基於深度學習之視障者影像識別輔助系統

Visual Recognition Image Recognition Assistant System Based on Deep Learning

鄭凱陽 [1]　劉遠楨 [2]
CHENG, KAI YANG[1]　LIU, YUAN CHEN[2]
[1] 國立臺北教育大學 資訊科學研究所 研究生
[1] National Taipei University of Education Graduate School of Computer Science Student
E-mail：g110632008@grad.ntue.edu.tw
[2] 國立臺北教育大學 資訊科學研究所 教授
[2] National Taipei University of Education Graduate School of Computer Science Professor
E-mail：liu@tea.ntue.edu.tw

摘要

為了幫助對視障者在面對未知的環境，除了運用有限範圍的觸覺及聽覺之外，還能借助科技的幫助，快速且方便的了解周遭的空間感，本研究以人工智慧與 Tensorflow 深度學習框架為基礎提出一套視障者影像識別輔助系統，該系統是可穿戴式裝置，可以將鏡頭掛於胸口處或任意地方，按下拍照擷取盲人視角的影像透過 USB 連接到 Raspberry pi 由處理器進行深度學習影像識別後提供目標物件名稱、數量、相對位置的語音輔助提示，實驗結果顯示針對廁所場域中的物件，使用自行收集的圖片與自行訓練的深度學習模型為該系統影像識別核心，從擷取影像到輸出語音提示只需要一秒時間就能為視障者提供服務，幫助視障者透過影像視覺科技獲取更多有用的資訊，為生活帶來更多安全與便利。

關鍵字：深度學習、Tensorflow、視障者、物體識別、樹莓派

Abstract

For helping people who is blind to face unknown environment with hearing and touching merely, new technology can support not only fast but also convenient service. This research combines with AI and tensorflow in a system of image recognition in a device to serve those blind people, this device is wearable, which is carried on a body. When the person presses the button of device, it will take a photo which represent the person's vision. The photo will be taken to Raspberry Pi form USB by processor which is processing deep learning and image distinguish. After this process of deep learning and image distinguish, the device will provide prompt voice to remind the person with information of target's name, amount as well as opposite location. This result of research focus on special object in bathroom. For example,

toilets, skins and squat toilets. This data which be collected and the model of deep learning which train with itself are the core in this system. The device goes through from taking photo to supplying voice prompt is just in one second. This device can make invisible people have a better life with image vision tech.

Keywords：Deep Learning、Tensorflow、Blind、Object Detection、Raspberry Pi

壹、 前言

根據內政部統計，全台視障者約六萬名，以國際導盲犬聯盟建議理想視障者與導盲犬比例約 1:100 計算，台灣至少需要 600 隻導盲犬，但台灣目前投入服務導盲犬僅 42 隻，遠遠低於國際理想標準[1]。因此現階段在導盲犬數量嚴重不足與視覺障礙者不斷增加的情況下，迫切需要開發相關能夠快速普及化的儀器來協助這些視障朋友。

視障者在未知的新環境下，除了運用有限範圍的觸覺及聽覺之外，對於周遭的事物一無所知，雖然視障者可以尋求周圍視力正常的人協助，但這會影響視障者的獨立性，且在人口稠密度較低的地區或周圍沒有他人情況下無法尋求幫助，因此完全限縮生活圈及社交能力，到台北市立啟明學校，透過訪談了解一位視障新生，必須花幾個月的時間透過旁人一對一的協助，來認識陌生校園從而依靠記憶建立心理地圖。另外，還面臨行動中對方位的確定與把握不易、環境過度複雜適應不來、環境隨意改變無法預知等問題。因此，若能透過輔助裝置協助快速了解周遭環境，將為視障朋友帶來非常大的便利性。

貳、 文獻探討

2.1 樹莓派

Raspberry Pi 是由英國的 Raspberry Pi Foundation 所開發，其研發目的是希望提供廉價的硬體及自由軟體來促進學校在電腦科學教育的推動[2]，卻也因為它便宜具競爭力的價格，擁有強大的效能，很快地受到廣泛的運用。

2.2 Tensorflow

Tensorflow 是 Google 於 2015 年 11 月在 GitHub 上開源的一套深度學習框架。使用 Tensorflow 建立的深度學習模型應用場景相當廣泛，包括電腦視覺、自然語言處理、資料數據分析等，早已相當依賴深度學習模型。

2.3 CUDA (Compute Unified Device Architecture)

CUDA(Compute Unified Device Architecture) 是 NVIDIA 研發的通用並行計算架構，在深度學習的運算過程中對於提升計算效率有非常大的幫助[3]。

2.4 cuDNN (cu Deep Neural Networks)

cuDNN(cu Deep Neural Networks)是用於深度神經網絡的 GPU 加速函式庫。提供經過調校優化常用於 DNN 應用的常式，例如：CNN、Pooling、Softmax、ReLU 等，使用 GPU 訓練模型，一般會採用這個加速函式庫[4]。

2.5 CNN (Convolutional Neural Networks)

卷積神經網路(Convolutional Neural Networks)擺脫了傳統機器學習方法預處理及構造特徵的繁瑣過程，同時大幅減低了因角度、遮擋等因素造成的誤檢和漏檢，複雜場景的適應性更強，為目前圖像訓練中提取特徵的主流技術，其主要架構包含:卷積層、池化層、全連接層[5]。

2.6 TTS (Text-to-Speech)

TTS 文字轉語音合成技術(Text-to-Speech)是透過電腦處理，對文字進行即時分析轉換，在其特有智慧語音控制器作用下，生成對應的合成語言。

2.7 Opencv

Opencv 是一個開源函式庫，包含了 500 多個用於圖像和影片分析的優化算法。為一套跨平台函式庫。

2.8 視障者與科技的應用

目前市面上有客服人員結合手機 app 的應用，視障者使用手機傳輸影像與客服人員進行語音或視訊通話，透過客服人員的雙眼給予視障者協助。

英國牛津大學（University of Oxford）研發一款智慧型眼鏡，鏡框備有紅外線攝影鏡頭與微型電腦，可把鏡頭前的障礙物化為簡單線條投射在鏡片上，讓非全盲視障人士透過強化後的線條分辨出眼前的物體。

參、 系統架構設計

本研究提出使用 Logitech C525 鏡頭當作擷取影像的裝置，以視障者視角去做影像辨識，將眼前景象輸入物體檢測深度學習系統，該系統裝載於攜帶方便節能的樹莓派 3B+主板，將深度學習運算過程在本地裝置運行，不受外在的網路環境限制，採用自製數據庫，該圖片庫包含 4 類目標，5269 張圖片，使用神經網路 MobileNet 搭配速度較快的物件檢測技術 SSD（Single Shot Multibox Detector），將檢測結果包含物體名稱、數量、物體相對位置資訊透過 Espeak TTS 引擎轉換成語音告知視障者，透過裝置上的鏡頭可以描述周遭的世界，在 Kyu-Dae Ban、Chen,Ya-Hsin 等人[6][7]研究基礎上更進一步識別廁所中所存在的物體，使其能更清楚知道目前所在空間環境，幫助盲人『聽』見世界。系統使用示意圖如圖 1。

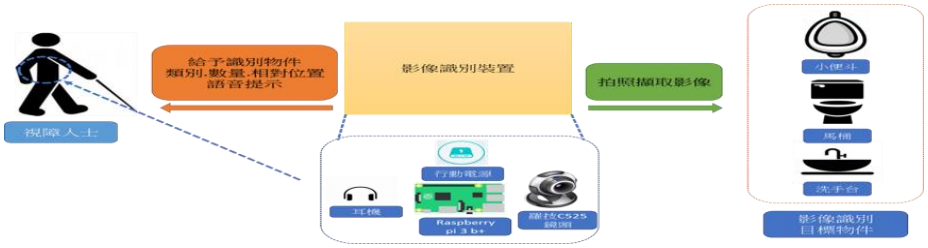

圖 1 系統使用示意圖

一、 系統架構

Google 開源的 Tensorflow 為最受世界開發者和研究人員歡迎的深度學習框架。由於 Python 有大量用於機器學習的簡單易用的工具包，Tensorflow 搭配 Python 為深度學習開發的主流方式

本研究使用 Python + Tensorflow 的架構進行學習和訓練，從而保證訓練的準確性和高效性，系統分為兩部分第一部分為數據準備與模型訓練，流程圖如圖 2，第二部分為識別語音系統設計。

圖 2 數據準備與模型訓練流程圖

3.1 數據準備與模型訓練

3.1.1 圖片數據準備

要完成盲人物件識別輔助系統，必須先實現物件的辨識，如馬桶、洗手台等目標物件。這部分我們將使用 Google 於 2017 年 6 月發布的基於 Tensorflow 的 Object Detection API 來開發訓練自己的物件識別模型。此 API 提供目前許多主流的預訓練模型供開發者使用如 faster_rcnn、ssd_mobilenet、faster_rcnn_nas、mask_rcnn_inception 等，這些模型的準確性、識別時間、訓練速度上都略有不同，考量到盲人對於物件的識別需要在最短的時間內給予語音的輔助以及樹莓派裝置本身運算能力有限的情況，本研究採用基於 CNN 的深度學習模型 ssd_mobilenet_v1_coco，此模型專門設計搭載於行動裝置，有識別速度較快準確性適中的特性。

選定好預訓練模型之後，就要開始準備訓練模型所需的圖片數據，因為識別物件的數據取得不易，且各國常見的馬桶、小便斗、洗手台等款式也會因為國家與地域性有所差異，如我們在台灣廁所常見的蹲式馬桶，在西方世界幾乎看不到，為了符合在台灣場域的適用性與提升識別的準確性，本研究實地從台灣的各個地方如台北捷運、高雄捷運、台灣鐵路、百貨公司、校園、餐廳等地收集了 5269 張圖片數據，使用 LabelIMG 經過手動標記後完成了原始的訓練數據，數據包含”urinal”、”toilet”、”asiantoilet”、”sink”四類，分別代表小便斗、坐式馬桶、蹲式馬桶、洗手台，這四種類別為在識別系統中需要找出的物件，標記後的圖片會產生 xml 文件，內容包含圖片大小、彩度、標記標籤、標記位置等資訊供訓練模型使用，接著我們必須將這些數據轉換為 Tensorflow 可以使用的.record 格式，才能進行模型的訓練。

3.1.2 模型訓練

準備好圖片數據之後，接著進行模型訓練，此研究識別系統模型訓練基於

Win10 作業系統，CPU 為 i5 8300H、RAM 20G、 GTX 1050 4G 顯示卡，由於模型的訓練需要大量且複雜的計算，強調平行化的計算能力，因此本研究使用 Tensorflow 的 GPU 版本進行訓練，NVIDIA 所生產的 GPU 支援 CUDA 為一個並行計算架構，可使繪圖處理單元（GPU）的運算能力大幅提升，一般我們在使用 CUDA(Compute Unified Device Architecture)架構還會搭配 cuDNN(cu Deep Neural Networks)使用，cuDNN 是用於深度神經網絡的 GPU 加速函式庫。它強調性能、易用性和低記憶體消耗。透過上述 GPU 的模型訓練架構，與原先使用 CPU 訓練模型縮短了 10~15 倍時間，大大的提升模型訓練效率。

3.2 識別語音系統設計

本研究識別語音系統搭載於樹莓派 3B+裝置，系統分為三大部分組成，第一部分是前端影像擷取與圖像預處理，第二部分是導入預處理後影像從模型輸出識別出結果，第三部分解析圖像中各類別數量、位置透過 TTS 引擎轉換語音輸出給予視障者協助。

3.2.1 影像擷取與預處理

視障者攜帶的樹莓派裝置連接一組 Logitech C525 鏡頭，以視障者視角透過 Opencv 擷取圖像，對擷取到的圖像進行正規化與 Median Filter 等圖像預處理，增強圖像品質，以利提升識別準確性。

3.2.2 模型輸出識別出結果

預處理完成的圖像輸入到前一節所訓練好的深度學習模型 ssd_mobilenet_v1，進行檢測與分類，識別出圖像中的坐式馬桶、洗手台、小便斗、蹲式馬桶四個種類物件。

3.2.3 解析圖像透過 TTS 引擎轉換語音輸出

深度學習模型 ssd_mobilenet_v1 將識別結果輸出後，需要分析出該圖像中包含多少個目標識別物件，以及各個目標識別物件的中心座標依據擷取圖像的尺寸判斷該物件的相對位置，本研究將圖像分為左半邊及右半邊，例如左邊有兩個馬桶右邊有一個洗手台，將上述資訊輸出文字透過 ESpeak TTS 引擎轉換成語音，給予視障者輔助。

肆、 實驗結果與討論

本研究開發的物體檢測深度學習語音輔助系統，其所使用的硬體規格如表 1 所示。

表 1 硬體規格

主板	Raspberry PI 3B+
CPU 型號	Broadcom BCM2837 Quad Core 1.4GHz
記憶體	1GB
記憶卡容量	32GB
作業系統	Raspbian Stretch 2018-11-13
鏡頭	Logitech C525

該系統是可穿戴式的裝置，可以將鏡頭掛於胸口處按下拍照擷取盲人視角的影像透過 USB 連接到 Raspberry pi 由處理器進行深度學習影像識別後提供目標物件名稱、數量、相對位置的語音輔助提示。

4.1 模型準確率評估

本研究所自行收集的數據集總計有 5269 張圖片，以 9:1 的比例分拆為訓練集與測試集，本節我們使用 526 張圖片做為測試集驗證本模型準確率，在訓練模型設定迭代次數(Training step)是 130000 次隨著迭代次數(Training step)增加，每 10 分鐘保存一個 checkpoint，針對每個 checkpoint 使用測試集驗證該模型的準確性，我們使用 tensorboard 工具得到測試集對這個模型評估的可視化介面，得到使用目標檢測(object detection)常見的準確率評估指標 mAP(mean average precision)，在迭代次數(Training step)為 92222 次時，模型達到最佳準確率為 91.66% mAP，如圖 3 所示。

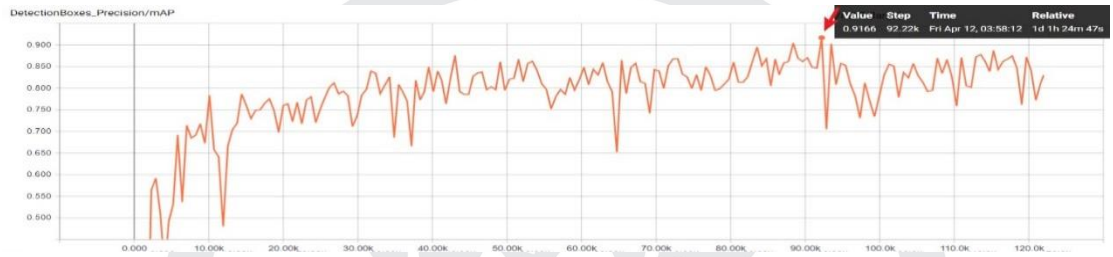

圖 3 目標檢測準確率 mAP

4.2 實際場景識別成果與準確率

本研究實際於廁所針對目標物件進行實驗，透過自行收集圖片數據與自行訓練的深度學習模型進行識別，正確的將目標物件蹲式馬桶、洗手台、小便斗、坐式馬桶於圖像中找出給予分類標籤並且框定位置，實驗結果如圖 2 所示。

圖 2 物件識別結果

表 2 為實際於校園、捷運站廁所，以正面角度，距離目標物件 110cm，開始每向後退 10cm 擷取一張圖像，針對各物件擷取 10 張圖像，計算成功識別、置信度、漏檢率、誤檢率數據，平均 4 個物件與兩個場景數據得到 98.75%的正確識別率。

表 2 正面角度-無遮擋識別準確率

正面角度-無遮擋	正確率	置信度	錯誤結果	
			漏檢率	誤檢率
校園-小便斗	100%	96.7%	0%	0%
捷運-小便斗	100%	97.6%	0%	0%
校園-坐式馬桶	100%	94.1%	0%	0%
捷運-坐式馬桶	100%	97.71%	0%	0%
校園-蹲式馬桶	90%	97.55%	10%	0%
捷運-蹲式馬桶	100%	87.9%	0%	0%
校園-洗手台	100%	92.33%	0%	0%
捷運-洗手台	100%	92.61%	0%	0%
平均	98.75%	94.56%	1.25%	0%

4.3 分析圖片與文字轉語音技術成果

本研究進行影像識別後，分析識別出的物件名稱、數量、座標、相對位置，其中相對位置部份我們找出圖像 X 軸方向的中心點將圖像進行垂直切割，把圖片分為左右兩邊之後，計算目標物件 bounding box 中心座標位於哪個區塊，如圖 3 所示，我們識別出圖片中左側的洗手台及右側的坐式馬桶，接著透過本研究演算法分析圖片中的上述資訊，給出文字提示左邊有一個洗手台，右邊有一個坐式馬桶，最後再透過 TTS 文字轉語音引擎將資訊以語音的方式給予視障者協助，更能了解眼前所未知的世界。

圖 3 分析目標物件相對位置

伍、 結論

本研究提出一款基於深度學習的視障者影像識別語音輔助系統，本系統搭載於樹莓派 3B+裝置上，除了有輕量節能擴充性能佳的優點之外，樹莓派主板

價格也是極具優勢，結合深度學習影像識別與文字轉語音技術，從鏡頭拍照到語音提示的輸出平均僅耗費 1 秒的時間，透過實地場景的準確性驗證達 98.75%，能夠迅速且正確的協助視障者適應環境，讓視障者獲得更多的資訊，了解眼前未知的環境，增加視障者的自主獨立性，減少因為對周遭環境的陌生、不確定性而產生畏懼，利用影像視覺科技為視障者帶來更美好的生活。未來如果能夠進一步增加數據集的種類，做到更多樣化的分類。另外，透過其它硬體的搭配及演算法的改進，告訴視障者與目標物件的距離有多少，甚至提供找尋目標物件的導航服務，將能為視障者帶來更大的便利。

參考文獻

一、 中文部分

[1]台灣導盲犬協會 Taiwna Guide Dog Association(2018)。上網日期:2018 年 12 月 26 日，檢自: http://www.guidedog.org.tw/aboutguidedog/about-1.html
[2]維基百科樹莓派(2018)。上網日期:2018 年 12 月 16 日，檢自: https://zh.wikipedia.org/wiki/樹莓派
[7]陳雅歆(2017)。視障者輔助系統之標誌偵測與識別的研究。國立雲林科技大學電機工程研究所,雲林縣

二、 英文部分

[3] Li Gang,.Li Gang.,Luo Yujun.(2013).CUDA based parallel wavelet algorithm in medical image fusion. International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA), 23-24 Dec.
[4] Yuki Ito,.Ryo Matsumiya,.Toshio Endo.(2017).ooc_cuDNN: Accommodating convolutional neural networks over GPU memory capacity.IEEE International Conference on Big Data (Big Data), 11-14 Dec.
[5] Ahmed Ali Mohammed Al-Saffar.,Hai Tao.,Mohammed Ahmed Talab.Review of deep convolution neural network in image classification.(2017).International Conference on Radar Antenna Microwave Electronics and Telecommunications (ICRAMET), 23-24 Oct.
[6] Kyu-Dae Ban.,Ho-sub Yoon.,Jaehong Kim.(2013).Public signs detection in subway station images. International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Page s: 595 – 596.