基於影像處理及深度學習的兩階段人流偵測系統

林泓邦[1]*、林仁信[2]、廖伯翔[3]
[1]財團法人車輛研究測試中心
[2]財團法人車輛研究測試中心
[3]財團法人車輛研究測試中心
*Email: hplin@artc.org.tw

摘要

本文提出了一種基於傳統機器學習和深度學習演算的人員計數方法。所提出的方法主要使用在特定公共場所，例如：公共汽車候車亭或接駁廣場等，用來分辨人流高峰或離峰的估算訊息，其優勢在有效節省管理成本與智慧運輸應用的升級。

本系統基於邊緣計算的架構設計分成前台系統和後台系統，包含下述的兩種方法。第一階段是主要用來進行人流量的概數計算並簡單分級，後台僅處理第二個級別的行人識別，在最後的實驗結果可以發現，本系統能有效地降低計算複雜度與整體的時間花費，其中，前台人流分級的正確率為94.28%，後台人數計數正確率為94.04%。

關鍵字： 人流、人數、2階段

1. 前言

全球車輛技術發展趨勢正逐漸往先進駕駛輔助系統(Advanced Driver Assistance Systems, ADAS)發展，而智慧運輸也是關鍵的一環，透過統計公共候車亭等場所的人數，可對於安全監控或營運分析的策略更有效率，監視系統用途通常用來統計特定區域的行人、遊客數量或流量，以此來達到監控或管制的目的[1]，傳統計算人流的方法可能會在入口處利用計數器手動計算，或者透過閘門式機械設備逐一計算，常見的方法為紅外線感應或旋轉門計數，但是對於公車候車亭等開放式、半開放式的場域而言，缺少固定入口來協助逐一計算，因此本研究利用影像處理方法達到智慧監控的目的。

在過去眾多的研究方法中，支持向量機(Support Vector Machine, SVM)與Adaptive Boosting (AdaBoost)是較於著名的兩種分類器方法，利用分類器這種機器學習的方法處理行人的影像辨識，再經過計數藉此達到流量的計算 [2, 3, 4]；另一方面，近年來，利用深度學習來達到影像辨識的方法亦被採用[5, 6]，深度學習的辨識率通常更高、抗遮蔽效果更好，最大的問題在於周邊的硬體運算效能要提高，成本亦會增加。

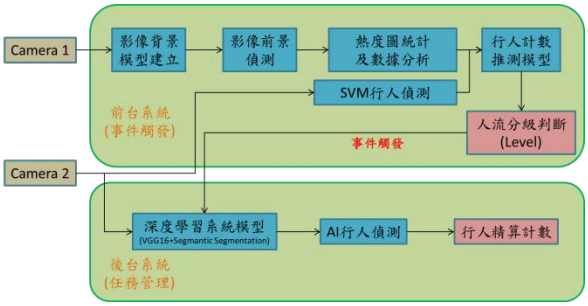
本研究主要針對候車亭在開放式或半開放式場域進行人流、人數的識別，我們採用辨識率較高的深度學習作為統計手段，而深度學習亦有許多種分類，例如YOLO[7]、Fast R-CNN[8]等，而為了解決行人在影像中交疊遮擋的問題，採用抗遮蔽效果較強的Mask R-CNN[9]演算法，Mask R-CNN的精準度高，且能夠對物體進行更精細的分割，甚至具有姿態分析功能的擴充性，但是需要更高規格的硬體與效能來處理複雜運算。

為了解決計算機處理深度學習的負擔，本論文提出一個分段式計算方法，區分前台與後台兩個部分，前台進行候車人流的密度估測，使用高斯混合模型(Gaussian Mixture Model, GMM)取出行人區塊加上SVM的行人偵測，再利用熱度圖統計估算流量等級(稀疏Level 0、擁擠Level 1)；後台則是在前台發生擁擠事件之後才會啟動，使用Mask R-CNN的語意分割演算法來進行更精細的人數計算，以如此兩階段方式來實現邊緣運算，以達到減少效能使用與時間花費的效果。

2. 系統架構

本系統具有兩支攝影機，攝影機Camera 1為俯視角度， Camera 2為側視角度，並結合前、後台影像處理與影像辨識演算法加以運算。

首先由Camera 1擷取即時影像畫面交給前台的影像處理單元處理訊號，Camera 1影像藉由影像圖學等演算法取出前景資訊，同時Camera 2的畫面也進行SVM行人偵測的處理，接著藉由進行感興趣區域(Region of Interest, ROI)區域劃分，並以熱度圖計算進行人流密度統計分析，當人流密度較高時，將影像訊號與前台影像偵測結果輸出至後台進行深度學習運算獲得人數統計資訊，系統流程圖如圖一所示。



圖一：系統架構流程圖

3. 前台系統(人流分級估算)

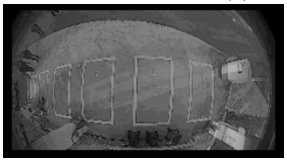
利用兩攝影機擷取兩不同角度之影像，分別為上方俯視與側方斜視，如圖二，上方俯視影像(Camera 1)主要將影像上感到有興趣的物件提取出來，並計算其物件面積比例，而側方斜視影像(Camera 2)則是以分類器演算法技術開發行人偵測，兩者偵測結果可進行後續之分級判斷。

圖二：系統之攝影機架設方式

3.1 影像背景模型建立及前景偵測

本系統將利用GMM把背景/前景分離，如圖三所示，最後再藉由圖三(c)針對ROI進行前台第一階段覆蓋面積計算。

(a) 原始影像



(b) GMM背景影像

(c) GMM前景影像

圖三

本論文主要應用於公車站牌區域，屬於室外場景，容易受時間影響，白天、下午與晚上的環境差異顯著，所以本論文採用GMM[10]建立背景，此方法可隨時間調整背景。使用C個高斯分布來描述像素隨著時間n的像素座標x灰階值 $I_{0,x}$， $I_{1,x}$， … $I_{t,x}$ ，因此由t時刻的高斯分佈組成的混合模型可以表示為：

$$P(I_{t,x}) = \sum_{k=1}^{C} \omega_{t-1,x,k} N(I_{t,x}; \mu_{t-1,x,k}, \sigma^2_{t-1,x,k}) \quad (1)$$

$$N(I_{t,x}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{(I-\mu)}{2\sigma^2})} \quad (2)$$

其中 $N(I_{t,x}; \mu_{t-1,x,k}; \sigma^2_{t-1,x,k})$ 為高斯密度函數(Gaussian probability density function)如式(2)， $\omega_{t-1,x,k}$ 為第k個高斯模型的權重值， $\mu_{t-1,x,k}$ 為平均值， $\sigma^2_{t-1,x,k}$ 為標準差，當此像素符合背景模型才進行更新，更新方式如式(3)(4)：

$$\mu_{t,x,k} = (1-\rho)\mu_{t-1,x,k} + \rho I_{t,x} \quad (3)$$

$$\sigma^2_{t,x,k} = (1-\rho)\sigma^2_{t-1,x,k} + \rho(I_{t,x} - \sigma^2_{t-1,x,k}) \quad (4)$$

其中 ρ 為平均值與標準差的學習率(learning rate)，而權重則是以此像素是否符合模型進行更新，如式(5)：

$$\omega_{t,x,k} = (1-\alpha)\omega_{t-1,x,k} + \alpha M_{k,t} \quad (5)$$

其中 α 為權重的學習率，若模型匹配成功則 $M_{k,t}$ 為1，反之則為0。

3.2 HOG特徵及SVM行人偵測

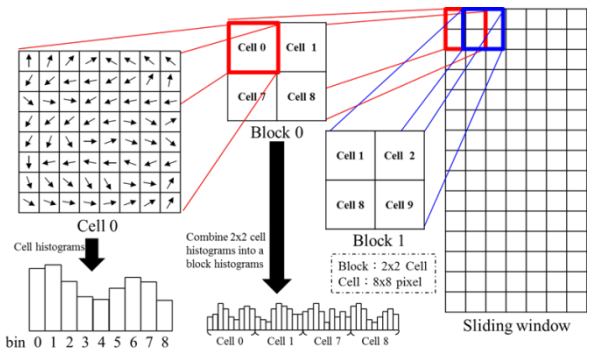
前台另外還採用了行人分類器演算法技術，使用Histograms of Oriented Gradients (HOG)[11]描述行人邊緣之相關性，對物體進行特徵擷取，並透過基於SVM之分類器訓練與分類。

方向梯度直方圖特徵計算方法首先以遮罩(mask) $Mask_x = [-1 \quad 0 \quad 1]$ 與 $Mask_x = [-1 \quad 0 \quad 1]^T$ 計算影像梯度向量：。

$$G_x(x,y) = I(x+1,y) - I(x-1,y)$$
$$G_y(x,y) = I(x,y+1) - I(x,y-1) \quad (6)$$

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$$
$$\theta = \arctan(\frac{G_y(x,y)}{G_x(x,y)}) \quad (7)$$

其中，x、y為影像座標(pixel coordinates)， G_Y 為垂直梯度向量， G_x 為水平梯度向量， G 為梯度向量振福(gradient magnitude)， θ 為梯度向量角度(gradient angle)。接續統計cell的直方圖，定義一個cell有8*8個pixel，而block有2*2個cell，將所有梯度角度0°~180°分成等距9個bin，並記錄各個cell之梯度向量。依據滑動視窗(sinding window)大小計算特徵，並建立HOG描述子區塊(HOG descriptor block)，如圖四：



圖四：HOG特徵圖示

SVM主要為以一超平面將學習樣本分成正樣本與負樣本，此平面如式(8)所示，使其任意樣本點到平面距離大於等於1，如式(9)所示。

$$w^T X + b = 0 \tag{8}$$

$$y_i(w^T X_i + b) = 0 \tag{9}$$

其中w, b分別為超平面(hyperplane)之法向量與截距。接續找出w, b滿足式(10)之約束條件，主要條件為超平面至正負樣本距離為最大值。

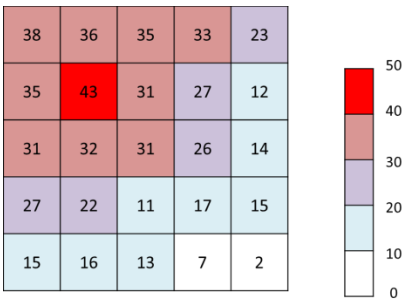
$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} L_i,\ L_i = \max[0, 1 - y_i(w^T X_i + b)] \tag{10}$$
$$\text{s.t. } y_i(w^T X_i + b) \geq 1 - L_i,\ L_i \geq 0$$

其中C為正規化係數。最後得到一組w, b，將特徵點輸入至此組參數後即可分類成正負類別，如式(11)所示。

$$\begin{aligned}(w^T X_i + b) &\geq +1, \quad \Rightarrow y_i = +1 \\ (w^T X_i + b) &\leq -1, \quad \Rightarrow y_i = -1\end{aligned} \tag{11}$$

3.3 數據分析及人流分級判斷

兩組不同角度攝影機可對應至相同ROI區域內，透過GMM方法可分離出前景影像計算出上方俯視影像之覆蓋面積，進一步利用熱度圖的方式進行區域(cell)劃分，熱度圖是一種統計的手段，其優勢在於可降低大量分散的資料並做級別的區分，而此處則是將ROI分為若干cell，並針對每個cell計算覆蓋面積來決定每個cell的強度，如圖五，進一步可定義流量參數值為式(12)，w為每個cell的權重。



圖五：熱度圖示意

$$H = \sum_{i=0}^{m} w_i * cell_i \tag{12}$$

而行人分類器則可取得側視影像偵測人數n，假設一人份的H比值為v，則可估算出側視影像之估計值C如式(13)。

$$C = n * v \tag{13}$$

最後加總進行人流分級，如式(14)。

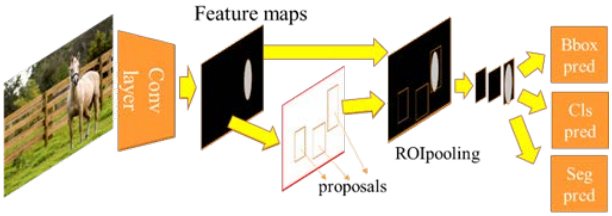
$$level = \begin{cases} 0, \text{if } normal(C + H) \leq th \\ 1, \text{if } normal(C + H) > th \end{cases} \tag{14}$$

4. 後台系統(Deep Learning人數計算)

後台接收前台分級資訊再判斷是否需進行後台運算，主要以深度學習方法進行精準的人數統計。

4.1 深度學習及行人精算計數

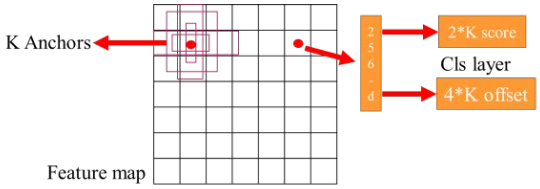
本論文使用的深度學習方法是Mask R-CNN，首先影像特徵擷取，將特徵圖輸入於Region Proposal Network 提取障礙物資訊，再對每個障礙物資訊進行分類並微調候選框，最後每個候選框生成Semantic Segmentation，如圖六所示。



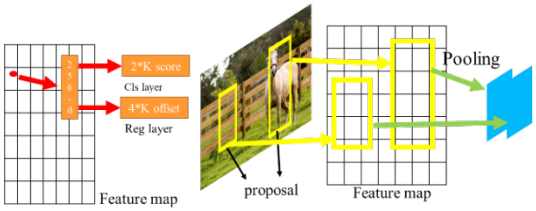
圖六：Mask R-CNN Algorithm

其演算法步驟如下：.

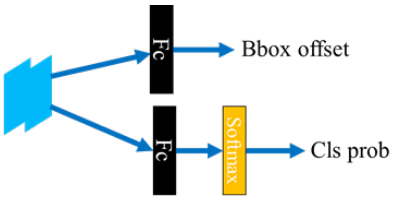
- 特徵擷取：以 ResNet-101 作為特徵擷取器
- Proposal Network：將特徵圖輸入於 Region Proposal Network 提取障礙物資訊，以 K 個 anchors 放在特徵圖的 cell 上，每個 anchor 進行預測前景機率(Cls layer)與預測 bounding box 位移量(Reg layer)，如圖七所示。
- ROI Pooling：將提案框預測位置對應回輸入影像，再從輸入影像投影到特徵圖上，最後框選到的 feature pool 成固定尺寸，如圖八所示。
- 分類與候選框微調：使用 Fully Connected Layer 預測每個提案類別(Cls prob)與框的微調值(Bbox offset)，如圖九所示。
- Semantic Segmentation：使用卷積層對每個提案進行語意分割(1~2 層 Convolution Layer)，類別分成前景與背景，如圖十所示。



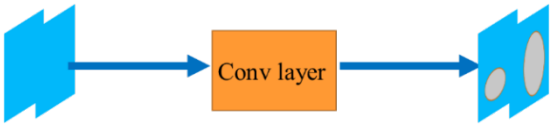
圖七：Proposal Network



圖八：ROI Pooling



圖九：Cls prob & Bbox offset



圖十：semantic segmentation

5. 結果與討論

測試所使用之攝影機解析度均為1280*720，前台與後台所使用之運算平台不同，後台由於深度學習方法的運算量較高，所以使用的是NVIDIA GeForce GTX 1660Ti配合64GB的記憶體，而前台則使用NVIDIA GTX 1050及8GB的記憶體。

5.1 實驗及分析

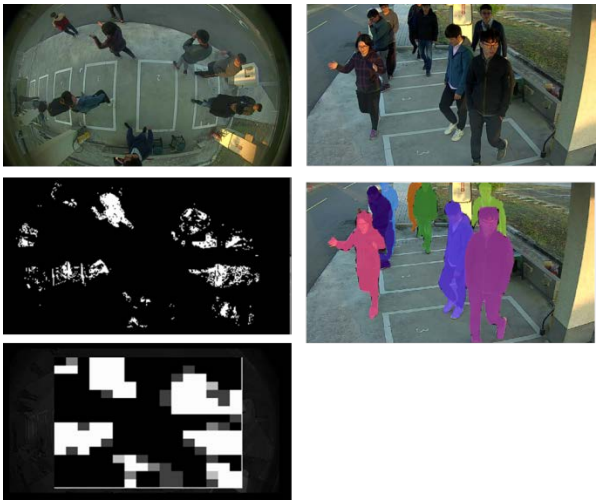
前台行人流量分級0~3人列為稀疏與3人以上列為擁擠，我們測試影片包含：1~10位行人隨機移動進出目標範圍計3321張；行人1~5人戴帽、背包等特殊物品隨機移動進出目標範圍計834張，共4233張影像，影片中包含稀疏與擁擠之狀態，前台偵測結果如圖十二(a)，影像統計正確計算行人流量比率之正確率為94.28%(正確人流分級張數/總張數)，如表一所示；後台深度學習偵測結果如圖十二(b)，分析後台系統人數計數之人數偵測率為94.04%(正確人數/實際總人數)，誤判率為2.26%(誤判人數/實際總人數)。

表一：前台正確率

	總數	正確數	正確率
隨機1~10人行走	3321	3157	95.06%
戴帽	427	387	90.63%
背包	485	447	92.16%
整體分析	4233	3991	94.28%

表二：後台偵測率及誤判率

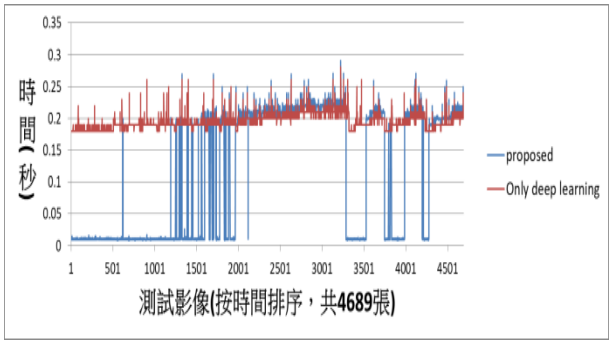
實際總人數	16884		
正確人數	15879	偵測率	94.04%
誤判人數	383	誤判率	2.26%



(a)前台偵測結果　　(b)後台偵測結果

圖十二

時間花費逐一影像觀察如圖十三，單就每個影像執行Mask R-CNN時，每張影像所花時間平均約為0.19秒(紅線)；而利用本論文的方式可以看出執行前台所花時間約0.02秒，當驅使Mask R-CNN的條件發生後，前台才喚醒後台進行其運算工作，所花費時間相加約0.21秒(藍線)，因此系統將會根據觀察到的人流狀態進行分工，以達到節省時間與資源的效果。



圖十三：系統時間比較

6. 結論

本論文提出的系統架立在前後台兩階段的架構上，前台以兩角度影像進行人流分級，分成兩級，當前台人流分級偵測發現為擁擠階段才會執行後台系統，並以Mask R-CNN進行人數計數，經過實驗數據統計後，整體而言前台人流分級正確率為94.28%，後台人數計數正確率為94.04%。本論文方法與單獨以Deep learning偵測方法比較數據可知，系統可以根據前後台分工來有效降低整體的執行時間。

由於候車亭具有離峰時段與尖峰時段，本論文方法因此有明顯人數變化上可達到有效降低執行時間的效果。未來智慧運輸的發展可以朝多前台系統結合一後台系統方向，並將前台系統整合至嵌入式平台，可有效降低成本。

7. 參考文獻

1. Jingwen Li, Lei Huang and Changping Liu, "An efficient self-learning people counting system", The First Asian Conference on Pattern Recognition, pp. 125–129 (2011).
2. Xi Zhao, Emmanuel Delleandrea and Liming Chen, "A People Counting System Based on Face Detection and Tracking in a Video", 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 67-72 (2009).
3. Elvira Sukma Wahyuni, Rizqi Renafasih Alinra and Hendra Setiawan, "People counting for indoor monitoring", 2017 International Conference on Computing, Engineering, and Design (ICCED), pp. 1-5 (2017).
4. Fabio Dittrich, Luiz E. S. de Oliveira, Alceu S. Britto Jr. and Alessandro L. Koerich, "People Counting in Crowded and Outdoor Scenes using a Hybrid Multi-Camera Approach", Cornell University Computer Science, (2017).
5. Peiming Ren, Wei Fang and Soufiene Djahel, "A novel YOLO-Based real-time people counting approach", 2017 International Smart Cities Conference (ISC2), pp. 1-2 (2017)
6. Misbah Ahmad, Imran Ahmed, Kaleem Ullah and Maaz Ahmad,"A Deep Neural Network Approach for Top View People Detection and Counting", 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, pp. 1082-1088 (2019)
7. Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, " You Only Look Once: Unified, Real-Time Object Detection", IEEE conf. on Computer Vision and Pattern Recognition, pp.779-788 (2016).
8. Ross Girshick, "Fast R-CNN", IEEE Intl. Conf. on Computer Vision, pp. 1440–1448, 2015.
9. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, "Mask R-CNN", International Conference on Computer Vision, pp. 2961-2969 (2017).
10. Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction", Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.
11. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 886–893 (2005).