應用卷積神經網路與雙目視覺於番茄機器人採摘之研究

黃柏喻

國立中興大學 生物產業機電學系

摘　要

本研究將機器學習方式導入至牛番茄採收，以期能達到省時、省工的目的並以更精確的方式減少人為上的誤差。將針對溫室內番茄進行成熟度與位置的辨識，此研究包含兩個部分：影像擷取與物體偵測模型之建立、三維位置計算。第一部分將於溫室內進行不同時間取像並使用物體偵測進行成熟果實的辨識；第二部分將雙目視覺計算辨識到的成熟果實相對位置，並對其結果進行校正，最後計算手臂移動所需的六軸資訊。本研究在偵測成熟果實的準確率與召回率均高於 95%，立體空間位置誤差平均為 0.5 cm，且果實實際大小與邊界框大小的 R^2 高於 0.9，該採摘系統所進行的採摘一回流程僅需耗時 25 秒。

關鍵詞：牛番茄、物體偵測、深度學習、雙目視覺、機械手臂、採收。

緒　論

設施蔬果栽培特質為投入的資本較高，如設施需具備光源、溫溼度等環境變數的調整，確保栽培蔬果的品質，因此在設施中栽培之蔬果售價、品質均會高於一般露天栽培者，使設施栽培者收益會有較高的收益。目前農村勞力缺乏與老化問題十分普遍，設施蔬菜栽培也不例外。不過大型固定設施因現成結構之存在，所以可以配合許多機械化裝置，以減少重複性操作或較粗重勞力工作，將可節省勞力，提高工作舒適躓。

過往已有許多水果採收機器人的研究，使用 3-D 攝影機進行取像並進行採收[1]。在影像擷取上有許多設備能進行不同方面的資料取得，如彩色攝影機、光譜攝影機、熱影像攝影機等[2]。本研究將使用雙目視覺攝影機，拍攝的影像為 RGB 的彩色影像，並以計算視差取得 RGB-D 的彩色與深度影像。

對於水果採收判斷成熟果實的影像前處理有使用 HSI[3,4]或色彩平衡[5]等方式，將果實從背景中分離，使用邊緣演算法進行果實外觀的詮釋[6]。得到的詮釋資料再使用支持向量機(support vector machine, SVM)或人工神經網路(Artificial Neural Network, ANN)進行非線性運算[2,7,8]。近年來新的深度學習演算法在影像辨識上有很好的成功率，因此有相當多的人用來辨識水果的定位與成熟度[9,10,11,12]。

材料與方法

一、Depth Camera

為達成水果採收的目的，機械手臂在移動上不僅只有平面上的移動，而是一個三維的立體空間，平面上的位置能由一般的 RGB 攝影機計算出來，但攝影機距離目標的深度使用單一個攝影機不易計算出來，若單一攝影機搭配滑軌進行線性移動，則要考慮馬達與滑軌設備，在校準上會有誤差。

本研究選擇使用 ZED mini 雙目視覺攝影機，此攝影機透過 USB 3.0 Type-C 介面與電腦連接，拍攝影像的解析度為 1280 x 720，三維深度感測距離為 0.15 m-12 m。

在深度計算部分，本研究應用 Mask R-CNN[13]推論的邊界框使用 ZED 官方的函式庫進行三維位置的計算，如圖 1 所示，其深度攝影機之三維空間資訊，X 方向為左鏡頭中心向右鏡頭的方向；Y 方向為垂直於左鏡頭之 X 方向向下延伸；Z 方向則是垂直於左鏡頭 XY 平面向前延伸。

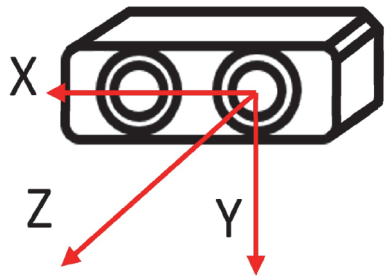

圖 1、VGG Image Annotator 操作介面

二、系統架構

本研究設計一番茄採收設備裝設於一行走載具上，包含深度攝影機、秤重設備、中控主機、機械手臂、採收爪具等。本研究將影像擷取、物件偵測、雙眼立體視覺等技術結合，擷取一張良好的影像後使用物件偵測模型分析，將處理完後之資訊計算其三維立體空間之座標，將其轉換至手臂坐標系，以完成番茄採收之目的。

三、訓練資料集建立

在影像中，以人工完成將要偵測的物件進行樣本標記的工作，再將影像本身與標記完成的資料輸入神經網路進行訓練，利用 Mask R-CNN 將判斷的結果與實際的結果進行計算誤差，根據其誤差值反過來修正演算法內部的參數。然而機器學習受到輸入資料的影響很大，未被訓練過的影像很容易出現誤偵測或沒找到目標的情況，於是為了評估一模型的訓練成效，在收集到的資料上會分為訓練集(training dataset)與驗證資料集(validation dataset)。

本研究將由深度攝影機取得的 255 張影像進行訓練與驗證資料的分類，將 205 張影像歸類為訓練集資料，50 張影像歸類為驗證集資料，彼此不相重複，且為了增加資料量，在輸入神經網路之前會進行左右翻轉的動作，能使數據量增為 2 倍。

四、樣本標記

本研究使用 VGG Image Annotator(VIA)[14]進行成熟樣本的標記。

VIA 為輕量型卻功能豐富的人工標記程式，在網頁瀏覽器上便能輕易進行標記的工作，且能輸出不同的資料格式提供使用者使用。

VIA 的輸出標記資料格式選用 json(JavaScript Object Notation)，標記完成的資料包括影像檔案名稱、影像大小、標記框形式集其資訊、標記類別等資料。本研究將經 VIA 標記完成的資料再轉換成 Microsoft common objects in context(MS COCO)資料格式輸入神經網路。

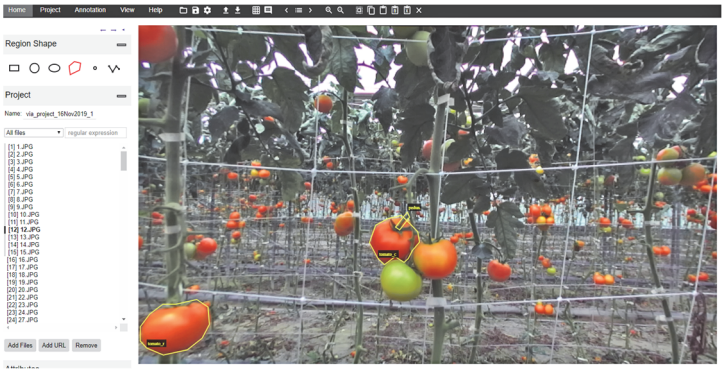

圖 2、VGG Image Annotator 操作介面

考慮到實際番茄生長狀況不一，在其生長過程中受到其他果實的推擠、網子的遮擋、枝幹的遮擋等情況，未必是每一顆成熟的番茄皆能進行採摘，故在此本研究將目標分為三類：成熟且無遮擋番茄(Full)、成熟但被遮擋的番茄(Covered)、番茄梗(Peduncle)。

如圖 2 所示，中間的成熟番茄被下方未成熟的綠色番茄遮擋，便對分類為遮擋的番茄，並對其梗進行標記；左下即是成熟且無遮擋番茄，沒有對其標記梗的原因為，當人類都不能確認其目標確實位置時便不對其進行標記。

五、Mask R-CNN 參數設定

本研究模型使用 ResNet-101 在 COCO 訓練過的預訓練模型(pre-trained model)進行遷移式學習(transfer learning)，COCO 資料集有 80 個分類與數萬張的影像，預訓練模型即是使用此數據集先行進行訓練，因 COCO 的數據集數量龐大且種類多元，故預訓練模型已經有各種類別的資訊，且抓取特徵的能力強大；未來在不同研究領域下，只要套用此模型，再加上自

已準備要應用的場合進行最後的訓練，將模型轉化成所需的專用模型即可，如此便能節省在開發不同領域時，訓練完整模型的時間。

本研究將訓練分為三個階段，如表 3-5 所示。第一階段使用 0.001 學習率，僅訓練 RPN、分類與 mask 層；第二階段使用 0.001 學習率，訓練 ResNet 第 4 部分之後；第三階段則是以 0.0001 學習率訓練全部的網路。

訓練過程將會記錄訓練資料集與驗資料集的整體誤差、RPN 誤差、mask 層誤差等。提供最後模型選擇的指標之一。

表 1、訓練階段資料表

Stage	Epoch	LR	Layers
1	1-40	0.001	heads
2	40-120	0.001	ResNet Stage 4 and up
3	120-160	0.0001	all

六、機械手臂定位控制

本研究使用手臂為協作型機械手臂 UR-5，其特點為有六軸轉動機構，分別為 x、y、z、Rx、Ry、Rz，XYZ 由深度攝影機提供，剩下三個資訊 Rx、Ry、Rz 為旋轉向量，利用角度的變化再經由羅德里格旋轉公式去計算實際的旋轉向量，得出下列公式：

$$x = \frac{r_{32} - r_{23}}{2\sin\theta} \times \theta \quad (1)$$

$$y = \frac{r_{13} - r_{31}}{2\sin\theta} \times \theta \quad (2)$$

$$z = \frac{r_{21} - r_{12}}{2\sin\theta} \times \theta \quad (3)$$

根據上述計算結果表示，只要能擁有角度資訊便能使用羅德里格旋轉公式將旋轉矩陣轉換至旋轉向量供機械手臂使用；因此本研究將影像分為四個區塊，共九個頂點，以手動方式移動機械手臂末端至前述頂點，記錄其腕關節 3 所架設之二指夾爪正面面對果實的三個腕關節分別的旋轉角度。紀錄完的頂點使用雙線性插值取得目標點的旋轉角度，最後使用轉換式計算機械手臂腕關節所需的旋轉向量之值。

結果與討論

一、Object Detection

本研究使用 Mask RCNN，backbone 選用 Resnet101 的 COCO pre-trained model，learning rate 設定 0.001，使用隨機梯度坡降法（SGD），輸入圖片大小為 1024 x 704，訓練 160 epochs。

總影像數量為 255 張，隨機挑選 205 張作為訓練樣本，剩餘 50 張作為驗證樣本，使用 VIA 進行多邊形框選標註，在全部照片中共有 1538 顆有標註之番茄。

為了評估模型的成效與閾值的選擇，使用下列公式進行計算：

$$\text{Precision} = TP/(TP + FP) \tag{4}$$

$$\text{Recall} = TP/(TP + FN) \tag{5}$$

$$\text{IoU} = Intersection/Union \tag{6}$$

TP 為正確判斷的數量，FP 為錯誤判斷的數量，FN 為沒判斷到的數量。

Intersection over Union(IoU)為辨識的邊界框與實際框的重疊比率。

使用 Mask R-CNN 訓練的目標物分為：完整的番茄、被遮擋的番茄、梗。如圖 3 所示，驗正樣本在 IoU@90 情況下，即 IoU 大於 0.9 判斷為正確，小於 0.9 則為錯誤判斷，mAP 為 1.0，在 IoU@95 下才有失敗，此時訓練結果 Precision 為 98.6%，Recall 為 98.6%，圖 4 則為 IoU@95 時，三個分類的 AP。

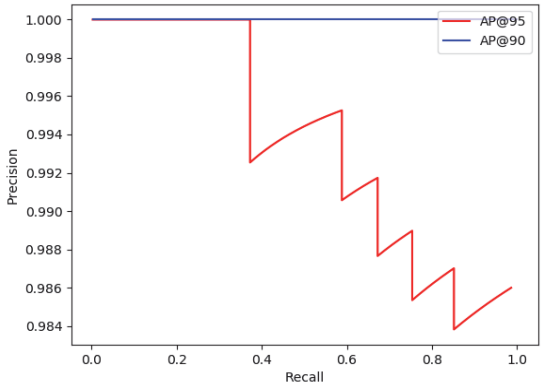

圖 3、Precision-recall curves of two IoUs.

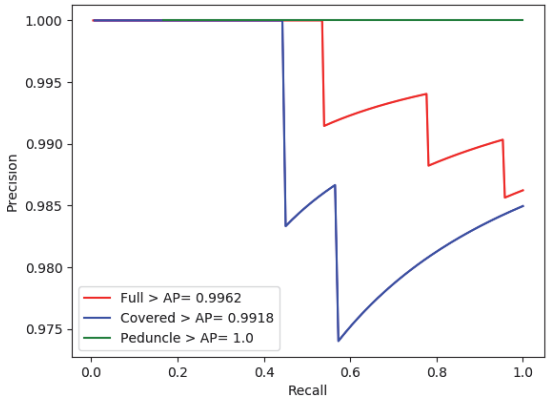

圖 4、Precision-recall curves of three classes.

二、深度攝影機位置計算

本研究使用 ZED mini 雙目視覺深度攝影機作為研究材料，其深度可計算範圍為 0.15-12 m 間，在本研究的場域攝影機裝設於自走載具上，攝影機至番茄果實距離約為 0.4-0.7 m 間，因此對深度攝影機計算之三維空間資訊與實際位置進行比對與計算誤差。如表 2 所示，最大誤差為 1.85 cm，平均誤差約為 0.53 cm；XY 方向的最大誤差為 1.14 cm，平均為 0.48 cm，深度 Z 方面最大誤差為 1.86 cm，平均誤差為 0.67 cm。

表 2、深度攝影機三維空間計算與實際位置比較

(cm)	Prediction			Measurement		
	X	Y	Z	X	Y	Z
1	-1.52	-13.14	48.08	-2.00	-12.00	48.00
2	-5.17	-11.46	56.07	-5.50	-11.00	56.00
3	8.46	-7.93	49.80	8.00	-8.00	50.50
4	14.64	-10.57	64.35	14.00	-10.50	62.50

三、機械手臂定位控制結果

本研究為取得旋轉向量之值，使用角度轉換成弧度，建構一旋轉矩陣並以此矩陣轉換至旋轉向量；因此需要量測其 2 x 2 網格的頂點角度值，在記錄數值時發現在 50 cm、55 cm、60 cm 間，機械手臂腕部 3 個關節的旋轉角度在相同頂點下變化性不大，因此本實驗使用 55 cm 的深度紀錄頂點的數值，如表 3 所示。

表 3、腕關節於頂點旋轉角度表

頂點位置		1			2			3		
腕關節		1	2	3	1	2	3	1	2	3
頂點位置	1	89.31	43.12	86.10	95.37	47.19	100.84	90.04	44.08	90.06
	2	89.09	43.32	94.08	92.70	45.41	91.94	93.32	43.26	97.04
	3	92.11	44.20	93.93	91.37	46.25	94.87	88.65	48.95	93.26

根據上表即可以雙線性插值取得影像中任意點的機械手臂腕關節需轉動的角度值，最後計算出轉換成旋轉像量的數值後，便能以此 6 個值傳送至機械手臂並驅動機械手臂至該定點，根據上表即可以雙線性插值取得影像中任意點的機械手臂腕關節需轉動的角度值。

結　論

本研究整合物件偵測網路與深度攝影機的運用，建立一成熟番茄果實位置定位與預估其大小。物件偵測網路使用具有高精確度的 Mask R-CNN 搭配 ResNet 網路進行成熟果實的檢測，其像素級別的檢測能很好的對被遮擋的物體進行預測，對於農業上來說有很好的應用空間。

本研究對該番茄採摘系統進行採摘試驗，經試驗結果顯示在距離為 50 與 60 公分時，在可採取的上方範圍會有採取位置偏移的情況，但仍可進行採摘的動作。本研究從影像擷取、影像偵測、果實採摘至秤重的一回流程總耗時為 25 秒。使用的機械手臂使用的是商用手臂，以實際應用來看，該機器手臂 UR-5 的價格過於高昂，建議未來改採用自製的機械手臂來替代。

誌　謝

本研究由科技部計畫編號 MOST 109-2321-B-002-052 經費支持。

參考文獻

1. Quan, Q., Lanlan, T., Xiaojun, Q., Kai, J., & Qingchun, F. (2017, April). Selecting candidate regions of clustered tomato fruits under complex greenhouse scenes using RGB-D data. In 2017 3rd International Conference on Control, Automation and Robotics (ICCAR) (pp. 389-393). IEEE.
2. Gongal, A., Amatya, S., Karkee, M., Zhang, Q., & Lewis, K. (2015). Sensors and systems for fruit detection and localization: A review. Computers and Electronics in Agriculture, 116, 8-19.
3. Hayashi, S., Shigematsu, K., Yamamoto, S., Kobayashi, K., Kohno, Y., Kamata, J., & Kurita, M. (2010). Evaluation of a strawberry-harvesting robot in a field test. Biosystems engineering, 105(2), 160-171.
4. Ji, W., Zhao, D., Cheng, F., Xu, B., Zhang, Y., & Wang, J. (2012). Automatic recognition vision system guided for apple harvesting robot. Computers & Electrical Engineering, 38(5), 1186-1195.
5. Nguyen, T. T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J. G., & Saeys, W. (2016). Detection of red and bicoloured apples on tree with an RGB-D camera. Biosystems Engineering, 146, 33-44.
6. Bulanon, D. M., Kataoka, T., Ota, Y., & Hiroma, T. (2002). AE—automation and emerging technologies: a segmentation algorithm for the automatic recognition of Fuji apples at harvest. Biosystems Engineering, 83(4), 405-412.
7. Xiong, J., R. Lin, Z. Liu, Z. He, L. Tang, Z. Yang and X. Zou. 2018. The recognition of litchi

clusters and the calculation of picking point in a nocturnal natural environment. Biosystems Engineering 166: 44-57.

8. Yossy, E., J. Pranata, T. Wijaya, H. Hermawan and W. Budiharto. 2017. Mango Fruit Sortation System using Neural Network and Computer Vision. Procedia Computer Science, 116, 596-603.
9. Bargoti, S., & Underwood, J. P. (2017). Image segmentation for fruit detection and yield estimation in apple orchards. Journal of Field Robotics, 34(6), 1039-1060.
10. Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: fruit counting based on deep simulated learning. Sensors, 17(4), 905.
11. Wang, C., Zou, X., Tang, Y., Luo, L., & Feng, W. (2016). Localisation of litchi in an unstructured environment using binocular stereo vision. Biosystems Engineering, 145, 39-51.
12. Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture, 163, 104846.
13. He, K., Gkioxari, G., Dollár, P., & Girshick, R. 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision: 2961-2969.
14. Dutta, A., & Zisserman, A. 2019. The VIA annotation software for images, audio and video. In Proceedings of the 27th ACM International Conference on Multimedia: 2276-2279.