Journal of Science and Engineering Technology, Vol. 19, No. 1, pp. 47-53 (2023)

基於深度學習與雙目視覺水果採摘機械手臂控制之應用

黃登淵 [1*]　許景貿 [2]

[1] 大葉大學電機工程學系

515006 彰化縣大村鄉學府路 168 號

[2] 宏新科自動控制有限公司

528011 彰化縣芳苑鄉三合村工區一路 45 巷 96 號 1 樓

[*]kevin@mail.dyu.edu.tw

摘要

電腦視覺結合深度學習控制機械手臂在工業界是一個重要研究領域。目前，大部分的機械手臂控制系統需要使用昂貴的距離感測器來估測物體位置。然而，這些感測器的價格高昂，限制了機械手臂系統的應用。因此，本文提出一個結合深度神經網路模型和雙目視覺的控制系統，用於水果採摘應用。這個系統包括了使用 Yolov5 目標偵測模型判斷物體座標位置，並使用 OAK-D 雙目相機進行雙目深度估測來估測物體距離。實驗結果證實，本文所提出的方法可以有效地控制機械手臂，並且可以減少使用昂貴感測器的成本，進而提高了機械手臂系統的泛用性。

關鍵詞：人工智慧，雙目視覺，深度學習，機械手臂

Application of Robot Arm Control System for Fruit Harvesting Based on Deep Learning and Binocular Vision

DENG-YUNG HUANG[1*] and JING-MAU SHIU[2]

[1]Department of Electrical Engineering, Da-Yeh University

No.168, University Rd., Dacun, Changhua 515006, Taiwan, R.O.C.

[2]Hong-Sin-Ke Automatic Automation Co., Ltd.

No. 96, Ln. 45, Gongqu 1st Rd., Fangyuan Township, Changhua County 528011, Taiwan, R.O.C.

[*]kevin@mail.dyu.edu.tw

ABSTRACT

The use of computer vision and deep learning techniques is essential for effective robotic arm control systems. Current robotic arm control systems rely on expensive distance sensors to determine the location of objects. The high cost of these sensors is a barrier to the widespread application of robotic arm control systems. Therefore, this paper proposes a control system that combines a deep neural network model and binocular vision for fruit-picking robots. In this system, the YOLOv5 object detection model is used to determine object coordinates, and an OAK-D stereo camera is used for depth estimation. The experimental results indicated that the proposed system was successful in

effectively controlling a robotic arm by using low-cost sensors, making it a suitable option for robotic arm control systems.

***Key Words*:** Artificial intelligence, binocular vision, deep learning, robotic arm

一、文獻探討

近年來深度學習技術的快速發展，人們對於利用深度學習技術來解決機器視覺問題的興趣日益增加。同時，農業產業的快速發展，自動化的水果採摘機器人也成為熱門的研究方向。本文將探討基於深度學習與雙目視覺的水果採摘機器人的特點，並分析相關研究成果。

電腦視覺應用廣泛，其中較為常見的包括影像分類 [3, 11]、物體偵測 [8, 9, 12]。物體偵測是指在數位影像或影片中自動地識別出不同物體的位置，並對其進行標記和分類。它是電腦視覺領域的重要研究方向之一，應用於自動駕駛、智慧安防、智慧監控等眾多領域。近年來，深度學習的興起推動了物體偵測技術的發展。這些方法在準確率和速度方面取得了很大的進展，並且已經被廣泛應用於各種場景中。然而，物體偵測依然面臨著很多挑戰和問題，如遮擋、低對比度、光線變化、多尺度等，未來仍需要更加高效和精確的方法來應對這些問題。

對於 3D 環境重建 [4]、機器人 [2]、自動駕駛 [10]等，深度感測是必要的技術。在機器人研究領域中，執行探索任務的重要關鍵是影像深度的配合。而為了進一步提高模型的準確度，需要進行微調（Fine-tuning）[1]，即針對特定應用場景對現有的物體偵測模型進行微調。通過微調可以將模型在特定領域的表現提高到一定水平，提高模型的實用性。另外，為了使模型能夠在邊緣設備上運行，需要進行模型佈署 [5, 6]。在邊緣運算裝置中，由於硬體資源和能耗的限制，需要對模型進行優化和壓縮，以提高運行效率。目前常用的模型佈署方法包括 TensorFlow Lite、PyTorch Mobile 等。

機械手臂結合影像識別一直以來也都是機器人領域研究的問題 [7]。機器視覺與機械手臂結合應用於許多場域，例如機器人自動化操作、工業生產等。這項研究主要著重於將機器視覺與機械手臂結合，實現機器人對複雜環境的自動操作。其中，機器視覺主要用於對環境進行感知和識別，並提供相應的物體位置信息。而機械手臂則可以根據物體位置信息進行精確的抓取和操作。此外，機械手臂研究還關注了運動軌跡的優化，以實現更加高效和準確的操作。

二、研究方法

本文所提的控制應用系統架構如圖 1 所示，本文以溫室中的小黃瓜為檢測目標，透過蒐集現場實際小黃瓜影像後做影像資料增強來擴充訓練影像資料。使用影像標註工具來標註目標物位置後產出針對溫室環境中小黃瓜影像的客製資料集，後輸入 Yolov5 模型進行 fine tuning training。經重新訓練後的輕量化權重模型將嵌入 Nvidia Jetson Nano 裝置健行邊緣運算後控制機械手臂抓取目標水果。

在本文中，於實際現場蒐集了 81 張小黃瓜影像，並透過影像增強，如圖 2 所示。旋轉原影像由-15˚~15˚ 每 5˚取一張，共 6 張樣本數，可將原始 81 張影像樣本擴充至 486 張。

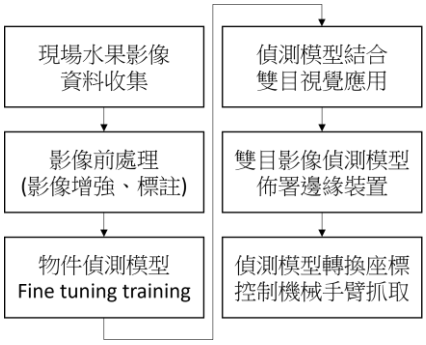

圖 1. 本文所提機械手臂影像控制系統架構圖

圖 2. 經影像旋轉後所得影像增強以擴充樣本數

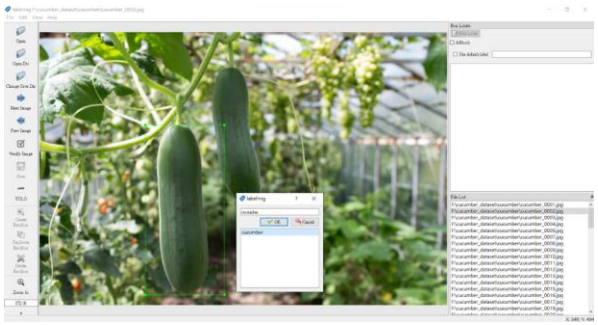

圖 3. LabelImg 影像標註結果

透過影像標註工具 LableImg 可對原始影像進行標註，如圖 3 所示。將前處理後的擴充影像樣本做標註定位，並產出以 Yolov5 訓練格式的目標影像訊息檔，Yolov5 的標註檔為 txt 檔，標註格式為影像中標註有小黃瓜位置候選框的四個數值（x, y, H, W），其中：x, y 分別為物體中心點的 x 與 y 座標，H, W 分別為物體的高與寬。

本文為將目標檢測模型嵌入邊緣運算裝置中執行，使用的目標檢測模型需在檢測速度與精準度中取得一個平衡。我們以 Yolov5s 模型作為 fine tuning 的 base model。由於 Yolov5 預訓練模型沒有小黃瓜這個類別，因此需使用本研究自製的小黃瓜資料集提供訓練，再做二次模型調校。以 388 張訓練集以及 98 張驗證集小黃瓜影像資料，設定預訓練權重模型 Yolov5s，以 300 epochs、16 batch size、416 x 416 Image Size 進行二次訓練，最終訓練結果為 0.98 mAP_0.5，訓練驗證結果輸出圖如圖 4 所示。

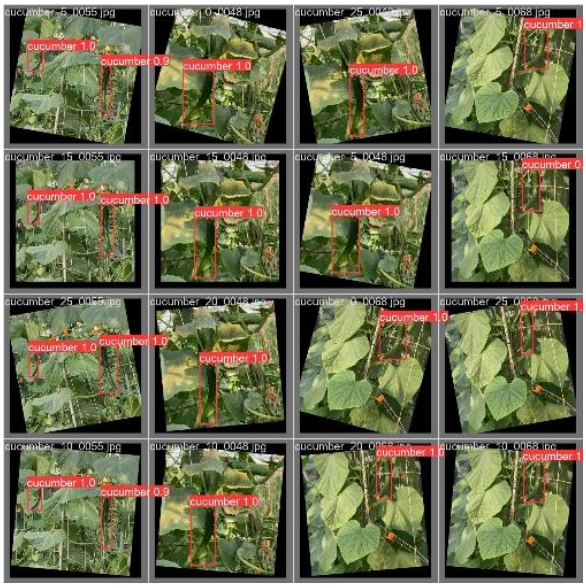

圖 4. 自製訓練資料集的圖片

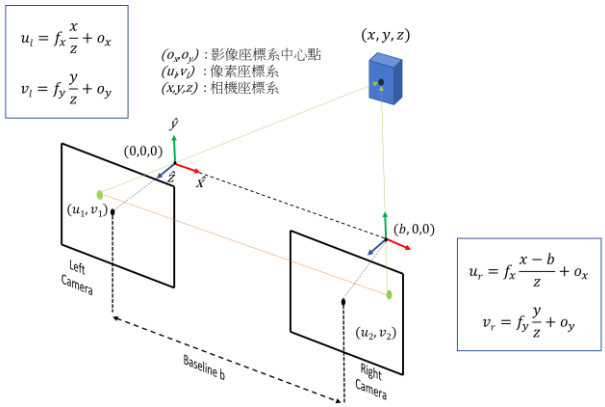

圖 5. 雙目影像座標與物體深度轉換示意圖

經訓練驗證後，將訓練完成的目標檢測模型結合雙目立體視覺做應用，本文使用 OAK-D 雙目相機實現深度立體視覺擷取空間影像資訊，透過 OAK-D 左右兩顆 100 萬畫素黑白 CCD (Charge Couple Device)，雙目相機影像畫面透過 Yolov5 目標偵測模型於各自畫面檢測出小黃瓜的位置(u, v)，透過深度計算公式換算，可取得其相對於雙目相機的深度值 Z。最小立體深度距離為 19.6 cm，其影像深度計算如圖 5 所示。

由圖 5，其中 f_x, f_y, u, v, O_x, O_y 是以像素為單位的已知參數。由於圖像感測器中的基本感光元件形狀有可能不是正方形的，因此在水平與垂直方向的焦距 (f_x, f_y) 有可能不相同。(o_x, o_y) 為光軸與像平面相交的點。相機中心之間的線稱為基線。(u_l, v_i) 和 (u_r, v_r) 分別為世界座標 (X, Y, Z) 在左、右圖像平面中的投影點。藉由上述方程式得到以下 X, Y, Z 的值 ：

$$X = \frac{b(u_l - o_y)}{(u_l - u_r)} + o_y \tag{1}$$

$$Y = \frac{bf_x(u_l - o_y)}{f_y(u_l - u_r)} + o_y \tag{2}$$

$$Z = \frac{bf_x}{(u_l - u_r)} \tag{3}$$

透過其 ROI（Region of Interest）於左右兩影像中的視

差可計算出世界座標系中的深度值 Z，此為物體的深度值。由於視差和深度成反比，隨著視差的減小，深度會呈指數級增長，具體取決於基線和焦距。如果視差值接近於零，視差的微小變化在深度值的估算上會產生比較大的變化。

Yolov5s 模型在經過深度學習主機完成二次訓練後，再結合雙目相機演算，可將訓練完成後的模型嵌入到邊緣運算裝置，本文採用的邊緣運算裝置為 Nvidia 的 Jetson Nano，選用此模組的原因是因為它的體積小，有利於結合機械手臂的操作。此外，本模組提供 128 個 NVIDIA CUDA 核心，為影像處理及物體辨識提供足夠的運算效能。基於 Jetson Nano 的效能支持，透過嵌入 Yolov5 模型檢測利用 OAK-D 雙目相機檢測小黃瓜位置，獲取小黃瓜於圖像座標轉換至世界座標 X,Y,Z 值後，通過 TCP/IP 將座標資訊經處理後轉換傳輸至機械手臂以下達抓取指令。

將邊緣運算裝置及雙目相機安裝於手臂末端執行器後，需先進行手眼校正來取得相機座標與機械手臂座標的轉換關係式，以眼在手上（Eye in hand）的方式移動機械手臂，對校正板取多張角度來進行手眼校正，如圖 6 所示。

透過 OpenCV 影像函式庫先針對棋盤格影像做相機校正取得其內參、相機位移向量及旋轉矩陣後，再同時獲取當前機械手臂末端的座標位置及姿態，即可求出相機與機械手臂末端的座標轉換關係。假設機械手臂末端執行器的初始位置為ε_1，相機的初始位置為 C_1，同時假設相機 C_1 相對於機械手臂末端執行器ε_1的變換矩陣為 $X_{C_1}^{\varepsilon_1}$。由於ε_1是相對於機械手臂底座 R 的座標，因此若 ε_1 已知，我們即可求得 C_1 到標定板 H 的座標，其如圖 7 所示（取自 https://www.torsteinmyhre.name/snippets/robcam_calibration.html）。

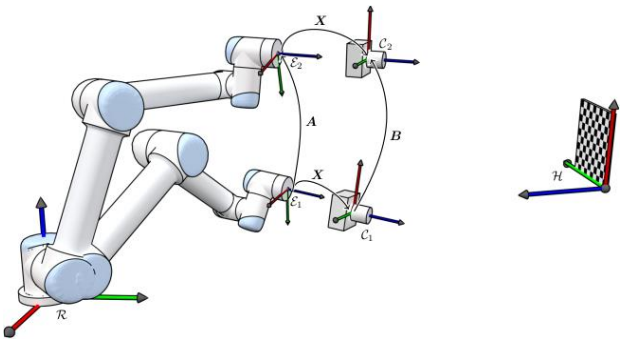

圖 7. 眼在手上（Eye in hand）標定示意圖

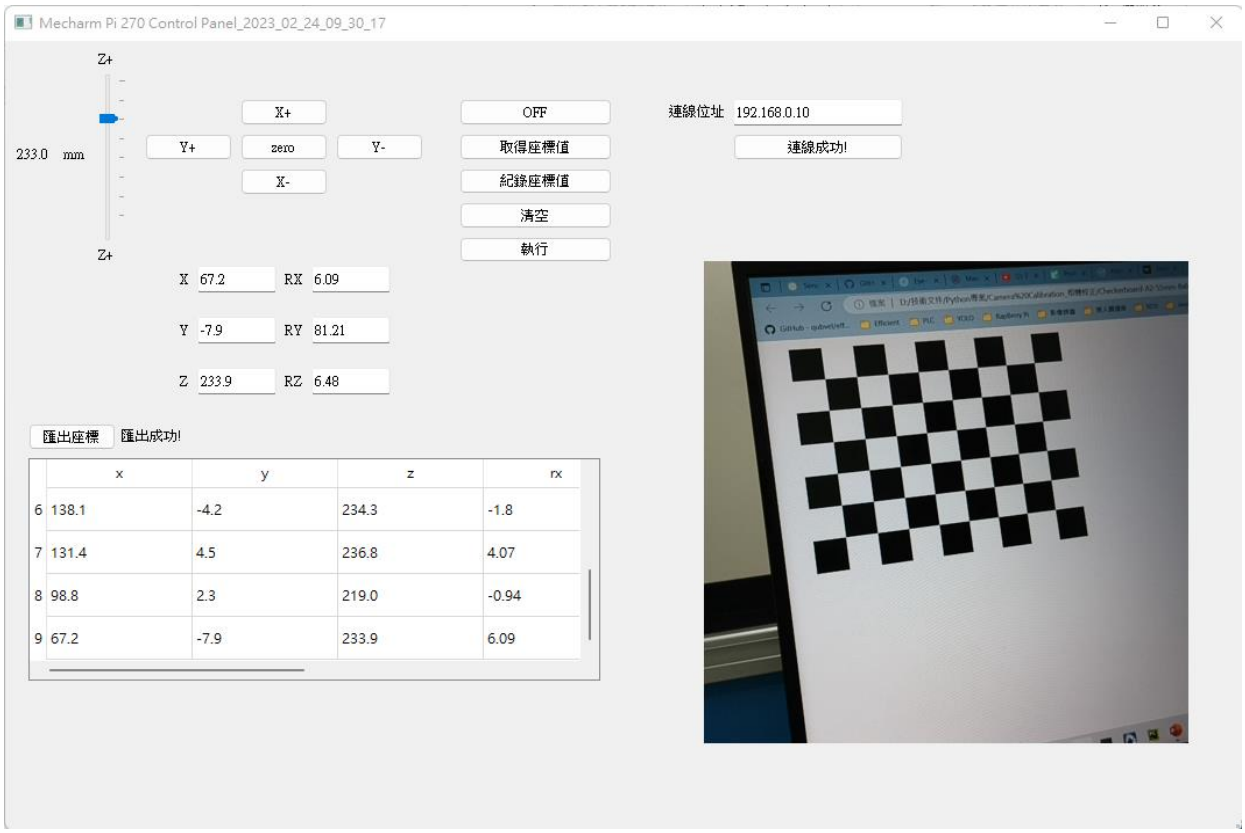

圖 6. 本文所開發的手眼校正軟體界面

假設機械手臂末端執行器在下一個時間移動到 ε_2 的位置，相機的位置也跟著移動到 C_2 的位置（注意此時相機相對於末端執行器的位置仍是固定的）。另外，相對於機械手臂底座 R，標定板 H 的位置是固定不動的。因此，透過以上的觀察，我們可以得到以下的變換式，其如式(4)至式(5)所示。

$$T_H^R = T_{\varepsilon_1}^R \cdot X_{C_1}^{\varepsilon_1} \cdot T_H^{C_1} \tag{4}$$

$$T_H^R = T_{\varepsilon_2}^R \cdot X_{C_2}^{\varepsilon_2} \cdot T_H^{C_2} \tag{5}$$

其中符號：T_A^B 表示座標系{A}相對於座標系{B}的變換矩陣，其通常可表示為 $T = \begin{bmatrix} R & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}_{4\times4}$，其中 R 與 **t** 表示兩個座標系之間的旋轉矩陣與平移向量。X 變換矩陣的定義與 T 變換矩陣相同，只是它被用來描述相機與末端執行器之間的座標變換而已。由於式(4)等於式(5)，因此我們可以得到式(6)。

$$T_{\varepsilon_1}^R \cdot X_{C_1}^{\varepsilon_1} \cdot T_H^{C_1} = T_{\varepsilon_2}^R \cdot X_{C_2}^{\varepsilon_2} \cdot T_H^{C_2} \tag{6}$$

另外，由於相機相對於末端執行器是固定的，故可知 $X_{C_1}^{\varepsilon_1} = X_{C_2}^{\varepsilon_2} = X$。另外，由於 $T_A^B = \left(T_B^A\right)^{-1}$，因此可再推得式(7)至式(8)的結果，其如下式所示。

$$T_R^{\varepsilon_2} \cdot T_{\varepsilon_1}^R \cdot X = X \cdot T_H^{C_2} \cdot T_{C_1}^H \tag{7}$$

$$T_{\varepsilon_1}^{\varepsilon_2} \cdot X = X \cdot T_{C_1}^{C_2} \tag{8}$$

因此，由上述式(8)可以得到手眼標定轉換公式 AX=XB。藉由標定結果，可以得到相機與末端執行器間的變換矩陣 X，藉此可以取得以上兩者間的旋轉矩陣 R 與平移向量 **t**。在本系統偵測到目標水果後，即可將偵測到的像素座標 (u,v) 與雙目相機測定的深度值，轉換為機械手臂末端執行器的座標值 (X,Y,Z)，藉此可透過控制電路發出控制訊號命令機械手臂末端執行器來執行採摘目標水果的命令。

三、結果與討論

本文實驗系統採用 Nvidia Jetson Nano 邊緣運算裝置執行機械手臂末端執行器夾取目標水果，開發環境為 Python 3.7.9，安裝函式庫 Pytorch 1.7.0、opencv-python 4.1.1。硬體使用 MechArm 270 pi 機械手臂結合 OAK-D 雙目相機。

本文透過實際至溫室場域拍攝小黃瓜影像以客製化資料集，針對 Yolov5 的預訓練權重進行 Fine tuning training，以 300 epochs、16 batch、416x416 Image Size 進行二次訓練，最終訓練結果為 0.98 mAP_0.5，訓練設備為 Nvidia RTX2080 Ti，訓練耗時 32 分鐘，以 Fine tuning training 方式進行二次訓練，不僅訓練耗時低且預測準確度及檢驗穩定度都達到不錯的效果，如圖 8 驗證結果，且針對不同水果再訓練也有同樣的效果。

再將其偵測結果結合 OAK-D 雙目相機，計算目標物體的影像深度進而取得目標物體的 3D 空間座標值，其如圖 9 所示。

圖 8. 本系統所提方法目標水果偵測結果

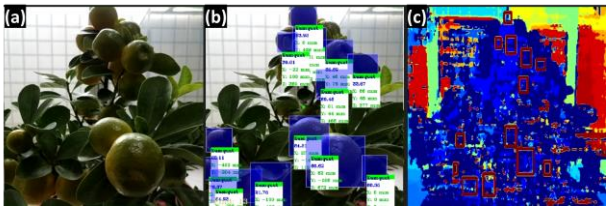

圖 9. 偵測模型結合 OAK-D 雙目相機結果：
(a)原圖、(b)模型辨識金桔的空間座標位置、
(c)OAK-D 雙目相機深度影像

當獲取影像座標後，將其透過手眼標定取得的旋轉矩陣與位移向量關係轉換為機械手臂的座標系位置，機械手臂座標系以手臂末端執行器相對於基座坐標系的位移向量及歐拉角的旋轉角度組成，所以在手眼校正函數中，需針對末端執行器的歐拉角（Euler angle）將之轉換為旋轉矩陣 R，並以此來與相機做姿態校正，藉以獲得相機與機械手臂末端執行器間座標系的轉換關係。

在邊緣運算裝置的選用中，一開始本文採用 Raspberry Pi 4B + Yolov5 作為測試結果驗證，其結果如表 1 所示。表中的 Yolov5 s 與 m 分別表示是 small and medium model 之意。由表中可知，Raspberry Pi 4B 的執行速度連 1 fps 都達不到，顯示無法應用於即時影像偵測。透過調整，本文改採用 Nvidia Jetson Nano，相較於 Raspberry Pi 4B 實驗板，Nvidia Jetson Nano + Yolov5 至少可以達到 5 fps 以上的幀數，顯示其可即時應用於實際水果採摘系統的物體偵測上。

經過實驗結果評估，本文整合視覺深度學習偵測系統與機械手臂通訊控制系統，並將之佈署到邊緣運算裝置 Nvidia Jetson Nano 上。透過該整合系統，再結合輕量型深度學習物體偵測模型，可以達到低耗能、高精度的檢測結果，這是本文研究的最主要目的與實現的成果。

四、結論

本文提出應用 Fine tuning training 對 Yolov5 模型做二次訓練，針對不同的水果客製不同的訓練集，以進行基於深度學習的物體偵測模型建模，藉以達到不同場景應用之目的。為了成功地將開發的系統佈署於邊緣運算裝置，需要考慮許多因素，例如訓練數據的質量和數量、模型的複雜度和性能、運算裝置的計算能力等等，最後將其嵌入至邊緣運算裝置上以實現輕量化負載之目的。 物體偵測模型在與雙目相機和機械手臂控制的配合下，面對不同的使用情境也能夠展現快速且良好的泛用性。

致謝

本文承科技部產學合作計畫「應用視覺深度學習於水果採摘機器人智慧控制盒之開發，NSTC 111-2622-E-212-003-」補助完成，在此表達十分感謝之意。

表 1. 針對 Raseberry pi 4b 與 Nvidia Jetson Nano 在 Yolov5 不同權重大小下執行速度測試比對結果

	Raspberry pi	Nvidia Jetson Nano
Yolov5 s	0.33 fps	5.28 fps
Yolov5 m	0.13 fps	3.79 fps

參考文獻

1. Hespeler, S. C., H. Nemati and E. Dehghan-Niri (2021) Non-destructive thermal imaging for object detection via advanced deep learning for robotic inspection and harvesting of chili peppers. *Artificial Intelligence in Agriculture*, 5, 102-117.
2. Horn, B., B. Klaus and P. Horn (1986) *Robot vision*, 1-3, MIT press, Cambridge, MA.
3. Krizhevsky, A., I. Sutskever and G. E. Hinton (2017) Imagenet classification with deep convolutional neural networks. *Communicaitons of the ACM*, 60(6), 84-90.
4. Kuhnert, K. D. and M. Stommel (2006) Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing.
5. Li, D., X. Wang and D. Kong (2018) Deeprebirth: Accelerating deep neural network execution on mobile devices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1),2322-2330.
6. Ota, K., M. S. Dao, V. Mezaris and F. G. D. Natale (2017) Deep learning for mobile multimedia: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s), 1-22.
7. Rai, N., B. Rai and P. Rai (2014) Computer vision approach for controlling educational robotic arm based on object properties. 2nd International Conference on Emerging Technology Trends in Electronics, Communication and Networking, Surat, India.
8. Redmon, J., S. Divvala, R. Girshick and A. Farhadi (2016) You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV.
9. Ren, S., K. He, R. Girshick and J. Sun (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
10. Ribler, R. L., J. S. Vetter, H. Simitci and D. A. Reed (1998) Autopilot: Adaptive control of distributed applications.

Proceedings of the Seventh International Symposium on High Performance Distributed Computing (Cat. No. 98TB100244), 172-179. Chicago, IL.

11. Simonyan, K. and A. Zisserman (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
12. Sermanet, P., D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

收件：112.03.03　修正：112.03.16　接受：112.03.27