Yolo + Positioning

Chi Yen Lee
r06631026@ntu.edu.tw
National Taiwan University
Taipei, Taiwan (R.O.C.)

Chun Yen Tai
r07631030@ntu.edu.tw
National Taiwan University
Taipei, Taiwan (R.O.C.)

Figure 1: The diagrammatic sketch of research purpose.

ABSTRACT

In order to improve the positioning of the automated vehicle in the greenhouse, we use a monocular camera to capture the image containing the vehicle, and Yolo v3 to find the position of the vehicle in the image.Then use the bird-eyes transformation and the full connection layer training model convert the position of the vehicle in the image to the world coordinates,the position of the vehicle in the greenhouse.The results confirmed that the average recognition accuracy of the vehicle for Yolo v3 can reach 91.66%, and the positioning accuracy can be controlled within 5 cm to 20 cm. This system uses a low-cost monocular camera combined with deep learning to improve the accuracy of indoor positioning.

KEYWORDS

Indoor positioning, neural networks

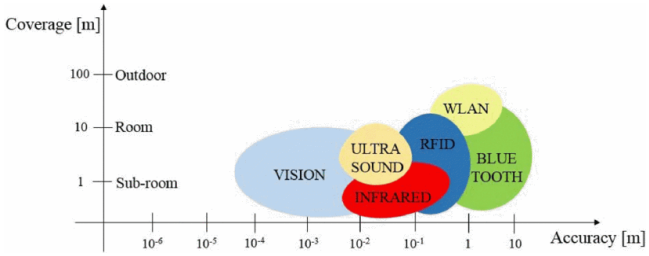

Figure 2: Graphical overview of the technologies enabling the indoor localization in dependence of accuracy and coverage.[2]

1 INTRODUCTION

Taiwan is located in the subtropical zone. Because of the environment which is warm and moist, it is suitable for crop growth in this country. Nevertheless, it is also susceptible to pests, diseases and natural disasters. Those make the agricultural products in reduced circumstances and then the cost would increase. Furthermore, the production of crops has seasonal problems, so that we can only rely on import cargo in in certain seasons.

Therefore, Council of Agriculture planned to promote the five-year plan for facility-based agriculture from 2017.The plan assist farmers in the construction of strong-type greenhouse. It is scheduled to increase the facility area by 2,000 hectares.Council of Agriculture also promoted the âĂIJSmart Agriculture 4.0âĂİ from 2017, and raised the need for automation. Hence, it is important to build the system of automated vehicles in the greenhouse. The system depends on the positioning system, obstacle avoidance function and path planning. We can find the indoor positioning is the most important part in this system.

A number of technologies have been proposed for indoor positioning systems which include Bluetooth, laser, RFID, Wi-Fi, ZigBee and UWB. For this case, according to the Figure 2, only Bluetooth and lasers technologies meet the demand.

Therefore, this paper intends to study a visual positioning system and then add it to the original laser positioning. The visual

positioning system consists of two parts, one is object tracking, and the other is to convert the results obtained in the first part to the world coordinates.

2 APPROACH

2.1 System introduction

This paper proposes a visual positioning system model, which consists of two parts, one is object tracking, and another is to convert the results obtained in the first part to the world coordinates. In the first part, we used the Yolo v3 method. In the second part, we compared two methods, one is the bird-eyes view transformation, and another is the deep learning model of fully connected layer. The method of bird-eyes view transformation is relatively easy and fast. We only need to know the transformation matrix through the correction plate. The deep learning method need to establish the training data set by recording the actual position information. Therefore, it is complicated in setting.

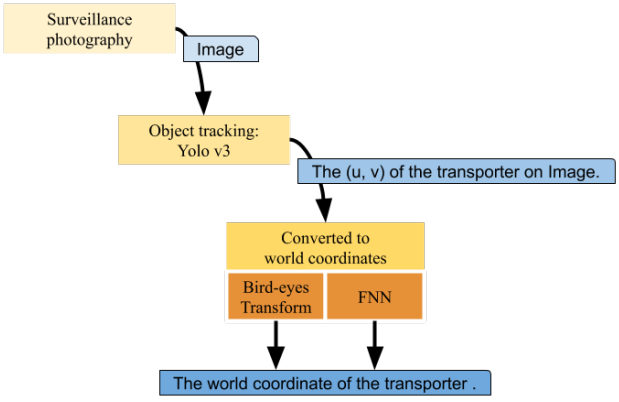

Figure 3: The structure of the visual positioning system.

2.2 Setting

We chose four different locations reported in Figure 4 to set our camera. The reason is that we need to test whether the same Yolo v3 model is suitable for different shooting angles, when we only use the data set of the upper left surveillance camera to train the Yolo v3 model.

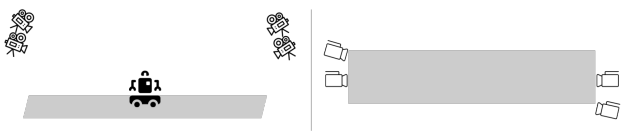

Figure 4: The modes of camera setting.

2.3 Object tracking

We use Yolo v3 to capture the position of the automated vehicle to improve the positioning accuracy of the automated vehicle. Yolo v3 can directly predicts the object position and the class of the object according to the input test image.The model structure is as shown in the Figure 5, and its structure consists of three parts: darknet53, Feature Pyramid Networks and output layer[1]. Darknet53 uses the convolutional layer to extract the features of the image and adds a residual block to prevent the gradient diverge.Feature Pyramid Networks concat better features for location capture and better features for identifying class of objects for better results.The output layer converts features into specific dimension data for prediction.

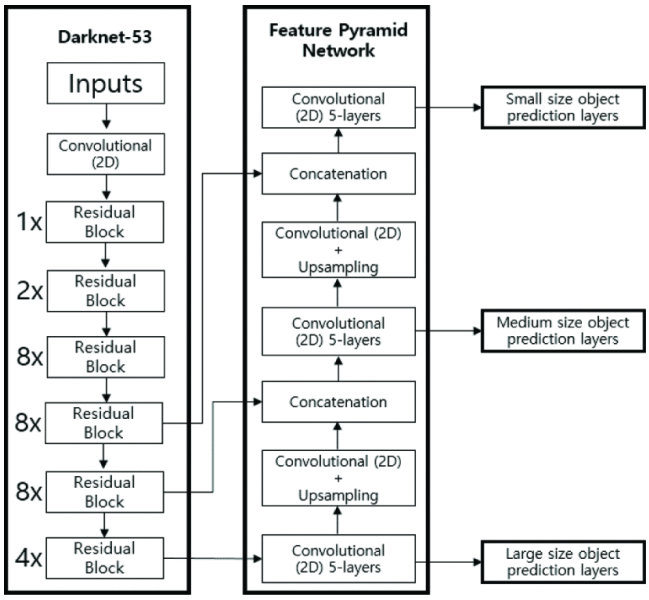

Figure 5: The structure of darknet53.[1]

2.4 Transformation

In the transformation of the world coordinates, we apply two different methods, one is the bird-eyes view transformation, and another is the deep learning method about fully connected layer.

2.4.1 Bird-eyes view transformation. The following formula is a transformation formula for converting image coordinates to world coordinates.

$$z\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} [R|t] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

We can apply it to convert the results obtained by object tracking to the world coordinates. For implement, we only need to use the calibration plate to perform the calculation of the transformation matrix. The transformation matrix for each shooting angle is shown in the Figure 6.

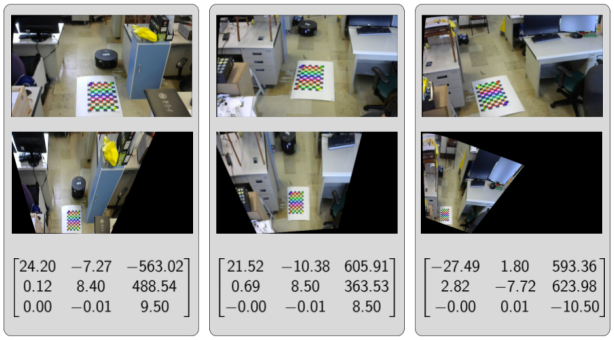

Figure 6: The transformation matrix for each shooting angle.

2.4.2 Fully connected layer. The result obtained by the Yolo v3 object tracking model is the input data of this model and output is the world coordinates of the target object, as shown in the Figure 7. On the preparation of the dataset, we set up a camera on the ceiling and then use the hough circles detection method to seek the real position of the object. We tested different layers and batch size. The results are shown in the Figure 8. It can be found that the effect when the number of layer is equal to 3 and the batch size is equal to 128 is the best performance. Therefore, we apply a 3 fully connected layer for this transformation.

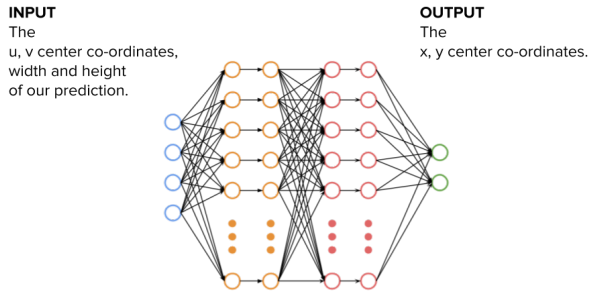

Figure 7: The structure of FCN.

3 EXPERIMENTS

3.1 Object tracking

We used the camera in the upper left position to train the Yolo v3 model. The results are shown in the table. In addition, we apply the same parameters to other positions. The results are shown in the table.

3.2 Transformation: Bird-eyes view transformation

The results of the bird-eyes view transformation are shown in the Figure 9. The average error on the x-axis is equal to 32.01, about 0.06402 meters. The average error on the y-axis is equal to 56.189, about 0.112378 meters. The average error of distance is equal to 0.13 meters.

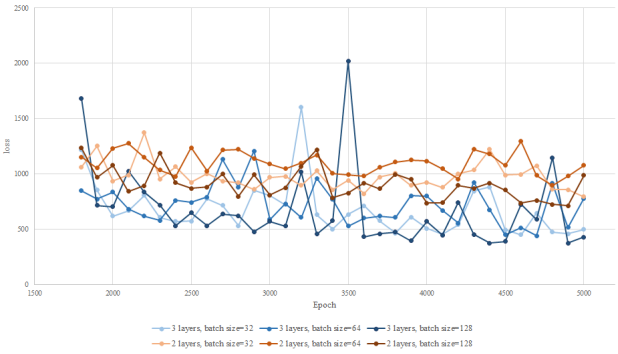

Figure 8: Loss(MSE) value of testing data.

Table 1: The result of object tracking by Yolo v3 model at the position where the training data set is included.

Position 0 Predict\Real	Object exists	Object doesn't exit
Object exists	2250(69.78%)	5(0.16%)
Object doesn't exit	107(3.32%)	862(26.74%)

Table 2: The result of object tracking by Yolo v3 model at 3 positions where the training data set is not included.

Position 1 Predict\Real	Object exists	Object doesn't exit
Object exists	2369(77.24%)	2(0.07%)
Object doesn't exit	317(10.34%)	379(12.35%)
Position 2 Predict\Real	Object exists	Object doesn't exit
Object exists	3663(79.15%)	0(0.00%)
Object doesn't exit	762(16.46%)	203(4.39%)
Position 3 Predict\Real	Object exists	Object doesn't exit
Object exists	2817(70.45%)	0(0.00%)
Object doesn't exit	60(1.50%)	1123(28.08%)

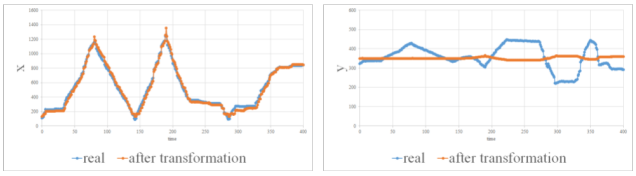

Figure 9: The result of bird-eyes view transformation.

Table 3: Comparison of the bird-eyes view transformation method and the fully connected layer transformation method

	Bird-eyes view transform	Fully connected layer
error of x(m)	0.06404	0.0326
error of y(m)	0.112378	0.0114
error of distance(m)	0.13	0.0345
Setting difficulty	Easier.	Harder.

3.3 Transformation: FCN

In the fully connected layer transformation method, the results are as shown in the Figure 10, the average error on the x-axis is equal to 16.30, about 0.0326 meters; the average error on the y-axis is 5.70, about 0.0114 meters; the average error of distance is 0.0345 meters.

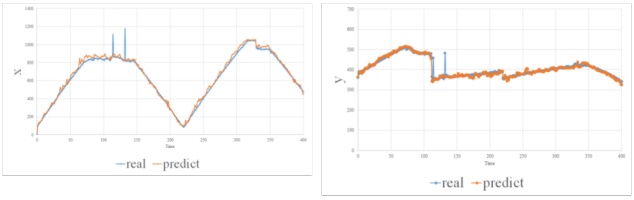

Figure 10: The result of FCN.

4 CONCLUSION

In this case, we apply the visual positioning method, which include the Yolo v3 model and the bird's-eye view transformation or fully connected layer transformation method. Positioning accuracy is about 5 cm to 20 cm. Table 1 is a comparison of the bird-eyes view transformation method and the fully connected layer transformation method. The accuracy of the fully connected layer transformation method is superior to the bird-eyes view transformation method. However, the bird-eyes view transformation method only needs to use the correction board to obtain the transformation matrix when the monitor camera node is set up. However, the fully connected layer transformation method needs to collect the actual positioning information to establish the training data set. In terms of complexity, the bird-eyes view transformation method is superior.

5 WORK DISTRIBUTION

- C. Y. Lee(50%):
  - Paper work.
  - Bird-eyes view transformation.
  - Predict the result within fully connected layer.
- C. Y. Tai(50%):
  - Paper work.
  - Build and Train fully connected layer model.

REFERENCES

[1] Kwang-Ju Kim, Pyong-Kun Kim, Yun-Su Chung, and Doo-Hyun Choi. 2018. Performance Enhancement of YOLOv3 by Adding Prediction Layers with Spatial Pyramid Pooling for Vehicle Detection. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[2] Luca Mainetti, Luigi Patrono, and Ilaria Sergi. 2014. A survey on indoor positioning systems. In *2014 22nd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 111–120.