International Journal of Advanced Technology and Engineering Exploration, Vol 9(97)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.876501

Research Article

Analysis on localization and prediction of depth chili fruits images using YOLOv5

M. N. Shah Zainudin[1*], M. S. S. Shahrul Azlan[1], L. L. Yin[1], W. H. Mohd Saad[1], M. I. Idris[1], Sufri Muhammad[2] and M. S. J. A. Razak[3]
Universiti Teknikal Malaysia Melaka, Faculty of Electronics and Computer Engineering, Hang Tuah Jaya, Durian Tunggal, Melaka, Malaysia[1]
Universiti Putra Malaysia, Faculty of Computer Science and Information System, UPM Serdang, Seri Kembangan, Malaysia[2]
MSJ Perwira Enterprise, Duyung, 75460 Melaka, Malaysia[3]

Received: 02-August-2022; Revised: 26-November-2022; Accepted: 28-November-2022

Abstract
Chili fruits are essential ingredients that Malaysians consider essential for cooking. Adding chili to a dish used to produce a second fiery flavour. In south-western Ecuador, one of the most important plant-growing regions on the American continent at the time. As a result, this provides evidence that people were using chili as an additional food element as early as 600 years ago. A traditional method of picking chili is common, but it is less precise and time-consuming. Incorrect picking and grading will cause the harvesting process to take longer. The advancement of computer vision and pattern recognition has demonstrated its effectiveness in image recognition. Because of their simplicity and low complexity, 2 dimensional images are frequently used in image recognition. As a result, advancement in automated picking systems with object detection is common. However, due to a lack of image information, such as depth, 2D images are thought to be difficult to identify the growing stages or maturity level of chili fruits. Object detection is prevalent for determining the localization and category of objects. One of the well-established methods such you look only once (YOLO) has widely used is in object detection. To anticipate this effort, the fast, reliable and able to recognize small object, YOLOv5 is proposed to localise and predict the category of chili fruits which allows the process to determine a chili's form and categories based on its colour. The proposed model is able to differentiate and localize the position as well as the colour of chili fruits with above 94% in average. Hence, our achievement has proven its effectiveness and becomes our greater goal of developing an autonomous chili fruit picking robot which could help farmers or agricultural sectors to reduce their labours during the grading process.

Keywords
Chili, YOLOv5, Semi-autonomous, Depth images, Object detection.

1.Introduction
Human food supplies are gradually increasing nowadays. To meet the world's increasing population's food demand, horticulture must find new ways to increase fruit and vegetable production [1]. Fruit harvesting is an essential part of the development and management of farmlands.

In order to address the issue of inefficient manual fruit harvesting, an intelligent and systematic, automatic harvesting robot has been developed in recent years. Traditional manual harvesting necessitates a large number of farmworkers, resulting in a high production cost [2].

In order to address the issue of inefficient manual fruit harvesting, an intelligent and systematic, automatic harvesting robot has been developed in recent years.

Horticulture must find new ways to increase fruit and vegetable production to keep up with the world's growing population [3]. Fruit harvesting is becoming

*Author for correspondence

This work funded by Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) under PJP/2020/FKEKK/PP/S01787 and INDUSTRI (MTUN)/ENDSTRUCT/ 2021/FKEKK/I00057 grant.

increasingly important as part of farmland development and management. Traditional manual harvesting requires a large number of farmworkers, resulting in a high production cost [4] and an increase in incorrect grading, particularly for small fruits like chili [5]. Furthermore, the characteristics of chili fruits have become a new challenge. Highly similar leaves, small sizes, and colour variations become critical for providing virtuous innovation. To address this issue, the invention of intelligent application exploration should be considered [5].

In recent years, an intelligent and systematic, automatic harvesting robot with artificial intelligence (AI) integration has been developed to address the issue of inefficient manual fruit harvesting.

One of the most common examples of a new innovation in the agriculture industry is the application of deep learning, an AI subfield that can learn from unstructured or unlabelled data and is built on the foundation of neural network architecture. The use of AI has significantly proven its ability to solve a variety of applications including human activity [6], industries [7], smart home [8], etc. In today's world, the use of AI or deep learning to solve various image processing and classification problems has become prominent. In agriculture, for example, You Look Only Once (YOLO) method is used to detect fruits such as tomatoes. In such cases, rather than using traditional sorting methods, the process of detecting its maturity, grades, and quality is highly efficient. As a result, this type of development should be pursued further in order to assist those industries in moving forward and competing with current technology.

Deep learning has proven to be a very useful technique due to its ability to handle large amounts of data. Hidden layer approaches, particularly in pattern recognition, have surpassed classical techniques in terms of popularity. Convolutional Neural Network (CNN) is one of the prominent deep neural network models [9-10]. CNN, for example, could recognise handwritten numbers, determine the type of cancer, recognise faces, and so on. It is primarily used in the postal industry to read handwritten zip codes, pin numbers, and other unique identifiers. However, deep learning model requires large amounts of dataset as well as the enormous amount of computational power to be trained. CNNs are a category of deep neural network that is commonly used in deep learning to interpret visual data. Deep learning has been demonstrated, its effectiveness in a variety of applications, including image and video recognition, image classification, image segmentation, medical image analysis, and natural language processing. CNN is a specialised multilayer perceptron.

The term "multilayer perceptron" refers to fully linked networks in which every neuron in one layer links to every neuron in above layer [7]. Because of their "full connectivity," these networks are prone to overfitting. A CNN receives input from a tensor of the form (number of heights × input inputs × input channels × input width). CNN has many exciting applications, including image classification. Object recognition, along with traditional image classification, is another intriguing challenge that computer vision. YOLO is an effective technique for real-time object detection [6]. YOLO proposes the use of an end-to-end neural network, which able to simultaneously predict bounding boxes and class probabilities, as opposed to previous object detection methods, which repurposed classifiers to do the detection. YOLO achieves cutting-edge object detection results by carrying a fundamentally different approach than existing real-time object detection algorithms [11] especially in agricultural industries.

Object detection and image recognition techniques are frequently combined with a wide range of applications, including agriculture, medicine [12], and manufacturing [13, 14]. Image recognition recognises the scenes or objects present in an image, whereas object detection identifies the instances and locations of such objects [15]. Image recognition can be used to automate these time-consuming processes, processing photos faster and more accurately than by hand [16]. Image recognition is a critical technique that is used in a wide range of applications to categorise images based on their properties. This is the driving force behind the development of AI systems such as deep learning. This is extremely beneficial in e-commerce applications such as recommender systems and image retrieval. Object detection has greatly evolved in the field of computer vision. It is one of the most challenging areas of computer vision to master because it involves both object classification and object localization. In layman's terms, the goal of this detection method is to identify the categories to which each object belongs and the location of each object within a given image, also known as object localization and object classification [17].

However, when it comes to exploring highly complex fruits like chilli or other small types of fruits, the

exploration of automated detection and classification is still in the early stages. Because of their small size and high similarity with their leaves, the detection and classification process become difficult [5]. In addition, another factor must be considered in order for this challenge to be relevant in a real-world setting. In order to incorporate AI inventions in the development of agricultural robots, the localization and the position of the fruits must be defined. Before the harvesting process can begin, the distance from the object (chilli fruits) must be calculated. As a result, the use of stereo images or depth images is critical in addressing this issue. As we all know, a single lens camera produces a 2-dimensional image in which every pixel is projected along two axes with values X and Y. However, multiple camera lenses produce additional values that represent the depth information of the images, where the distance between an image and the camera is also measured.

Depth cameras determine the distance of an object captured from a viewpoint by identifying the intensity of the image, providing information about the object such as shape, localization, classification, and distance in the real world [18]. The colour information in the depth image indicates the object's distance from the viewpoint. A variety of digital cameras output images as a 2-dimensional grid of pixels with x and y axes for information. Initially, each pixel in an image is assigned a value known as RGB—red, green, and blue. To indicate the colour code, the attribute value generated from each pixel ranges from 0 to 255; for example, pure bright red would be represented by the value (255, 0, 0) [19].

A depth camera, on the other hand, has an additional pixel value that is associated with a different numerical value. As additional information, the depth information, also known as the object's distance from the camera, is displayed. By combining an RGB colour space with a depth system (D), some depth cameras provide pixels with all four values, or Red, Green, Blue and Depth. The stereo depth camera has two sensors, one on each side, which are separated by a small amount. The distance between the two sensors is compared with the stereo depth camera. The two sensors in use are RGB and depth sensors. These sensors work by improving correspondence between the two different data streams and by aligning the RGB and depth sensors' fields of view. Since the distance between the two sensors is known, depth information is received [20].

There a few objectives to tackle aforementioned challenges in previous sections. 1) This work aims to localize the object (chili fruits in this case) using a well-known object detection method, YOLOv5. 2) The analysis is conducted by differentiating between green and red colours chili fruits. The experiments were conducted in two conditions: individual chili fruit detection (red and green chilies) and chili fruit detection by localising the position of the chili from the chili plants. As a contribution, this work has proven its effectiveness in localizing and detecting a chili fruit using the proposed model. The model is also able to differentiate between chili fruits and leaves from the plant images. This work is also capable of distinguishing between red and green colours chili. The structure of articles as follows: section 2 explains the literature review of previous related work, section 3 discusses the methodology proposed throughout this experiment, section 4 describes analysis of experimental result obtained, section 5 explains discussion of this entire experiment and section 6 discusses the conclusion and recommendation for future work.

2.Literature review

Plenty of work on fruit detection and classification using AI has been reported in recent years. Gongal et al. published their work on apple fruit size estimation in images. Using a charge-coupled device (CCD) camera and a time-of-flight (TOF) light-based 3-dimensional camera, they develop a machine vision system for estimating fruit size. The distance between all pixel pairs was calculated using 3D coordinates to implement the harvesting robot. However, due to the low resolution of 3D sensors, the segmentation and mapping procedures may be difficult [21]. Tian et al. also reported intelligent detection of apple fruits. The authors proposed an improvement to the YOLOv3 model in order to measure and observe the progression of apple growth stages. Because of the rapid changes in color, cluster density, and other characteristics, the author is motivated to carry out this analysis in order to address some other challenges such as changes in illumination, complex background, overlapping, and the appearances of branches and leaves. Dense Net is embedded in the ordinal YOLOv3 model, and the results are compared to the YOLOv3 and faster region based-convolutional neural network (Faster R-CNN) models using images with a resolution of 3000 x 3000 [22].

Plant disease detection is also important for improving and controlling crop yield production. Due

to a lack of sample images, an automated detection of apple diseases remains hazy. Tian et al. proposed a deep learning model called cycle-consistent adversarial network (CycleGAN) to interpolate images in order to increase the diversity of training data [23]. The proposed model in [21] is used to address the issue of low resolution in the original YOLOv3 model. As a result, the proposed model is capable of improving the effectiveness of detection models that are significantly more efficient than the Faster R-CNN model in identifying apple fruit diseases. Liu and Wang also reported on the early detection of diseases on fruits. The detection of diseases and pests is critical for controlling the growth process of tomatoes. Deep learning is proposed to address the shortcomings of traditional image processing models that require tedious steps such as pre-processing, feature extraction, and classification. This end-to-end structure is intended to improve the model's efficiency in expressing the object's attributes. To improve the speed and detection accuracy of the original YOLOv3 model, an image pyramid is proposed [24].

Kuznetsove et al. conducted an experiment on apple fruit detection to be used in the development of a fruit harvesting robot. The YOLOv3 model is used, and it is capable of recording a detection time of 19ms with an accuracy of error mistakenly detecting an apple of 7.8 percent. The error rate for unidentified apple has also been reported to be as low as 9.2 % [25]. However, this work only tackle on apple detection without differentiating its maturity stages. Lawal has proposed a YOLOv3 framework modification for automated tomato fruit detection. Due to environmental challenges such as branch and leaf occlusion, lighting variations, shading, and so on, the original YOLOv3 model will be unable to produce an outstanding performance. To reduce missed detection, the proposed dense architecture with mish activation and spatial pyramid pooling (SPP) is combined with the YOLOv3 model. The proposed architecture has a precision of more than 99 % and a high generalization ability [26]. Even though the high accuracy has obtained, the author has not taken into account in distinguishing the maturity level of tomatoes.

Fu et al. also created a fruit detection model using the YOLOv3 model. The YOLOv3 model has been improved due to the color similarity and illuminations of kiwi fruits. A deep YOLOv3-tiny (DY3TNet) model is proposed to address the problem while reducing model complexity with high precision. The images of kiwi fruits are captured in various illuminations and compared to other models such as YOLOv2 and Faster R-CNN. The experimental analysis can record more than 90% precision with an average of 34ms [27]. Due to texture of kiwi fruits, the image must be captured using flash in order to enhance its performance. Yao et al. have also investigated the same type of fruits. The new version of YOLOv5 is proposed to identify and detect kiwi fruit defects in real-time. Because the fruits are too small and difficult to identify, YOLOv5 model is used. In terms of speed and ability to detect small objects, this latest model outperforms the previous version of the YOLO model [28]. Yet, this reported work only detect kiwi fruits defect on its texture without detecting the fruits from the plant. Kuznetsova et al. also evaluated the performance of two YOLO models (YOLOv3 and YOLOv5) for apple detection under two conditions: general images and close-up images. As a result, original YOLOv5 outperforms version 3, where the model can detect apples precisely with a false positive rate (FPR) of 3.5 % [29]. However, the maturity or fruits grading has not included in their work.

Manan et al. has reported their experiment on classification of chili plant growth using transfer learning. To differentiate between chili and leaves, CNN has proposed. YOLOv4 Darknet model has been applied and detect the fruits from the plant images. YOLOv4 has achieved mAP 75%, which outperform other two models; Faster R-CNN and EfficientDent [30]. However, experiment on has been done for single color chili fruits. Hespeler et al. reported an experiment for robotic inspection and harvesting chili peppers. Due to low classification scores with heavy debris present in images, an improvement on pre-processing has proposed. Thermal images were used to extract more significant features for object detection and able to recognize the fruits with variant lighting and overlapping [31]. Only one class of fruits is been evaluated and could lead into overfitting. Sihombing et al. carried out experiment on chili classification using machine learning. Two features have been extracted; shape and color to categorize the fruits into five classes; cayenne pepper, green chili, big green chili, big red chili and curly chili. High accuracy has obtained in average of above 90% [32]. Yet, the experiment only been done for individual chili fruit images. Classification of dried chili pepper has also reported [33]. Dried chili pepper images have captured to measure the quality by detecting their defects. Artificial neural network (ANN) has applied to

classify the quality of dried chili pepper and able to obtain above 80% of performance. Even though a bundle of works has been reported to recognize and classify the fruits, there is too few works found in detecting the chili categories due to the challenge as stated in the earlier section.

3.Methodology

This work begins with the collecting of 3D images with a stereo camera, followed by pre-processing stages such as image alignment, depth information extraction, and image acquisition as depicted in *Figure 1*. The next step is to perform image labelling, which involves labelling an object from an image (chili fruits) before beginning the training process. Following that, the collected dataset is separated into two groups of subsets (training and testing/validation). This procedure is required to assess the training model's ability to predict and classify the new instance.

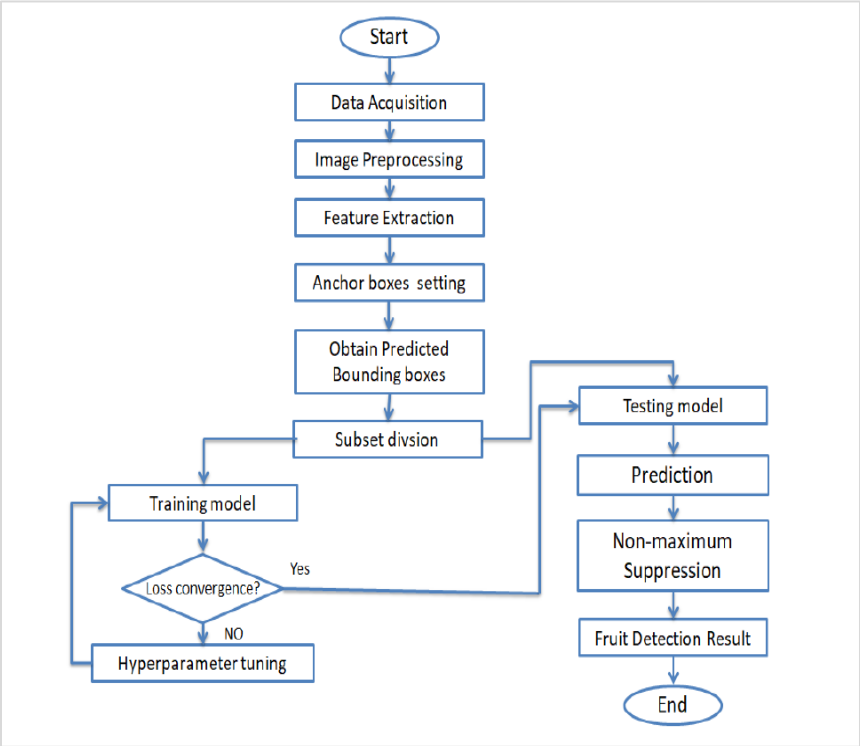

Figure 1 Methodology of the proposed work

3.1Data acquisition

Due to the difficulty in locating real farms to conduct our experiment, an artificial chilli is used as the dataset in this work. It is difficult to conduct an experiment in real-world conditions in Malaysia due to the country's unstable weather. As we all know, planting chilli fruits is difficult due to a variety of factors such as pest (fruit flies) and disease attack, insufficient pesticides, unstructured fertiliser, soil type, and so on [34]. As a result, we conduct our experiment with an artificial chilli. To increase diversity of the sample, image is captured in various positions. In this work, we conduct two sets of experiments, whereas differentiate between red and green individual chili images and to detect and locate the chili fruits from the chili plant images. An individual chilli and plant images are captured from various viewing angles.

3.2Image processing and labelling

An object must be labelled in order for the model to learn the characteristic of an object in object detection. Detection accuracy is highly reliant on a number of labelled objects. The labelling process allows you to draw visual boxes around an image or video using the annotation tools. In this part, labelling was done for both individual chili fruits and chili fruits from the plant. For both chilli colours images, annotation process is implemented using captured images as shown in *Figure 2*.

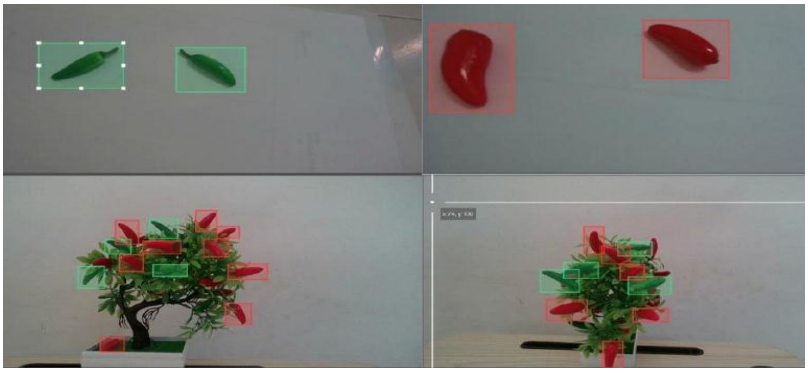

Figure 2 Labelling of red and green chili

3.3Feature extraction

The YOLOv5 network is divided into three sections as depicted in *Figure 3*: the backbone, the neck, and the head [35]. YOLOv5 built CSPDarknet as the Darknet's backbone by incorporating a cross stage partial network (CSPNet). The data is first fed into CSPDarknet for extracting features, then fed into PANet for data to be fused. Finally, detection result is obtained by YOLO layer outputs. The CSPNet used in YOLOv5 can address some issues; repeated gradient information in large-scale backbones by incorporating gradient changes into convolution layers, reducing model parameters and FLOPS (floating-point operations per second). This not only improves inference speed and accuracy, but also could reduce model size. YOLOv5 used PANet as a bottleneck to increase data throughput. To improve low-level feature propagation, a new feature pyramid network (FPN) topology with an improved bottom-up approach is employed into PANet. Furthermore, PANet is capable to improve the use of precise localization signals in lower layers while at the same time able to increase object location accuracy. The YOLO layer generates three different sizes of convolution layers (18×18, 36×36, and 72×72) to enable multi-scale prediction. The YOLO layer enables YOLOv5 to handle various sizes of object, including small, medium, and large.

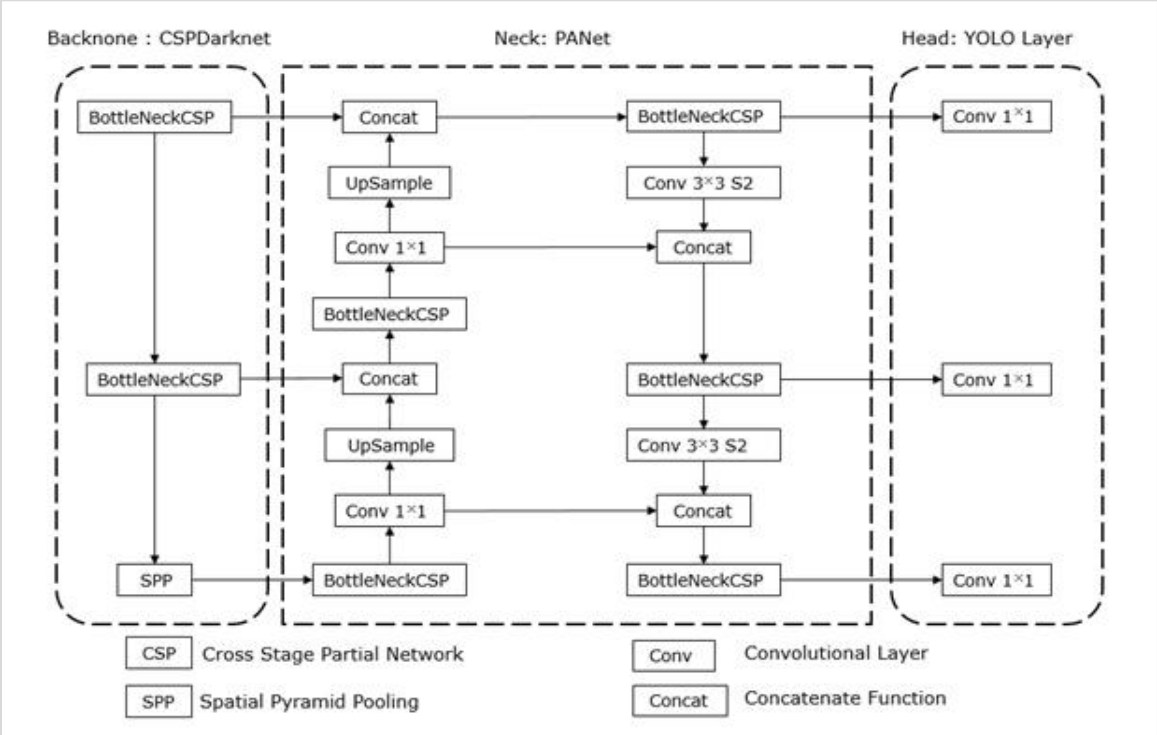

Figure 3 YOLOv5 network architecture

3.4Anchor boxes

Anchor boxes, as shown in *Figure 4*, are boundary boxes with a fixed height and width. K-means and genetic learning algorithm uses to generate learning anchor boxes by analysing distribution of bounding box in the dataset. For instance, when the distribution of bounding box size and location of the common objects in context (COCO) dataset significantly differ from predefined bounding box anchors, this makes the process difficult for custom tasks. YOLOv5 automatically learns all YOLO anchor boxes when custom data is entered. With their centres in the small cell, the anchor boxes are used in fruit detection to recognise a variety of things [36].

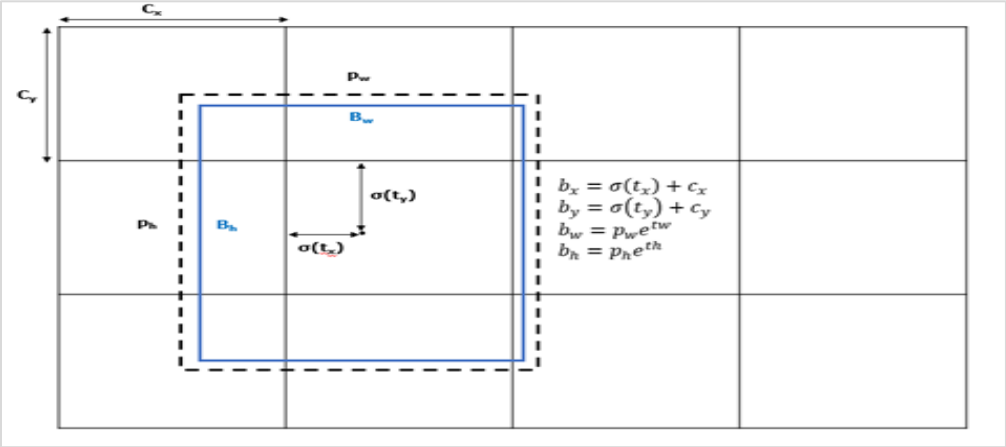

Figure 4 Anchor boxes

3.5Bounding box prediction

For each bounding box, YOLOv5 network predicts four coordinates: t_x, t_y, t_w, and t_h as shown in *Figure 5*. The cell is offset (c_x, c_y) from the top left corner of the image and the bounding box prior width and height (p_w, p_h), then predictions as (Equation 1-Equation 5).

$$b_x = \sigma(t_x) + c_x \quad (1)$$
$$b_y = \sigma(t_y) + c_y \quad (2)$$
$$b_w = p_w e^{t_w} \quad (3)$$
$$b_h = p_h e^{t_h} \quad (4)$$

YOLOv5 uses logistic regression to predict an object score for each bounding box. The values are 1 if the bounding box prior overlaps a ground truth objects by more than any other bounding box prior. Otherwise, the prediction is ignored when the model incapable to provide the best prior but overlaps a ground truth object by more than a certain threshold value. In order to validate the performance of classification, entire subset is divided into two portions of sub-subsets; training and testing. 80% from the sample used for training while 20% is allocated for testing.

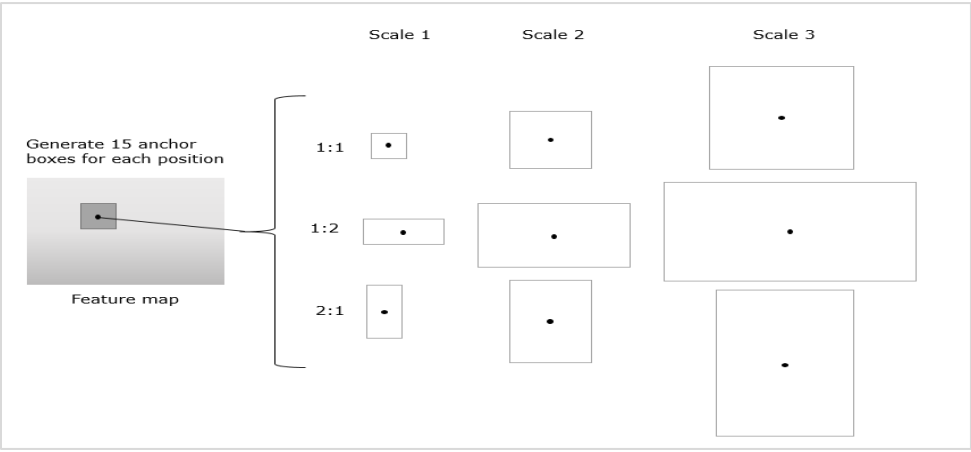

Figure 5 Bounding boxes with dimension priors and location prediction

3.6Object detection and classification

Image or object detection is a computer technology that processes images to detect objects in the same way that human brains do. There are a few distinctions between object detection and image classification. If we want to classify an image as an item, we use classification, whereas image detection is used to locate the item by detecting an object or the number of objects in the image. As an example, we want to know if there are any books in a single image. There could be several items such as a table, books, a cup, and so on. As a result, image detection will learn and discover whether or not a book is identified. AI performs image detection and classification by detecting an object in a labelled image and identifying its coordinates, class labels, and location. When a different location of the object is selected, the coordinate and size will change. To ensure that the class is correctly identified, AI will learn the pattern and characteristics of the image by referring to the labelled image. Object detection classification works by categorising a detection object with the labelled class. CNN is one example of a method for classifying an object [9, 17]. It employs a filter to learn and train image features. A number of features are relying on the type and total number of filters used. Technically, the features are generated in order to analyse an object after training.

3.7YOLO

YOLO, or You Only Look Once, is a well-known method for identifying various 'things' in real-time. CNN is used in YOLO to achieve real-time object detection by identifying an object in forward propagation and providing object class probabilities [37]. A single neural network is used in YOLOv1 to predict bounding boxes and class probabilities from an entire image [38]. When attempting to predict all possible bounding boxes, the network draws features from an image of its classes at the same time. However, the first version of YOLO was unable to detect small objects when images were grouped together in the same way. When the image dimension differs from the trained image, this early version of YOLO has difficulty generalising the objects within the image. The YOLO9000 or YOLOv2 is the second generation of the YOLO model. Because of its ability to identify over 9000 objects [37], it outperforms YOLOv1 in a variety of ways, such as use of CSPDarknet-19 as a backbone, batch normalisation and the use of a high-resolution classifier, fine-grained features, multi-scale training, and the use of anchor boxes to predict bounding boxes, among others. The introduction of the anchor boxes is one of the most noticeable changes in YOLOv2. Based on the information contained in these anchor boxes, the bounding box can be predicted. In comparison to its predecessor, YOLOv2 performs significantly better in terms of detecting smaller objects with greater precision [38].

In terms of speed, YOLOv3 outperforms YOLO and YOLOv2 by incorporating CSPDarknet-53 as a backbone for the feature extractor [39]. It operates at orders of magnitude faster speeds than other detection methods while retaining comparable performance for various applications including fruits recognition [36–41, 25–27]. In terms of average precision (AP), YOLOv3 can detect objects that are small, medium, and large. YOLOv4 significantly outperforms the other versions in terms of detection performance as well as speed [42]. YOLOv4 architecture comprising CSPDarknet53 backbone, spatial pyramid pooling (SPP) and PANet path-aggregation neck. CSPDarknet-53 is a backbone with the potential to improve CNN's learning capacity. In order to broaden the receptive field and distinguish the most important context features, SPP block is added on top of the CSPDarknet-53 block [43]. Planet is used for parameter aggregation for various detector levels rather than FPN, which were previously used in YOLOv3.

YOLOv5 is the most advanced object detection algorithm, capable of recognising objects with high accuracy [44]. A single neural network is used and divided into component pieces, bounding boxes, and probabilities for each piece are projected. To increase the prediction probability, a weight is assigned to each bounding box. In the sense of predictions are made after only one forward propagation passes through the neural network, the approach called as "just looks at the picture once". In other words, the method "only looks at the image once". It then delivers the items that were detected after non-max suppression (to make sure the object detection algorithm only identifies each object once). The memory-related YOLOv5 has both positive and negative connotations. 88 % more compact than the YOLOv4 and 180 % faster than the YOLOv4. YOLOv5 has a frame rate per second (FPS) of 140, while YOLOv4 only has an FPS of 50 [45–46].

3.7.1IoU and loss function

During the detection process, an anchor box is generated and matched. The centroid of the object is used to compute clusters, and the highest overlapping clusters are divided by non-overlapping for each anchor box. This is referred to as intersection over

union (IoU). The IOU is used to calculate object precision by comparing expected the bounding box to actual bounding box as shown in Equation 5.

$$\text{IoU} = \frac{\text{area(box(predicted)} \cap \text{box(truth)})}{\text{area(box(predicted)} \cup \text{box(truth)})} \quad (5)$$

The IOU is a normalised index with values ranging from 0 to 1. Object detection occurs when the value of the IOU exceeds 50%. Otherwise, no object is detected.

In order to increase object detection accuracy, the loss function is measured. The loss is estimated by YOLOv5 using bounding box positioning error, confidence error, and classification error. Loss is calculated by, Loss = L_{box} + L_{cls} + L_{obj} where L_{box} is bounding box positioning error, L_{cls} represents a confidence error and L_{obj} estimates the classification error. The difference between predicted and actual anchor box coordinates causes positioning error. The cross entropy of the probability target frame is used to calculate the confidence error. When the bounding box detects a target in the current box, the classification error is computed. Furthermore, as stated in Equation (6) to (8), the analysis of experiment is also measured by two additional parameters: recall and mAP (6).

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive+false positive}} \quad (6)$$

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive+false negative}} \quad (7)$$

$$\text{mAP} = \frac{1}{c}\sum_{k=i}^{N} p(k)\Delta R(k) \quad (8)$$

$C = \text{number of object categories}$

$N = \text{number of IoU threshold}$

$k = \text{IoU threshold}$

$p(k) = \text{precision}$

$R(k) = \text{recall}$

4.Results

This result focuses on two types of images: RGB images and depth images. Depth images of chili plants have only been used to determine an object from captured images. To localise the chili position, RGB images are used to label the chili. Two types of chilies are used in this work: green and red. Each object must be labelled before proceeding with the object localization process. Following the completion of the labelling process, the labelled images are divided into two subsets: training and validation. Another subset size has been set aside for testing. The testing subset is used to evaluate the model and measure the chili's localization accuracy based on its type.

4.1 Stereo camera images

We use stereo depth camera by Intel RealSense D455 model. The Intel RealSense Software Development Kit (SDK) 2.0 is required for the process of recording RGB and depth images. *Figure 6* depicts an example of an object captured by depth cameras. The images on the left represent the RGB version of the images, while the images on the right show the same image with depth information.

Figure 6 RGB image (left) and depth image (right)

4.2 Experimental result of chili prediction on single images and multiple colour images

Green chilies are labelled as 'green,' while red chilies are labelled as 'red.' YoLov5 is used to train the model. There are a few parameters that must be defined for this experiment. In this case, we used 16 batches with a total of 50 epochs with Stochastic Gradient Descent as an optimizer. We are unable to

evaluate the experiment with a large number of epochs due to limited resources. Hence, we used 213 model's layers is generated to evaluate a total of 56 sample images. The experimental analysis results of the chili detection experiment are shown in *Figure 7* and *8.*

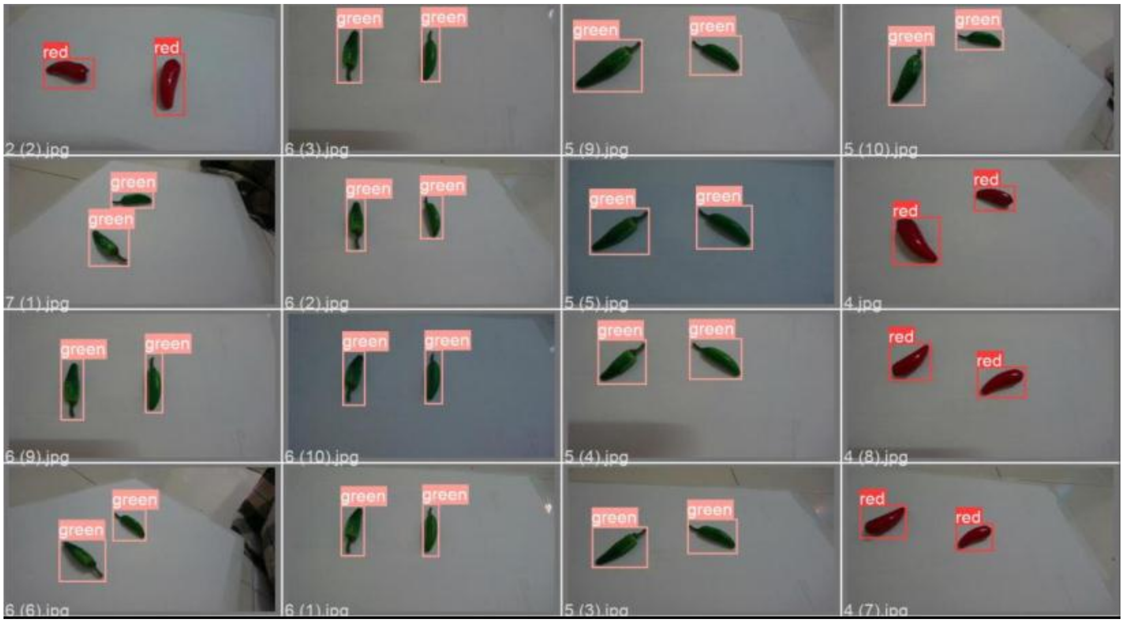

Figure 7 Training sample for both chili colours (green and red)

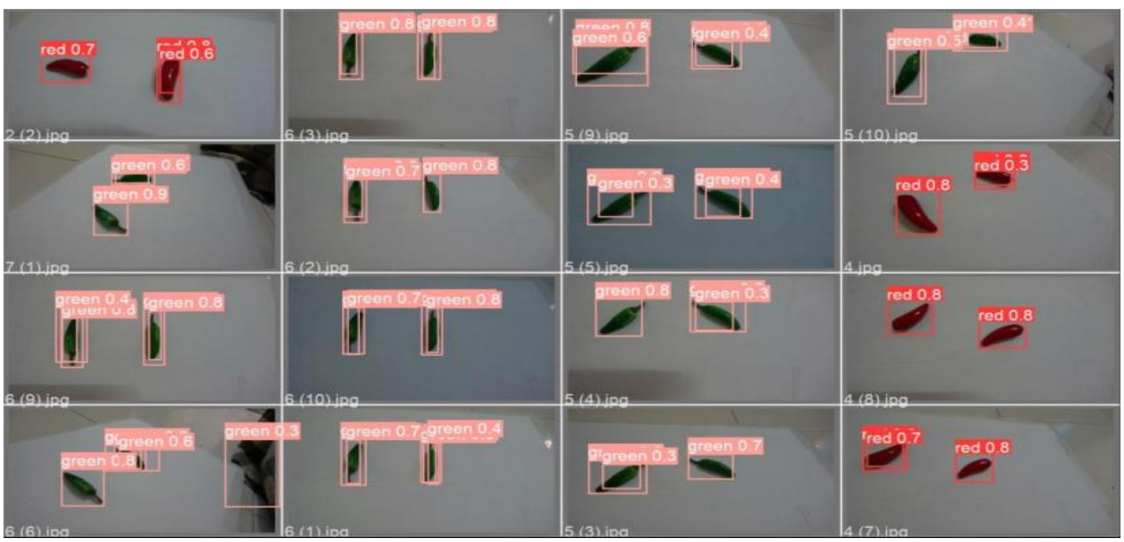

Figure 8 Prediction accuracy for both chili colours (green and red)

We separated the sample for each chili colour in this experiment. For both red and green chilies, two pieces of chilies were placed in the same image. Red chilies appear to be very bright due to their colour representation, whereas green chilies appear to be black in terms of colour. Even though these two chilies can be distinguished, the appearance of the chili may differ due to clutter or lighting. As a result, when the same group of chilies was tested, the accuracy performance was poor. The average precision for red chili is 99 %, while green chili is 83 %. Detail accuracy of detection as tabulated in *Table 1*.

Table 1 Result of single chili colours detection

Class	Precision	Recall	mAP
All	**0.830**	**0.990**	**0.940**
Red	0.927	1.000	0.994
Green	0.733	0.980	0.885

The second part combines for both chili colours into the same images. This experiment is required to analyse and investigate the model's ability to predict the colour of several chili fruits within the same image. The experimental results for both chili colours within the same image are shown in *Figure 9*.

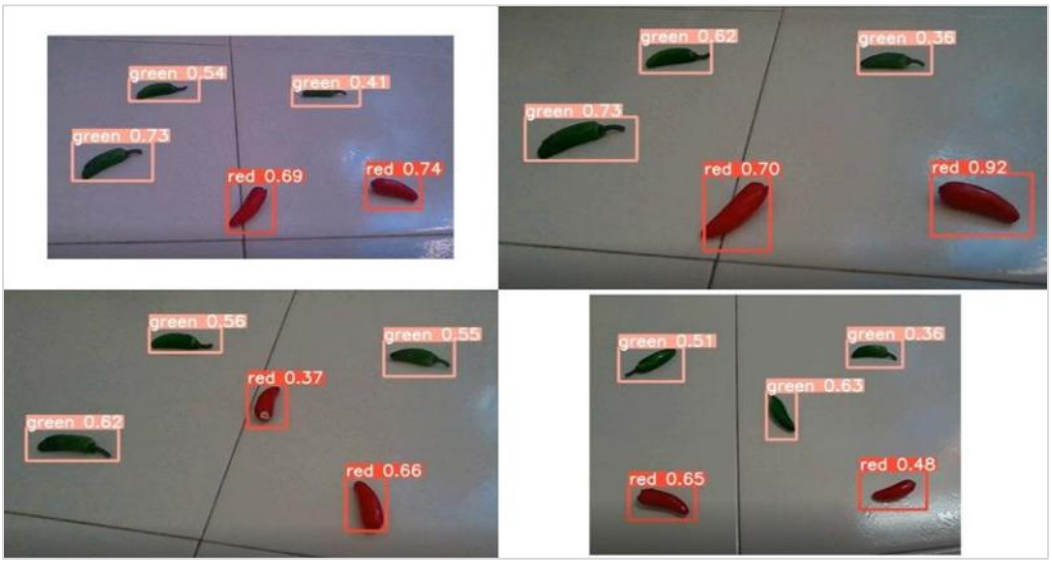

Figure 9 Prediction accuracy for both chili colours (green and red) within the same image

As illustrated in *Figure 8*, the proposed model is still capable of recognising the chilli fruits as an object. However, lighting, light reflection, and other factors such as position influence average prediction accuracy. Some images with very clear presentation achieve high accuracy, while others achieve low accuracy. Because the number of labelling images is insufficient, this issue is also critical for improving accuracy performance. YOLOv5 comes in four different sizes (s, m, l, and xl). The larger the network, intuitively, the more parameters that can be adjusted and the better the performance. To reiterate, having more parameters leads to longer training and inference times. The accuracy performance of multiple chili colours as presented in *Table 2*.

Table 2 Result of multiple chili colours detection

Class	Precision	Recall	mAP
All	**0.810**	**0.936**	**0.340**
Red	0.890	0.908	0.179
Green	0.880	0.965	0.501

The mAP on multiple chilli colour prediction degrades dramatically. The quality of the captured image, which is influenced by factors such as illumination, reflection of lighting, and shading, makes it difficult for the learning model to estimate the chilli colour. Furthermore, uncontrolled lighting darkens the appearance of green chilli.

4.3 Experimental Result of Chili Prediction on Chili Plant

The next part of our experiment will look into the YoLov5 model's ability to recognise and predict chili fruits from the chili plant. Due to limited resources, we used an artificial chili plant in our experiment. Both chili colours (green and red) are used in this section, and the image is taken from a single plant. The labelling images are the same as in the previous

experiment. A variety of plant images are captured from various angles or positions. The parameter used in this experiment is 16 batches with 300 epochs. The training sample for labelling the chili fruits from the plant is shown in *Figure 10.*

As shown in *Figure 10*, the model is capable of recognising and detecting an object (chili fruits) from the plant in both colours. However, due to the high similarity of green chili to leaves, the model is unable to recognise those fruits under certain conditions. The reason for the low number of green chili fruits detected is due to position, image distortion, and the appearance of the leaf. *Figure 11* depicts the prediction accuracy of distinguishing between green and red chili from the plant.

Figure 10 Training sample for labelling both chili colours (green and red) from the plant

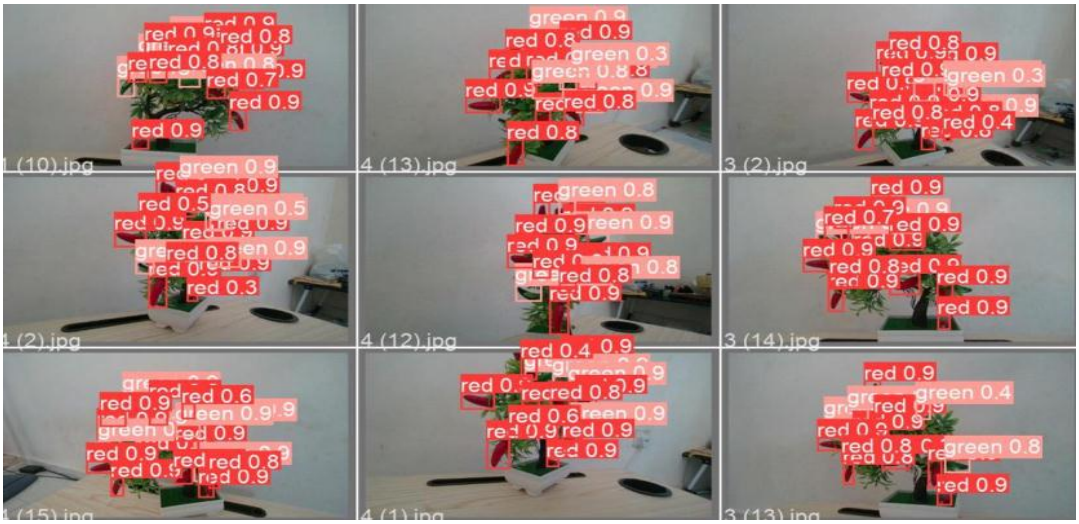

Figure 11 Prediction accuracy for both chili colours (green and red) from the plant

The proposed model clearly recognises and distinguishes between green and red chili fruits, despite having a lower average detection accuracy than the previous experiment. As we can see, the majority of the red chili recorded good accuracy above 80%. Because the appearance of the chili fruits interferes with other objects, some of the chili recorded with low accuracy. Because of interference from plant leaves, the model has difficulty distinguishing between green and red chili with the leaf. Green chili accuracy is slightly lower than the red chili accuracy since red chili is brighter and more

differentiated with plant leaves. *Figure 12* depicts the detection accuracy for both categories of chili fruits for clarity. As we can see from the first experiment, predicting individual chili fruits is far more accurate than predicting fruits from a plant. As shown in *Figure 10*, an average of more than 80% accuracy is obtained for both green and red chili. *Table 3* shows detail performance of detection chili fruits from the chili plant. As shown in *Table 3*, the average performance is slightly lower than that of individual chilli detection in *Table 1*. However, the performance was better than the combination of multiple colours chilli. Because the camera is so close to the plant, it improves the flaws mentioned in the previous section. When the camera is close, the lighting conditions and the effect of reflection improve significantly. Because of the high similarity with its leaves, green chilli scores somewhat lower.

Table 3 Result of chili detection on plant

Class	Precision	Recall	mAP
All	**0.825**	**0.940**	**0.810**
Red	0.841	0.956	0.820
Green	0.800	0.923	0.781

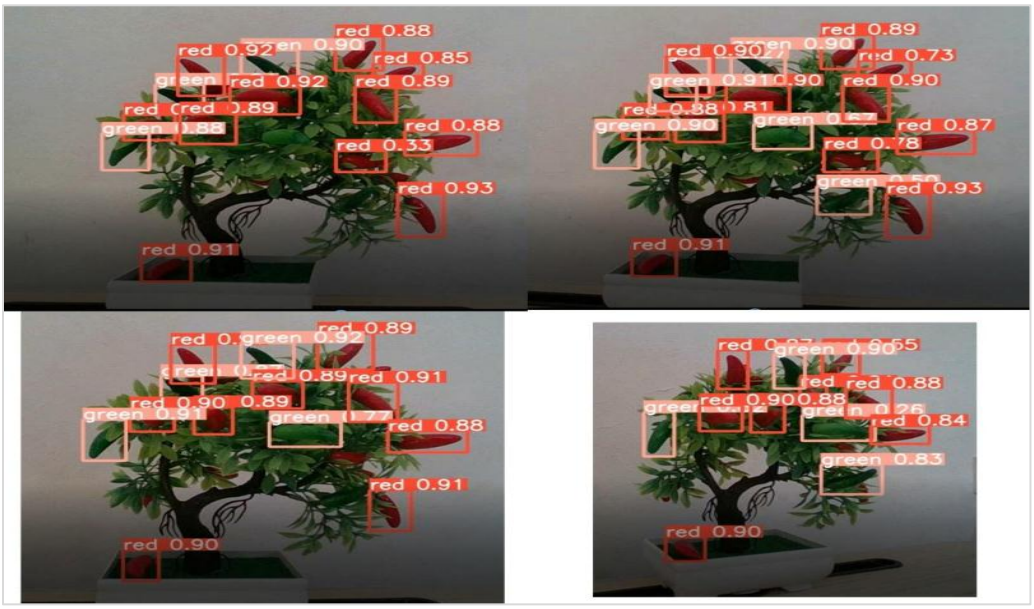

Figure 12 Prediction accuracy for both chili colours (green and red) from the plant

5.Discussion

We are conducting two types of experiments in this work: predicting individual chilli fruits and predicting chilli fruits from the plant. Referring to the results, the accuracy of prediction of only one colour chilli fruits averaged over 90%. When only one colour of chilli fruit is used, the model can predict green or red chilli fruits. Technically, high prediction accuracy is obtained when a number of objects of the same type are involved. In the second experiment, we use an analysis to predict the colour of the chilli fruits from the plant. Due to the difficulty in obtaining a real chilli plant, we decided to conduct our experiment with an artificial plant. The accuracy of prediction is affected by a few conditions. Because the colours of red chilli differ significantly from those of its leaves and green chilli, it can produce more than 80 % mAP on average.

However, when the experiment is expanded to include different types of objects (in our case, red and green chilli), the accuracy obtained is not particularly satisfying. There are several factors that contribute to poor accuracy performance. Variations in illumination, inconsistent lighting, and clutter will reduce detection accuracy. The camera's position is also important in ensuring that the image taken is consistent. The quality of the images has also been influenced by the distance and the position of the chilli fruits. Because we are using artificial chilli fruits made of plastic, the lighting reflection must be considered. As a result, one suggestion to address this concern is to use laboratory-controlled conditions. However, due to the colour similarity with the leaf, green chilli is recorded slightly lower. Furthermore, branch, leaf occlusion, shading, overlapping, and size made detection difficult. Furthermore, the appearance

of the chilli fruits without the intervention of leaf or branch contributes to the low detection accuracy.

A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future work

This work focuses on developing effective methods for distinguishing between the green and red chili from the chili fruits plant. An invention in agricultural industries for undergoing harvesting process may be critical for several reasons, including increasing productivity while lowering labour costs at the same time reducing the incorrectly grading mistakes. On the basis of stereo images, a well-known object detection method such as YOLOv5 is used to localise and predict the chili fruits. As stated earlier, we intend to conduct an experiment to determine the distance between the chilli and the camera. However, this experiment is still ongoing, and we present in our analysis an initial part of our analysis to localise and detect an object (chili fruits). Despite the fact that the green chili in cili padi is small and difficult to distinguish from the plant leaves, our proposed method is capable of producing an outstanding performance in predicting the fruits.

This work is part of our greater goal of developing a semi-autonomous chili fruit picking robot. We intend to project the position of the chilli fruits from the stereo camera based on the depth information provided. Since the depth information was obtained, the maturity of the chilli fruits could also be measured by calculating the image's point cloud to estimate the chilli size. As a result, we intend to test our incoming analysis by measuring the size of the chilli fruits in order to estimate their maturity levels. The size between two points on the stem that connects to the very top of the chilli fruits (calyx) and the rounded tip of the chilli fruits (apex) is calculated using the point cloud from 3D images. As a result, it will assist farmers in automating the sorting and harvesting processes. The system will be integrated with our agricultural robot to speed up the harvesting process. In addition, we are also planning to extend our experiment by comparing with different object detection algorithms with an additional parameter investigation such detection time and model size. Various criteria of images will be taken into consideration such as different front light, back light and occluded scenes.

Acknowledgment

This work funded by Centre for Research and Innovation Management (CRIM), Universiti Teknikal Malaysia Melaka (UTeM) under PJP/2020/FKEKK/PP/S01787 and INDUSTRI (MTUN)/ENDSTRUCT/ 2021/FKEKK/I00057 grant.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

M. N. Shah Zainudin, M. S. S. Shahrul Azlan and L. L. Yin: Responsible for entire study and analysis of experiment, **W. H. Mohd Saad and M. I. Idris:** Responsible in hardware setup and analysis, **Sufri Muhammad:** Responsible as in writing contributors including grammar checking. **M. S. J. A. Razak:** Experiment area/location provider.

References

[1] Gené-mola J, Vilaplana V, Rosell-polo JR, Morros JR, Ruiz-hidalgo J, Gregorio E. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. Computers and Electronics in Agriculture. 2019; 162:689-98.

[2] Wu G, Zhu Q, Huang M, Guo Y, Qin J. Automatic recognition of juicy peaches on trees based on 3D contour features and colour data. Biosystems Engineering. 2019; 188:1-13.

[3] Davies FT. Opportunities for horticulture to feed the world©. In proceedings of the 2014 annual meeting of the international plant propagators society 2014 (pp. 455-8).

[4] Brondino L, Borra D, Giuggioli NR, Massaglia S. Mechanized blueberry harvesting: preliminary results in the Italian context. Agriculture. 2021; 11(12):1-14.

[5] Zainudin MN, Husin N, Saad WH, Radzi SM, Noh ZM, Sulaiman NA, et al. A framework for chili fruits maturity estimation using deep convolutional neural network. Przegląd Elektrotechniczny. 2021; 97(12):77-81.

[6] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In proceedings of the conference on computer vision and pattern recognition 2016 (pp. 779-88). IEEE.

[7] Quinn J, Mceachen J, Fullan M, Gardner M, Drummy M. Dive into deep learning: tools for engagement. Corwin Press; 2019.

[8] Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Computer Science. 2021; 2(6):1-20.

[9] Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, et al. Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image. PloS one. 2021; 16(10):1-15.

[10] Choi RY, Coyner AS, Kalpathy-cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology. 2020; 9(2):1-14.

[11] Diwan T, Anirudh G, Tembhurne JV. Object detection using YOLO: challenges, architectural successors,

datasets and applications. Multimedia Tools and Applications. 2022;1-33.
[12] Zainudin MN, Mohd SM, Ismail MM. Feature extraction on medical image using 2D gabor filter. In applied mechanics and materials 2011 (pp. 2128-32). Trans Tech Publications Ltd.
[13] Sulaiman NA, Abdullah MP, Abdullah H, Zainudin MN, Yusop AM. Fault detection for air conditioning system using machine learning. IAES International Journal of Artificial Intelligence. 2020; 9(1):109-16.
[14] Ahmad HM, Rahimi A. Deep learning methods for object detection in smart manufacturing: a survey. Journal of Manufacturing Systems. 2022; 64:181-96.
[15] Kumar S, Balyan A, Chawla M. Object detection and recognition in images. International Journal of Engineering Development and Research. 2017; 5(4):1029-34.
[16] Buhmann JM, Malik J, Perona P. Image recognition: visual grouping, recognition, and learning. Proceedings of the National Academy of Sciences. 1999; 96(25):14203-4.
[17] Murphy K, Torralba A, Eaton D, Freeman W. Object detection and localization using local and global features. In toward category-level object recognition 2006 (pp. 382-400). Springer, Berlin, Heidelberg.
[18] www.intel.com/design/literature.htm . Accessed 20 June 2022.
[19] Kadambi A, Bhandari A, Raskar R. 3D depth cameras in vision: benefits and limitations of the hardware. In computer vision and machine learning with RGB-D sensors 2014 (pp. 3-26). Springer, Cham.
[20] Jeon HG, Lee JY, Im S, Ha H, Kweon IS. Stereo matching with color and monochrome cameras in low-light conditions. In proceedings of conference on computer vision and pattern recognition 2016 (pp. 4086-94). IEEE
[21] Gongal A, Karkee M, Amatya S. Apple fruit size estimation using a 3D machine vision system. Information Processing in Agriculture. 2018; 5(4):498-503.
[22] Tian Y, Yang G, Wang Z, Wang H, Li E, Liang Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computers and Electronics in Agriculture. 2019; 157:417-26.
[23] Tian Y, Yang G, Wang Z, Li E, Liang Z. Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. Journal of Sensors. 2019; 2019:1-14.
[24] Liu J, Wang X. Tomato diseases and pests detection based on improved YOLO V3 convolutional neural network. Frontiers in Plant Science. 2020; 11:1-12.
[25] Kuznetsova A, Maleva T, Soloviev V. Using YOLOv3 algorithm with pre-and post-processing for apple detection in fruit-harvesting robot. Agronomy. 2020; 10(7):1-19.
[26] Lawal MO. Tomato detection based on modified YOLOv3 framework. Scientific Reports. 2021; 11(1):1-11.
[27] Fu L, Feng Y, Wu J, Liu Z, Gao F, Majeed Y, et al. Fast and accurate detection of kiwifruit in orchard using improved YOLOv3-tiny model. Precision Agriculture. 2021; 22(3):754-76.
[28] Yao J, Qi J, Zhang J, Shao H, Yang J, Li X. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics. 2021; 10(14):1-13.
[29] Kuznetsova A, Maleva T, Soloviev V. Detecting apples in orchards using YOLOv3 and YOLOv5 in general and close-up images. In international symposium on neural networks 2020 (pp. 233-43). Springer, Cham.
[30] Manan AA, Razman MA, Khairuddin IM, Shapiee MN. Chili plant classification using transfer learning models through object detection. Mekatronika. 2020; 2(2):23-7.
[31] Hespeler SC, Nemati H, Dehghan-niri E. Non-destructive thermal imaging for object detection via advanced deep learning for robotic inspection and harvesting of chili peppers. Artificial Intelligence in Agriculture. 2021; 5:102-17.
[32] Sihombing YF, Septiarini A, Kridalaksana AH, Puspitasari N. Chili classification using shape and color features based on image processing. Scientific Journal of Informatics. 2022; 9(1):42-50.
[33] Cruz-domínguez O, Carrera-escobedo JL, Guzmán-valdivia CH, Ortiz-rivera A, García-ruiz M, Durán-muñoz HA, et al. A novel method for dried chili pepper classification using artificial intelligence. Journal of Agriculture and Food Research. 2021; 3:1-7.
[34] Sembiring A, Basuki RS, Rosliani R, Rahayu ST. Farmers' challenges on chili farming in the acid dry land: a case study from pasir madang-bogor regency, Indonesia. In E3S web of conferences 2021 (pp. 1-7). EDP Sciences.
[35] Xu R, Lin H, Lu K, Cao L, Liu Y. A forest fire detection system based on ensemble learning. Forests. 2021; 12(2):1-17.
[36] Ferguson M, Ak R, Lee YT, Law KH. Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning. ASTM International. 2018; 2(1):1-28.
[37] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 7263-71). IEEE.
[38] Gupta S, Devi DT. YOLOv2 based real time object detection. International Journal of Computer Science Trends and Technology. 2020; 8(3):26-30.
[39] Redmon J, Farhadi A. YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767. 2018:1-6.
[40] Hsieh KW, Huang BY, Hsiao KZ, Tuan YH, Shih FP, Hsieh LC, et al. Fruit maturity and location identification of beef tomato using R-CNN and binocular imaging technology. Journal of Food Measurement and Characterization. 2021; 15(6):5170-80.

[41] Sari AC, Setiawan H, Adiputra TW, Widyananda J. Fruit classification quality using convolutional neural network and augmented reality. Journal of Theoretical and Applied Information Technology. 2021; 99(22):5300-11.
[42] Bochkovskiy A, Wang CY, Liao HY. Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. 2020; 4(2004.10934):1-17.
[43] Yu J, Zhang W. Face mask wearing detection algorithm based on improved YOLO-v4. Sensors. 2021; 21(9):1-21.
[44] Liao Z, Tian M. A bird species detection method based on YOLO-v5. In international conference on neural networks, information and communication engineering 2021 (pp. 65-75). SPIE.
[45] https://github.com/ultralytics/yolov5. Accessed 24 April 2022.
[46] Song Q, Li S, Bai Q, Yang J, Zhang X, Li Z, et al. Object detection method for grasping robot based on improved YOLOv5. Micromachines. 2021; 12(11):1-18.

M. N. Shah Zainudin is a senior lecturer at Universiti Teknikal Malaysia Melaka. He received his bachelor's degree, master's degree in Computer Science from Universiti Teknologi Malaysia and doctorate in Intelligent Computing from Universiti Putra Malaysia. His current research interest includes Artificial Intelligence, Data Mining, Pattern Recognition and Machine Learning.
Email: noorazlan@utem.edu.my

M. S. S. Shahrul Azlan is an undergraduate student from Universiti Teknikal Malaysia Melaka. He received his bachelor's degree in Electronic Engineering from Universiti Teknikal Malaysia Melaka.
Email: B021820061@student.utem.edu.my

L. L. Yin is an post-graduate student from Universiti Teknikal Malaysia Melaka. Her received his bachelor's degree in Electronic Engineering technology and master's degree in Electronic Engineering from Universiti Teknikal Malaysia Melaka.
Email: M022010040@student.utem.edu.my

W. H. Mohd Saad is an associate professor at Universiti Teknikal Malaysia Melaka. He received his bachelor's degree in Electrical and Electronic Engineering and doctorate (PhD) in Multimedia System Engineering from Universiti Putra Malaysia. His current research interest includes Medical Signal and Image Processing, Embedded Artificial Intelligence, Deep Learning in Computer Vision.
Email: wira_yugi@utem.edu.my

M. I. Idris is a senior lecturer at Universiti Teknikal Malaysia Melaka. He received her bachelor's degree in electronic system engineering from Hiroshima University, Japan, master's degree in microelectronics from Universiti Kebangsaan Malaysia and doctorate (PhD) in Semiconductor Devices from Newcastle University, UK. His current research interest includes Fabrication of Semiconductor Devises (solar cell, transistor, MOS capacitor, diode) and IC Design.
Email: idzdihar@utem.edu.my

Sufri Muhammad is a senior lecturer at Universiti Putra Malaysia. He received her bachelor's degree, master's degree and doctorate (PhD) in Computer Science from Universiti Putra Malaysia. His current research interest includes Service-Oriented Architecture, Service Engineering, Semantic-Based Approach and Context-Aware Mobile Cloud Learning.
Email: sufry@upm.edu.my

M. S. J. A. Razak received his bachelor's degree and master's degree in Biology from Universiti T eknologi Mara. His current research interest includes Molecular and Biology Genetics. Currently he is managing and practicing fertigation chili farms for marketing and research purposes.
Email: solokfertigasi@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	AI	Artificial Intelligence
2	ANN	Artificial Neural Network
3	AP	Terms of Average Precision
4	CCD	Charge-coupled Device
5	CNN	Convolutional Neural Network
6	COCO	Common Objects in Context
7	CycleGAN	Cycle-Consistent Adversarial Network
8	Faster R-CNN	Faster Region Based-Convolutional Neural Network
9	FPP	Feature Pyramid Pooling
10	FPS	Frame Rate Per Second
11	IoU	Intersection Over Union
12	SDK	Software Development Kit
13	SPP	Spatial Pyramid Pooling
14	TOF	Time of Flight
15	YOLO	You Look Only Once