box objecteness scores, Equation (7), which gives us class-specific confidence scores for each box. These scores encode both the probability of that class appearing in the box and how well the predicted box fits the object.

In the proposed system, processing is performed to estimate which direction a person exists from the tractor using the spatial information of a person object output from the YOLO-v3 algorithm. Also, information on how many people exist at a given time can also be output. The information is finally sent to the controller of the autonomous tractor, and if a person is detected, a warning is provided to prevent an accident.

3. Results

3.1. Data Acquisition

Figure 5a shows an example of acquired data. As shown in Figure 5b, the acquisition of driving data was carried out at the LSMtron tractor driving test field in Hwaseong, South Korea.



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

Figure 5. (**a**) Example of acquisition data; (**b**) Data acquisition paths.

The image data were acquired at a storage rate of 5 fps. Faster data acquisitions were deemed unnecessary because of minimal changes between image frames due to the tractors slow operating speed of 3 to 4 km/h. In the process of acquiring the data, actors mimicked various agricultural worker postures, as well as positions that covered parts of the body with crops or other obstacles. A total of 8602 image frames were collected in total.

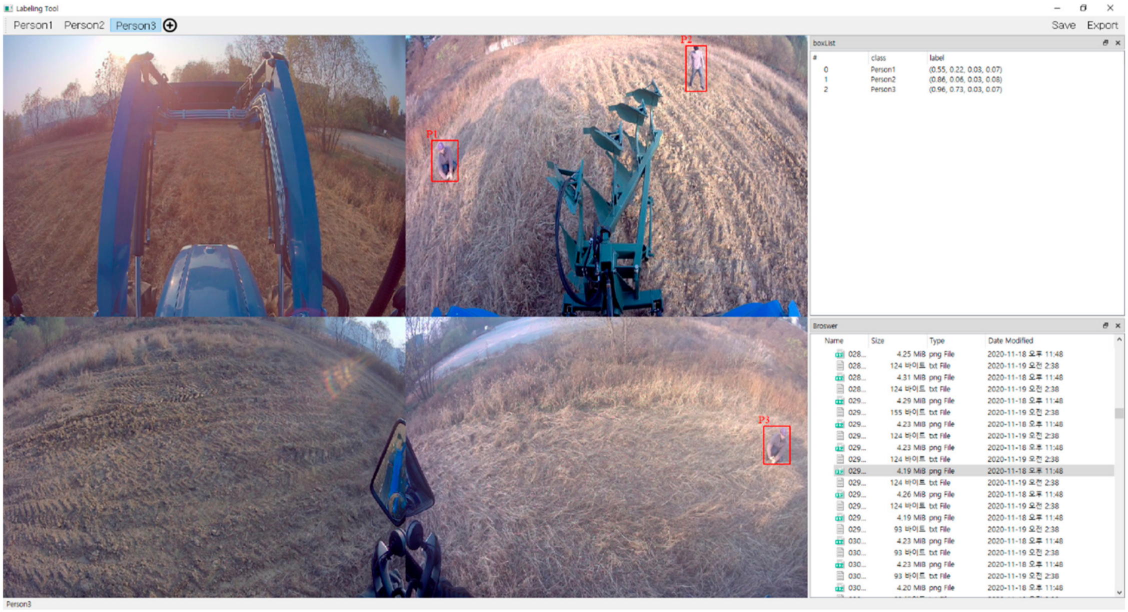
Using the self-developed annotation program introduced earlier, labelling work on person objects was carried out manually, and a total of 32,481 person object labels were obtained.

3.2. Data Annotation

The YOLO-v3 object recognition algorithm is a part of supervised learning. Therefore, for network learning, each learning image should be assigned a ground truth for the object to be recognized. The answer label consists of a pair of class labels for each object and a bounding box, called data annotation.

There are numerous algorithms for object recognition, and the annotations that each requires vary. The annotation format used by YOLO exists as a text file with the same name as the learning image file, and each text file contains labels for object classes, object coordinates, width, and height. Examples of tools that can be labeled according to the YOLO annotation format are Labelimg [45] and Yolo_mark. There is a limitation that these labeling tools can only be used for learning YOLO-based object recognition algorithms. Thus, a PyQT-based labeling tool was developed and used in this paper.

Figure 6 shows the execution screen of the developed labeling program. You can call up image or video data collected for learning and mark objects according to the class type that is color coded. In the case of video data, it can be automatically extracted on a frame-by-frame basis and saved as an image. The add-on features include zooming and scaling functions to enable precise labeling, and are designed to be easily adaptable to a variety of formats, not only the YOLO annotation format, enabling advantages in terms of scalability.



boxList

#	class	label
0	Person1	(0.55, 0.22, 0.03, 0.07)
1	Person2	(0.86, 0.06, 0.03, 0.08)
2	Person3	(0.96, 0.73, 0.03, 0.07)

Broswer

Name	Size	Type	Date Modified
028...	4.25 MiB	png File	2020-11-18 오후 11:48
028...	124 바이트	txt File	2020-11-19 오전 2:38
028...	4.31 MiB	png File	2020-11-18 오후 11:48
028...	124 바이트	txt File	2020-11-19 오전 2:38
029...	4.29 MiB	png File	2020-11-18 오후 11:48
029...	155 바이트	txt File	2020-11-19 오전 2:38
029...	4.23 MiB	png File	2020-11-18 오후 11:48
029...	124 바이트	txt File	2020-11-19 오전 2:38
029...	4.23 MiB	png File	2020-11-18 오후 11:48
029...	124 바이트	txt File	2020-11-19 오전 2:38
029...	4.19 MiB	png File	2020-11-18 오후 11:48
029...	124 바이트	txt File	2020-11-19 오전 2:38
029...	4.26 MiB	png File	2020-11-18 오후 11:48
029...	124 바이트	txt File	2020-11-19 오전 2:38
029...	4.19 MiB	png File	2020-11-18 오후 11:48
029...	93 바이트	txt File	2020-11-19 오전 2:38
030...	4.23 MiB	png File	2020-11-18 오후 11:48
030...	93 바이트	txt File	2020-11-19 오전 2:38
030...	4.23 MiB	png File	2020-11-18 오후 11:48
030...	93 바이트	txt File	2020-11-19 오전 2:38
030...	4.20 MiB	png File	2020-11-18 오후 11:48

Figure 6. LS Labeling Tool.

3.3. Training Data

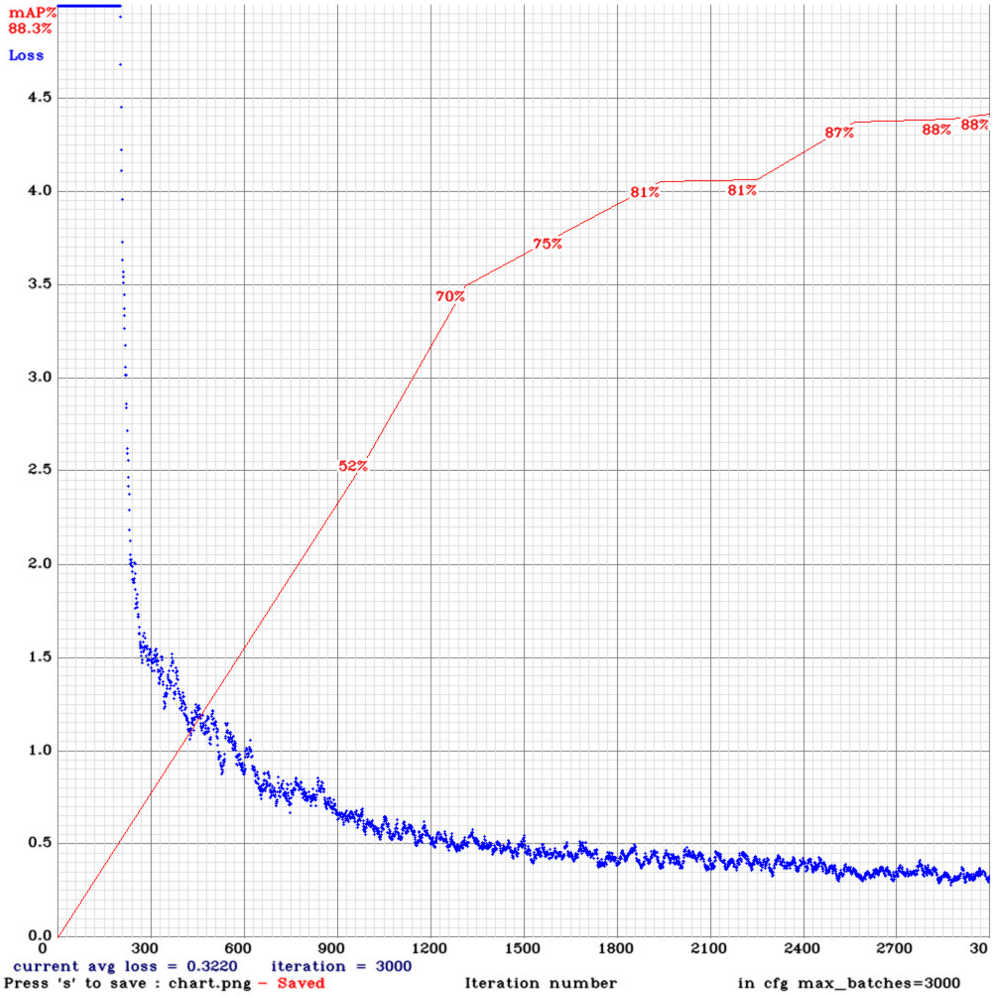
This paper focuses on increasing the safety of autonomous tractors by recognizing person objects. High-accuracy person object recognition requires a network learning process for various human images. The proposed system uses the MS-COCO dataset [46] for YOLO-v3 network learning. The MS-COCO dataset provides learning data, test data and validation data, such as object recognition, instance segmentation, key-point detection, and labeled annotation data.

The data for object recognition in the MS-COCO dataset includes labels with 80 object categories, and the proposed system used COCOapi to perform the extraction process for person objects only.

For the YOLO-v3 algorithm to recognize only human objects, deep learning neural networks were constructed based on Darknet-53. Darknet-53 is a newly developed network that adds a residual network to the existing Darknet-19 network, and is designed to have a total of 53 convolution layers with 3 × 3 and 1 × 1 configured to repeat with shortcut connections.

Hyperparameters set during learning were 0.0005 for the learning rate and 3000 for the max batch, and the number of classes and filters were changed to recognize a single object to proceed with learning.

Figure 7 shows the loss values (Loss) and *mAP* obtained in the course of learning only human objects in the MS-COCO dataset. After 1500 iterations, the loss value below 0.5 was attained and the mAP value of 88.3% was obtained on average from the final 3000 iterations.



current avg loss = 0.3220 iteration = 3000
Press 's' to save : chart.png - Saved
in cfg max_batches=3000

Figure 7. Training performance for person detection using YOLO-v3.

The mean average precision (*mAP*) is averaged over the $p(o)$ accuracy by Equation (8). $p(o)$ is the precision of the person detection.

$$mAP = \int_0^1 p(o)do \tag{8}$$

4. Discussion

For the test, a total of 8602 test datasets were obtained at LSMtron driving test fields, and it included a total of 32,481 person objects. We compared the person recognition performance of our YOLO-v3 (only person customized) with standard YOLO-v3 (include 80 classes) for the same test dataset.

The standard YOLO-v3 dataset classifies 80 different objects, but of those 80 objects, only person objects and cow objects are applicable to the tractor during agricultural work. The purpose of this study was to detect agricultural workers in proximity to the tractor, so our YOLO algorithm was trained using only one class, the person class. Our YOLO-v3 algorithm demonstrated a higher precision and a faster FPS when compared to the standard YOLO-v3.

To evaluation the validity of the model using the confusion matrix. Table 2 shows the confusion matrix for evaluating the person recognition model [47]. True indicates a person, while false indicates a non-person.

Table 2. Confusion matrix for evaluating the person recognition.

-		**Predicted Label**	
		True	False
True Label	True	TP (True Positive)	FN (False Negative)
	False	FP (False Positive)	TN (True Negative)

TP represents the number of accurately classified human objects, *FN* describes a non-human object being classified as a human object, *FP* occurs when a human object is not classified as such, while *TN* represents non-human objects correctly classified as non-human. Table 3 shows the person recognition performance of the actual driving environment displayed in a confusion matrix. The confusion matrix is a table for comparing forecasts and actual values to measure the performance of an object recognition model.

Table 3. Confusion matrix for person recognition.

-		**Predicted Label**	
		Person	Non-Person
True Label	Person	27,996 (0.86)	4485 (0.14)
	Non-Person	3372 (0.11)	29,109 (0.89)

Based on the confusion matrix, the precision and recall were calculated. Precision and Recall are illustrated by Equations (9) and (10).

$$Precision\ (P) = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)} \tag{9}$$

$$Recall\ (R) = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)} \tag{10}$$

TP represents true positives. FP is a measure of negative samples incorrectly classified as a positive sample. FN stands for positive samples incorrectly misclassified as negatives.

According to Table 4, our YOLO-v3 calculated 86.19% precision and 88.43% recall value by confusion matrix. The average processing time of 15.7 fps was verified as a result of algorithm execution for the test datasets by NVIDIA Jetson Xavier platform. By restricting the algorithm to learn only from a single object, persons, we have identified a 0.71% accuracy improvement and our YOLO-v3 is 2.3 fps faster than the standard YOLO-v3 using the same hardware and same test datasets.

Table 4. Performance comparison of standard YOLO-v3 and Our YOLO-v3 on test dataset.

	Precision	**Recall**	**FPS (Average)**
Our YOLO-v3	86.19%	88.43%	15.7
YOLO-v3	85.48%	88.52%	13.4

Figure 8 shows the output of the YOLO-v3 based person recognition algorithm for four-channel camera input images with a 1920 × 1088 resolution. The tests confirm that even when the camera's view of a person is partially obscured, a person is recognized as such with high accuracy.

Considering that the average working speed of autonomous tractors is 3~4 km/h, this image processing speed was judged to be sufficient to respond in real time.

Figure 9 shows an example of misdetection in the person recognition system. Misdetection was mainly shown at the boundary of a four-channel image, and there were cases where the side mirrors or exhaust pipes of a tractor were mistaken for persons.

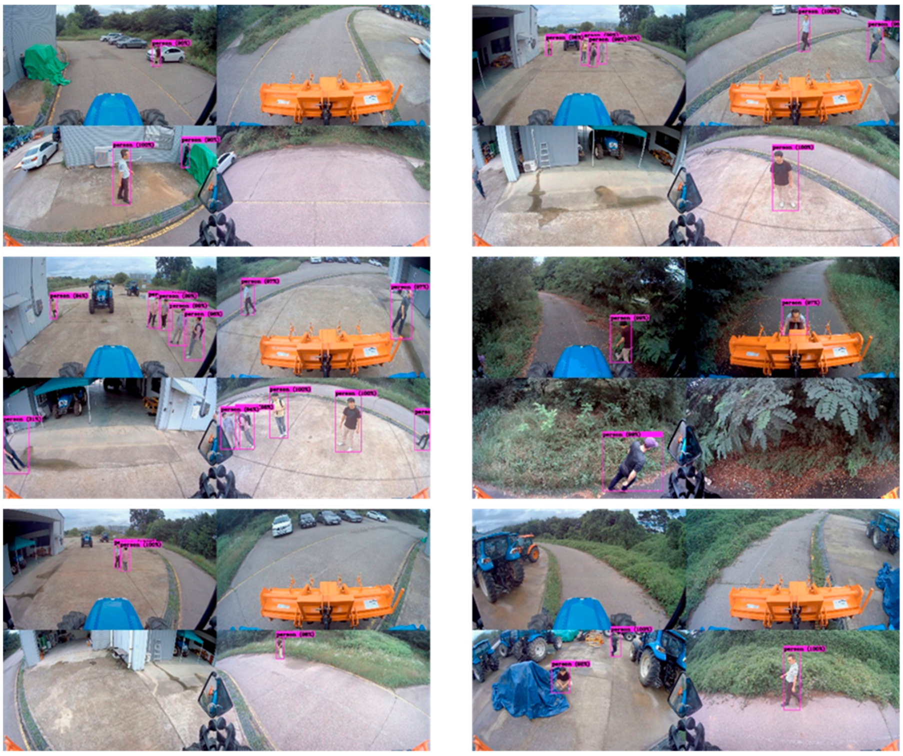


Figure 8. Person recognition test for autonomous tractor.

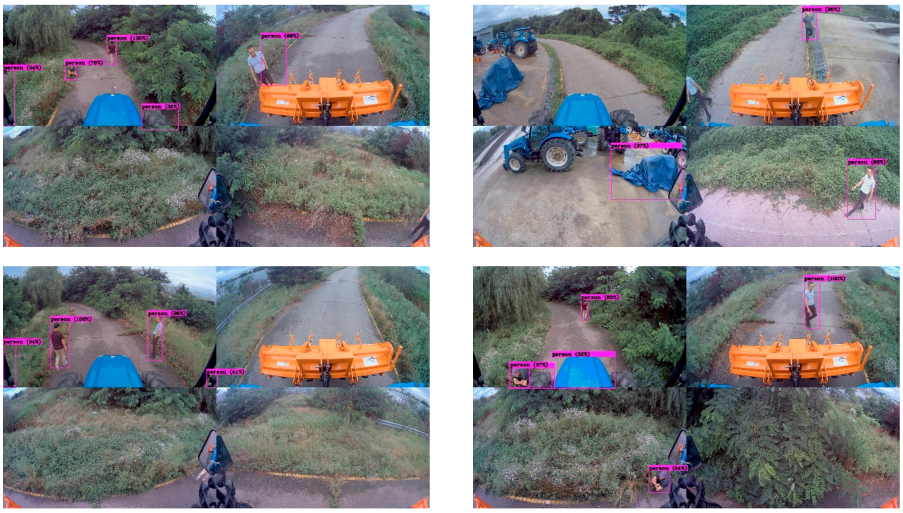


Figure 9. Example of misdetection result image.

5. Conclusions

In this paper, a recognition system for person objects appearing in the environment around autonomous tractors was proposed using cameras. Four wide-angle cameras with 120° viewing angle were used to detect the entire environment (360°) around the autonomous tractor, effectively minimizing blind spots to detect people. Frame synchronization and preprocessing were performed for the images entered from each camera, and four-channel split images were matched and used as an input to an object recognition algorithm. The object recognition algorithm used the YOLO-v3 model, and learning was performed using only person objects from the MS-COCO dataset. To evaluate recognition performance, actual autonomous tractor driving images were saved and evaluated, resulting in 88.43% precision, 86.19% recall, and an average processing speed of 15.7 FPS. This recognition performance is judged to be sufficient for the working environment of autonomous tractors. However, there remain problems that can reduce the efficiency of autonomous tractors, particularly because false positives sometimes appear, even though a person does not actually exist.

In the future, in order to overcome these problems, we need to conduct research to improve recognition performance by using learning data based on actual driving data in the field. We need to apply model lightening techniques, as well as deep learning acceleration techniques that enhance processing speed.

A limitation exists that our camera system was only tested in one type of working environment. In future studies we plan to evaluate the performance in more challenging agricultural environments. Additionally, while our system detected partially obscured human objects satisfactorily, our system cannot detect human objects if the camera lens is completely physically obscured. Additionally, camera sensors alone make it difficult to estimate the exact distance to an object. A fusion of sensor technologies could help detect objects in these cases.

Commercialization of autonomous tractors must be done to solve the problem of an aging population and decreasing labor in rural areas, and developing advanced technology for safety is a task that must be continuously solved in the future.

Author Contributions: Design framework, T.-H.J. and I.-K.C.; Data Acquisition, T.-H.J., I.-K.C., B.C., S.-H.L., and J.-M.C.; Mechanical Design, S.-H.L.; Data Annotation, T.-H.J., I.-K.C., B.C., S.-H.L., and J.-M.C.; Funding Acquisition, J.-M.C.; Writing-original draft, T.-H.J.; Writing-review and editing, I.-K.C., B.C., and J.-M.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Rural Development Administration, Korea, (Project Number: PJ013938022020).

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

References

1. Brown, M. Smart farming—Automated and Connected Agriculture. Engineering.com, 2018. Available online: https://www.engineering.com/DesignerEdge/DesignerEdgeArticles/ArticleID/16653/Smart-FarmingAutomated-and-Connected-Agriculture.aspx (accessed on 13 November 2020).
2. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
3. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
4. Viola, P.; Jones, M.J.C. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8–14 December 2001; Volume 1, p. 3.
5. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
6. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef]
7. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
9. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, Anchorage, AK, USA, 4–9 May 1998; pp. 2278–2324. [CrossRef]

10. Bargoti, S.; Underwood, J. Deep fruit detection in orchards. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, Singapore, 29 May–3 June 2017; pp. 3626–3633. [CrossRef]
11. Kragh, M.; Christiansen, P.; Laursen, M.S.; Steen, K.A.; Green, O.; Karstoft, H.; Jørgensen, R.N. FieldSAFE: Dataset for Obstacle Detection in Agriculture. *Sensors* **2017**, *17*, 2579. [CrossRef] [PubMed]
12. Zheng, Y.-Y.; Kong, J.-L.; Jin, X.-B.; Wang, X.-Y.; Zuo, M. CropDeep: The Crop Vision Dataset for Deep-Learning-Based Classification and Detection in Precision Agriculture. *Sensors* **2019**, *19*, 1058. [CrossRef] [PubMed]
13. Yang, Z.; Yu, W.; Liang, P.; Guo, H.; Xia, L.; Zhang, F.; Ma, Y.; Ma, J. Deep transfer learning for military object recognition under small training set condition. *Neural Comput. Appl.* **2018**, *31*, 6469–6478. [CrossRef]
14. Feng, C.; Liu, M.-Y.; Kao, C.-C.; Lee, T.-Y. Deep Active Learning for Civil Infrastructure Defect Detection and Classification. In *Computing in Civil Engineering 2017, Proceedings of the ASCE International Workshop on Computing in Civil Engineering 2017, Seattle, WA, USA, 25–27 June 2017*; American Society of Civil Engineers (ASCE): Reston, VA, USA, 2017; pp. 298–306.
15. Xu, J. A deep learning approach to building an intelligent video surveillance system. *Multimed. Tools Appl.* **2020**, 1–21. [CrossRef]
16. Liu, X.; Liu, W.; Mei, T.; Ma, H. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In *Proceedings of the Provable and Practical Security*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; pp. 869–884.
17. Wu, B.; Wan, A.; Iandola, F.; Jin, P.H.; Keutzer, K. SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 446–454.
18. Feng, D.; Haase-Schutz, C.; Rosenbaum, L.; Hertlein, H.; Glaser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, 1–20. [CrossRef]
19. Uçar, A.; Demir, Y.; Güzeliş, C. Object recognition and detection with deep learning for autonomous driving applications. *Simulation* **2017**, *93*, 759–769. [CrossRef]
20. Ferdowsi, A.; Challita, U.; Saad, W. Deep Learning for Reliable Mobile Edge Analytics in Intelligent Transportation Systems: An Overview. *IEEE Veh. Technol. Mag.* **2019**, *14*, 62–70. [CrossRef]
21. Tsai, C.-C.; Tseng, C.-K.; Tang, H.-C.; Guo, J.-I. Vehicle Detection and Classification based on Deep Neural Network for Intelligent Transportation Applications. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1605–1608. [CrossRef]
22. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
23. Wei, H.; Laszewski, M.; Kehtarnavaz, N. Deep Learning-Based Person Detection and Classification for Far Field Video Surveillance. In Proceedings of the 2018 IEEE 13th Dallas Circuits and Systems Conference, Dallas, TX, USA, 12 November 2018.
24. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
25. Wang, H.; Yu, Y.; Cai, Y.; Chen, X.; Chen, L.; Liu, Q. A Comparative Study of State-of-the-Art Deep Learning Algorithms for Vehicle Detection. *IEEE Intell. Transp. Syst. Mag.* **2019**, *11*, 82–95. [CrossRef]
26. Arabi, S.; Haghighat, A.; Sharma, A. A deep-learning-based computer vision solution for construction vehicle detection. *Comput. Aided Civ. Infrastruct. Eng.* **2020**, *35*, 753–767. [CrossRef]
27. Li, J.; Zhou, F.; Ye, T.; Li, J. Real-World Railway Traffic Detection Based on Faster Better Network. *IEEE Access* **2018**, *6*, 68730–68739. [CrossRef]
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
31. He, K.; Gkioxari, G.; Dollar, P.; Girshicket, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
32. Sermanet, P.; Eigen, D.; Zhang, Z.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv 1312.6229* **2013**.
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Proceedings European Conference on Computer Vision, Cham, Switzerland, 29 December 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.
34. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
35. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
36. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
37. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
38. van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv* **2018**, arXiv:1805.09512.
39. Qiu, Z.; Zhao, N.; Zhou, L.; Wang, M.; Yang, L.; Fang, H.; He, Y.; Liu, Y. Vision-Based Moving Obstacle Detection and Tracking in Paddy Field Using Improved Yolov3 and Deep SORT. *Sensors* **2020**, *20*, 4082. [CrossRef]
40. Ball, D.; Upcroft, B.; Wyeth, G.; Corke, P.; English, A.; Ross, P.; Patten, T.; Fitch, R.; Sukkarieh, S.; Bate, A. Vision-based Obstacle Detection and Navigation for an Agricultural Robot. *J. Field Robot.* **2016**, *33*, 1107–1130. [CrossRef]
41. Fleischmann, P.; Berns, K. A Stereo Vision Based Obstacle Detection System for Agricultural Applications. In *Springer Tracts in Advanced Robotics*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2016; pp. 217–231.
42. Ross, P.; English, A.; Ball, D.; Upcroft, B.; Wyeth, G.; Corke, P. Novelty-based visual obstacle detection in agriculture. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014; pp. 1699–1705. [CrossRef]
43. Hughes, C.; Glavin, M.; Jones, E.; Denny, P. Wide-angle camera technology for automotive applications: A review. *IET Intell. Transp. Syst.* **2009**, *3*, 19–31. [CrossRef]
44. Mittal, S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *J. Syst. Arch.* **2019**, *97*, 428–442. [CrossRef]
45. Tzutalin. "LabelImg". Git Code, 2015. Available online: https://github.com/tzutalin/labelImg (accessed on 6 March 2020).
46. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; pp. 740–755.
47. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.