Recognizing Individuals in Spaces by Fusing Computer Vision and Inertial Sensing Information

Chia-Ling Tsai [1], Wei-Chiao Chang [1] , Min-Chun Yeh [1,*] , Zhi-Cheng Liu [1] , Jen-Jee Chen [2], Simon C. Li [1]

[1] Department of Electrical Engineering, National University of Tainan, Tainan 700301, Taiwan
[2] College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan 71150, Taiwan

[*] E-mail : blue20120825@gmail.com

Abstract

Person identification is always one of the most popular technology applications. There are many devices and products have been sold to do person identification, such as radio frequency identification (RFID), face recognition, and iris recognition. However, most of identifications approaches, which are all based on single technology, have limitations when applying in the real environment. For example, they are strongly restricted by specific scenarios and spatial condition of places.

In this paper, we propose a data fusion method which combines three kinds of sensors, a camera, inertial sensors and compasses. The camera can capture the video of the whole space, with the video and AI algorithms, the record objects' positions and trajectories can be calculated and identified. Each user is equipped with a wearable device, and the wearable device can capture the user's motion without any space constraints. The video is not used for face or iris recognition so video quality is not concern here and privacy violation problem is prevented. In this paper, we propose a feature fusion algorithm, which not only considers the motion trajectory of the subject, but also the time characteristics. By the proposed methods, user and wearable devices are paired, so each user can be identified via his or her wearable device, which owns a unique id. According to experiments, our system reaches over 95% recognition rate. A prototype implementation is completed demonstrated to verify the feasibility of our proposed approach.

Keywords: Person identification, computer vision, data fusion, inertial sensors, wearable devices

---

* Corresponding author: blue20120825@gmail.com
DOI：10.53106/222344892023041301003

融合電腦視覺與慣性感測資訊實現在空間中識別身分

蔡佳玲[1], 張惟喬[1], 葉旻純[1,*], 劉致誠[1]
陳建志[2], 李世明[1]

國立臺南大學電機工程學系[1]
國立陽明交通大學智慧科學暨綠能學院[2]
*E-mail : blue20120825@gmail.com

摘要

身份辨識其應用場景非常廣泛，如：互動式機器人，目前市面上有許多相關的設備與產品可協助執行身分辨識，例如：無線射頻辨識(RFID)、虹膜辨識。但這些辨識大部分都是使用單一的設備，因此在應用在現實環境時，總會遇到諸多限制；例如：虹膜辨識和指紋辨識需要短距離或接觸操作。

在本論文中，我們運用三種感測器進行資料融合，分別為攝影機、慣性傳感器和電子證件。雖然使用 RGB 視覺攝影機協助取得場景的視覺資料，但是本論文並沒有使用到人臉辨識的技術，這樣的系統一來有助於解決人臉辨識所造成的隱私問題，二來不要求所使用的攝影機必須具備高解析度且不需要預先替場景中的人員建立其人臉生物資訊，最後是融合多感測器可校正慣性感測器的定位誤差問題且可有效降低影像遮蔽所造成的負面效應。在本文的系統中，我們提出二個特徵融合算法，進行感測資料融合，同時此演算法除了考慮受試者的運動軌跡，更加入受試者運動過程的時間特徵；算法的執行不需繁瑣的數據標籤和模型訓練。實驗數據顯示，我們的系統具有高達 95%以上的辨識率。我們實現並實現一原型系統來驗證我們的方法與可行性。

關鍵詞：身分辨識、電腦視覺化、資料融合、慣性感測器、穿戴式裝置

1. 緒論

身份辨識能協助提供許多個人化的服務。透過得知人的身分 ID，互動式的設備便可以提供客製化的服務內容以及提供友善的用戶體驗[1]。因此，身分辨識甚至可以成為 HRI 領域的感知度量[2]。不僅可以改善人與機器人之間的互動關係，更可以將此功能實現於工廠的設備控制或操作員的識別。

為了識別人們的身份，最直觀的方法是使用人臉辨識。。但是，這些基於攝影機的識別方法會涉及一些隱私、遮蔽和解析度不高之問題，因此並不適合動態環境。除此之外，人臉辨識需要很大量的標記數據集以訓練分類模型。手動註冊用戶更是非常耗時而且系統的可擴展性弱[4,5]。 雖然還有其他種類的人員識別系統，如指紋辨識和虹膜辨識。

在本論文中，我們針對動態環境提出整合電腦視覺與電子證件及慣性感測資訊的身份辨識。雖然目前市面上已有幾個基於此，針對機器人或 HRI 領域開發的應用程序，例如：機器人透過人臉辨識為客戶提供個性化服務，並提供類似人類的互動[7]。目前已有許多研究提出處理人物識別的混合系統。研究[6,7]介紹了一種將電腦視覺與射頻技術相結合的混合系統，這些作品利用深度相機處理空間信息和接收到的 RFID 標籤信號強度來進行配對。

在本文中，我們提出了數據融合的方法，整合攝影機、慣性傳感器和電子羅盤，個別提供資訊進行空間中人物身分識別。圖 1 顯示應用情境，一群人在空間中進行活動，空間中架設攝影機回傳即時影像，每個人都配戴電子證件或穿戴式設備，該證件或設備已整合個人註冊的唯一 ID，以及慣性感測器及電子羅盤。

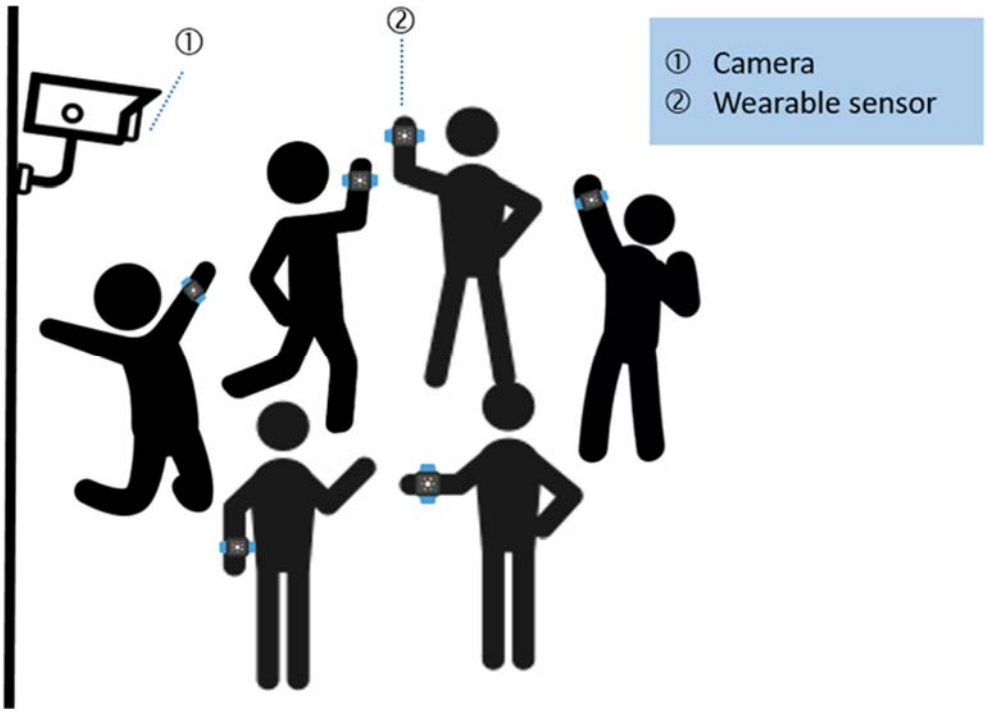

圖 1-1 辨識場景

我們所提供的系統可以提取影像中的人物及位置特徵，和穿戴裝置的活動及身分特徵，接著融合這些數據及特徵跟 ID，匹配影像中的人物。為了進行影像中人物與 ID 的

匹配，我們提出兩種融合算法，分別是 3D-DTW 算法和 state 算法。3D-DTW 方法擴充原有 DTW(Dynamic Time Warping軌跡比對方法，加入軌跡點的時間特徵，同時考慮影像與感測器軌跡間時間的相關性與空間中移動軌跡的相關性，分別是 3D-DTW 算法和 state 算法。3D-DTW 方法擴充原有 DTW(Dynamic Time Warping軌跡比對方法，加入軌跡點的時間特徵，同時考慮影像與感測器軌跡間時間的相關性與空間中移動軌跡的相關性，而 DTW and State 算法則是採取兩階段匹配的策略簡化複雜度，第一個階段先使用影像及感測器於空間上的軌跡進行 DTW 相似性配對，若有疑義則進入第二階段進行時間上的 state 特徵相似性比對。匹配過程中，量化相關性以確定軌跡間相似性，再根據相似性與其分佈每個組合進行評分，使用我們提出的配對算法，將人物影像與穿戴傳感器及 ID 進行一對一配對。

總結來說，本論文具有以下貢獻：

•提出了一個完整的架構，用數據融合方法結合不同感測資料提供的特徵解決人員識別問題。

•提出二種數據融合算法，同時考慮影像與感測器的時間空間特徵的軌跡數據之相似性。

•提出的配對算法不僅考慮各種配對的相似性得分，還應用統計指標提高配對準確率。

•本論文所提的系統可用於識別影像中人員的身份，但無需任何影像數據標籤和模型訓練，避免侵犯隱私權之疑慮。

接下來的論文中，在第二節，我們回顧了同樣融合視覺和慣性感測數據系統的一些論文研究。第三節展示並說明系統框架且介紹使用到的軟硬體的設備與架構。第四節說明本文提出的方法及其步驟以及配對算法。第五節則針對提出的方法進行績效評估並與其他方法做比較。最後，第六節總結本論文並描述未來的工作。

2. 參考文獻

電腦視覺與穿戴式裝置技術已廣泛應用於動作辨識[9,10]。甚至進一步可以實現許多應用並增強人機界面的體驗互動（HCI）[11]。然而，每種傳感器在現實條件下都有自身的局限性及不足之處。因此，便整合了視覺傳感器和慣性傳感器，以達到更穩健且優異的表現。

身分識別技術與許多應用息息相關，故長期以來一直受到重視，如相關的生物識別技術，例如虹膜辨識[15]、指紋辨識和人臉辨識，這些均是利用其獨特性來進行識別，

但是，虹膜辨識和指紋辨識需要短距離或接觸操作，因此，這兩種技術對於動態場景是不可行的；雖然人臉辨識可以支持更廣泛的操作範圍，但是，其影像資料需高解析度，且受到大數據的要求限制與訓練辨識模型相當費時，這些限制使得臉部辨識並不適合公共場合人數多的情況。

為了辨識人群和動態環境中的人，有幾項研究提出使用混合式系統，整合兩種以上的感測器資料。參考文獻[7]為了實現類似的互動目的，試圖增強人機交互的體驗，它利用每個人配戴的 RFID 標籤及訊號計算其相對運動路徑以及用 3D 深度相機偵測到的影像運動軌跡，接著使用支持向量機（SVM）分類器來進行匹配物理運動路徑，但此實驗環境中只有單個 RFID 接收器而無法考慮移動的方向，且 RFID 訊號的漂移及雜訊問題嚴重。

3. 系統架構與問題定義

本論文假設空間中安裝攝影機監看環境，並利用攝影機紀錄用戶移動軌跡，多數在空間中行動與進出，使用者皆配戴了整合電子 ID 功能、慣性感測器、羅盤功能的穿戴式裝置，蒐集這些穿戴式裝置的數據集，經過適當的訊號處理或前置處理後，分析可得各裝置的行進方向與步伐數據等移動軌跡，接著可透過影像移動軌跡與穿戴式裝置的移動步伐數據進行相似度計算，根據評分，在使用我們設計的配對算法進行穿戴式裝置與影像配對。該算法以一對一的方式將影像軌跡與穿戴傳感器配對，辨識出監看影像中每個使用者的身分。

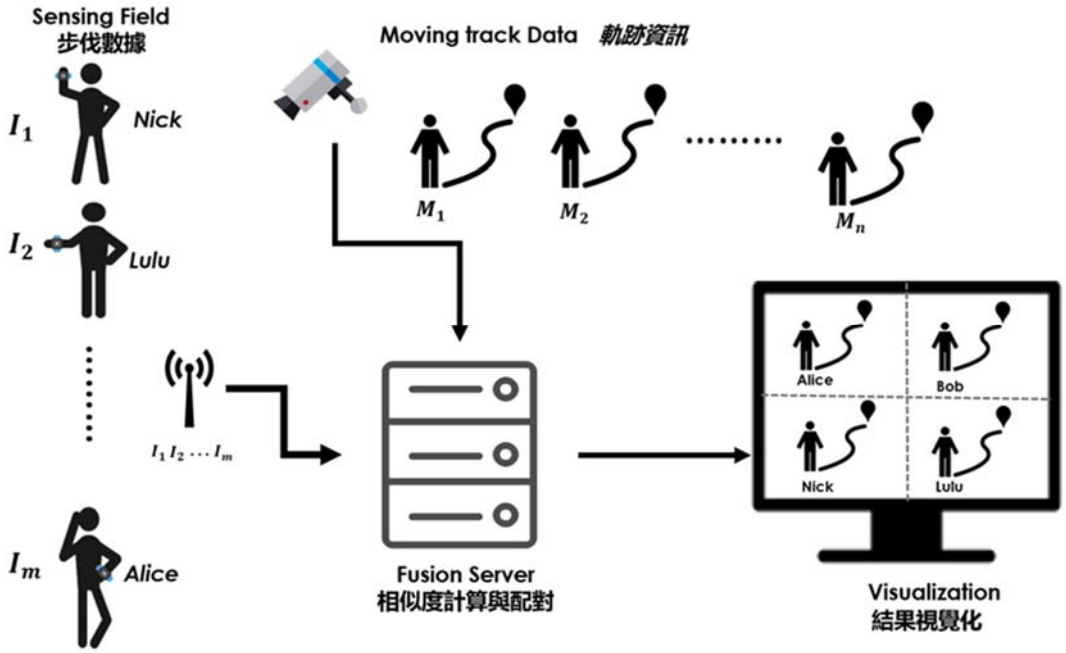

圖 2-1 系統架構圖

我們的人員辨識問題定義如下，一群人在架有攝影機的場景中活動及進出，攝影機捕獲這些人的移動軌跡數據使用$M_1, M_2, \dots M_n$表示，其中 N 是場景中的人數，每個用戶都佩戴穿戴式裝置，該裝置整合電子 ID、慣性感測器與羅盤等，其收集的數據表示$I_1, I_2, \dots I_m$，其中 M 是捕捉到的穿戴式裝置的數目。這些數據上傳到融合伺服器後，必須經過適當的訊號處理與前置處理，以過濾雜訊並轉化為有用的特徵I_m^P與M_n^P，這裡我們使用的特徵是步伐特徵I_m^P與影像軌跡M_n^P，接著 fusion sever 需要對I_m^P，$m = 1 \dots M$，與

M_n^P，$n = 1 \ldots N$，進行融合，此處我們必須設計方法計算I_m^P和M_n^P的關聯性，如果I_m^P和M_n^P之間具高相關性且沒有其他更好的配對，則可以產生一對配對（I_m, M_n），並從儲存在I_m中的用戶檔案中推斷出影像M_n的 ID。 最後，我們即可以在影像中標記其身分的 ID，本論文必須完成的部分包括感測器與影像資料的訊號處理，前處理，特徵轉換及融合演算法等。

為了記錄慣性感測數據以取得步伐及方向特徵，我們利用樹莓派 3 搭配 MPU6050 及 HMC5883L 作為穿戴式設備之雛形。並撰寫一個應用程序於樹莓派來記錄除重力之外的 3 軸加速力，藉由磁場和重力值，將加速度從設備坐標系旋轉到地球坐標系，採樣頻率為 10 Hz。影像資料則是使用 Logitech BRIO 4K HD 網路攝影機，透過內置彩色攝影鏡頭進行拍攝，搭配 YOLO v3 進行人物影像辨識，接著紀錄影像軌跡與追蹤，採樣頻率設為 20 FPS（Frames per Second）。

4. 身分辨識演算法

在本章中，我們將身分辨識演算法的 3 個階段作說明，包括

4.1 慣性感測器移動測

藉由測量人的行走步態並在檢測到步伐事件時輸出位移量$\vec{d}$。我們使用的穿戴式裝置是由樹莓派 3、MPU6050 及 HMC5883L 組成。用戶的移動軌跡由一系列 2D 步伐長度和行走方向[17,18,19]建模。 其特徵在於長度S和方向θ。 位移是$\vec{d} = S \cdot [cos\,\theta\,, sin\,\theta]$。如果在一個時間間隔內沒有檢測到運動，則回傳$S = 0$。

◆ 步伐事件檢測

事件的偵測是透過監控慣性感測器其隨時間變化的加速度幅度來完成。加速度定義為式子（4-1），其中 x、y、z 為感測器所讀取到的三軸速度值。

$$a(t) = \sqrt{x(t)^2 + y(t)^2 + z(t)^2} \qquad (4\text{-}1)$$

接著加速度$a(t)$通過移動平均濾波器以去除高頻數據。移動平均濾波器定義如式子（4-2）所示：

$$\hat{a}(t) = \alpha \cdot a(t) + (1 - \alpha) \cdot \hat{a}(t-1) \qquad (4\text{-}2)$$

最後，若$\hat{a}(t)$在窗口 t 內檢測到大於門檻值Δa 的下降時，判定為步伐事件發生，如圖 4-1 所示。若偵測步伐事件發生，則進入下一步驟－步長估算，反之則回傳 $S = 0$。

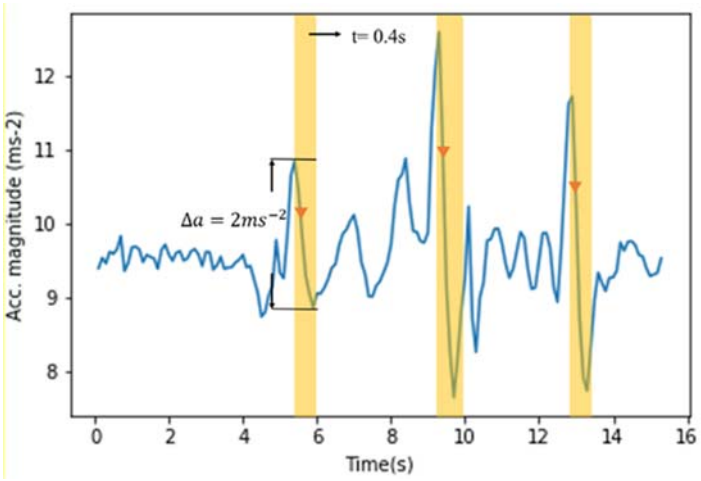

圖 4-1 加速度值大於門檻值Δa的下降

◆ 步長估算

由於用戶的移動軌跡由一系列 2D 步伐長度和行走方向[17,18,19]構成。因此，需要知道長度 S 和方向 θ，此階段目的在於得知此步伐的長度 S。我們採用頻率模型[18]，假設人的步長（S）和頻率（f）之間具有線性關係。如式子（4-3）所示，其中 a，$b \in R$ 是藉由線性回歸發現的常數。

$$S = a \times f + b \qquad (4\text{-}3)$$

為了發現這種線性關係，我們請一位受試者在 100 米的距離，分別進行步行、快速步行和跑步。藉由檢測到的步數得出頻率和長度，接著利用線性回歸來找到 a 和 b 的最佳解。表 4-1 顯示了來自同一位受試者的 12 個訓練集。其回歸關係為$S = 0.877 \times f - 0.931$。最後，請受試者沿著 60 米的矩形路徑以不同的速度行走以測試該模型的準確性。表 4-2 顯示了距離誤差。

表 4-1 頻率模型步長估計的訓練樣本

Trial	Casual walk		Fast walk		Running	
	f (Hz)	S (m)	f (Hz)	S (m)	f (Hz)	S (m)
#1	1.783	0.685	2.000	0.917	2.418	1.250
#2	1.937	0.752	2.094	0.926	2.433	1.205
#3	1.892	0.671	2.028	0.847	2.485	1.235
#4	1.944	0.676	2.089	0.901	2.489	1.205

表 4-2 60 米矩形路徑中頻率模型的距離誤差

	Casual walk	Fast walk	Running
Step frequency f (Hz)	1.88	2.04	2.56
Step length S (m)	0.72	0.88	1.31
Error distance (m)	0.53	1.21	3.20

◆ **步行方向偵測**

HMC5883L 電子羅盤感應器可測量傳感器周圍磁力線的方向，提供三軸個別的磁場強度值，作為方位角的計算基礎。計算方式為讀取X與Y軸的輸出，並套入$tan^{-1}(Y/X)$的公式中，得到方向角。

通過這種方法，我們可以獲得受試者的行走方位。受試者首先將他們的穿戴式設備朝向他們的前進方向，開始行走後，因為可以計算在第一步之前和行走開始之後的方向改變。最後，將步驟事件 i 期間的平均方位角值視為檢測到的方位θ_i。

經過上述這些步驟後，我們定義每位用戶 m 的移動軌跡$I_m(t)$為式子（4-4）：

$$I_m(t) = (S_m \times cos\,\theta_i, S_m \times sin\,\theta_i) + \vec{d}_m \qquad (4\text{-}4)$$

4.2 影像軌跡偵測

藉由攝影機拍攝人們的行走畫面，其輸出定義為$F(t)$，接著透過 YOLO v3 偵測到的 bounding box，bounding box 由 4 個元素組成，分別為 x、y、w、h，其中 x 和 y 為 bounding box 的左下角坐標，而 w 及 h 為 bounding box 的寬和高。每位用戶 n 的移動軌跡透過紀錄 bounding box 的中間底部而成，記為$\left(x_n(t) + \frac{w_n(t)}{2},\ y_n(t)\right)$。

◆ **影像做邊轉換**

若在現實空間中定義座標，如圖 4-2，接著透過攝影機拍攝人站在這些座標上的其 bounding box 中間底部$\left(x(t) + \frac{w(t)}{2},\ y(t)\right)$，再將其繪製出來，如圖 4-3。

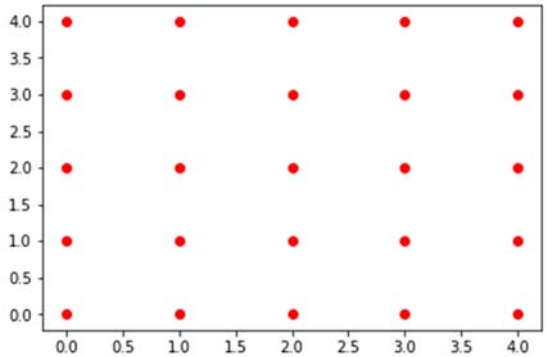

圖 4-2 現實座標

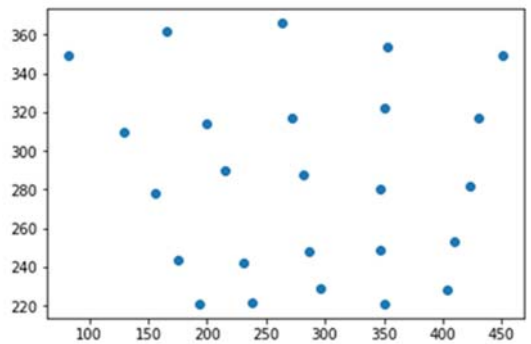

圖 4-3 影像座標

從圖 4-2 和圖 4-3，我們假設這兩個座標存在某種線性關係。因此我們利用多項式曲線擬合，同時考慮損失函數RMS$E(p)$與判定係數 R2 (Coefficients of Determination)，目的在於找出最符合的方程式，以準確地預測軌跡。多項式曲線擬合之計算公式，定義為式子（4-5）：

$$P_M(x,y) = \left[\sum_{j=0}^{M} x^j, \sum_{i=0}^{K} y^i, 1\right] \qquad (4\text{-}5)$$

其中$P_M(x,y)$為多項式特徵集，可以理解為對現有特徵的乘積，M 和 K 為多項式的最高次數，x^j代表 x 的 j 次冪，y^i代表 y 的 i 次冪。

我們用平方誤差和（sum of the squares of the errors）作為損失函數$RMSE(p)$，其樣本的數目為 N，對於每一個樣本$S_n = (x_n, y_n)$，其對應的真值(Ground truth)為t_n，損失函數RMS$E(p)$可表示為式子（4-6）：

$$RMSE(p) = \sum_{n=1}^{N}\{S_n - t_n\}^2 \qquad (4\text{-}6)$$

判定係數 R2 (Coefficients of Determination)用以來判斷依變數 Y 與獨立變數 X 線性相關的強度，亦稱為迴歸直線的配適度（goodness of fit），或稱為迴歸直線的解釋能力。其值介於 0 至 1 之間，值愈高代表 SSR 的值愈接近 Y 的總變異，表示樣本資料點的$S_n = (x_n, y_n)$值在迴歸直線附近變動的情況不大，我們稱之為迴歸模型的配適情況良好。

SSR稱為迴歸模式的變異量，SSR計算公式如式子（4-8），其中$\widehat{P_M}$為迴歸模式在P_M點的預測值，$\bar{P}$為所有P_M值的平均值。SSE 為誤差變異量或稱為誤差平方和，計算公式如式子（4-9）。SST稱為P_M值之變異量，計算公式如式子（4-10）。

$$R2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (4\text{-}7)$$

$$SSR = \sum_{i=1}^{n}\left(\widehat{P_M} - \bar{P}\right)^2 \qquad (4\text{-}8)$$

$$SSE = \sum_{i=1}^{n}\left(P_M - \widehat{P_M}\right)^2 \qquad (4\text{-}9)$$

$$SST = \sum_{i=1}^{n}(P_M - \bar{P})^2 \qquad (4\text{-}10)$$

從圖 4-4 可知，在 degree =3 時，損失函數RMS$E(p)$ 其值最小，判定係數 $R2$ 最接近於 1，由此可知用三階線性方程式其預測的座標會最接近我們在現實定義的座標。

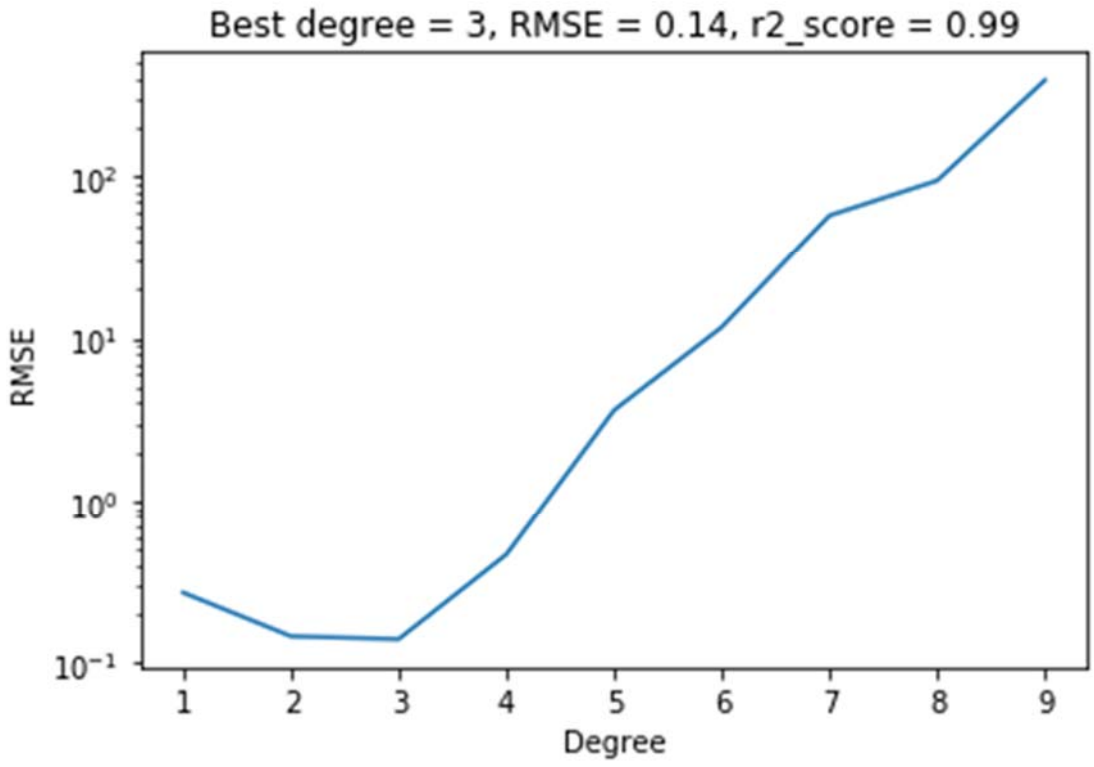

圖 4-4 各階線性方程式RMSE值與 R2

為了更清楚展示各階線性方程式的預測結果，圖 4-5 至圖 4-8 為各階線性方程式的預測座標與我們定義之現實座標的擬和狀況。也可更加確定三階線性方程式其預測的座標確實會最接近我們在現實中定義的座標。

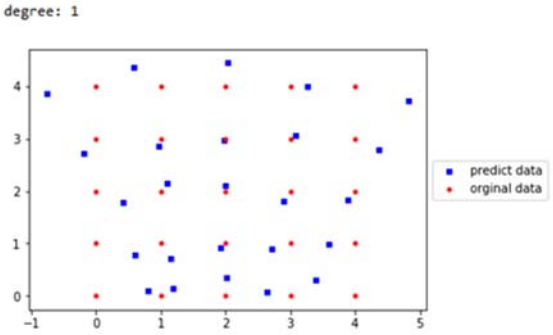

圖 4-5 一階線性方程式擬合

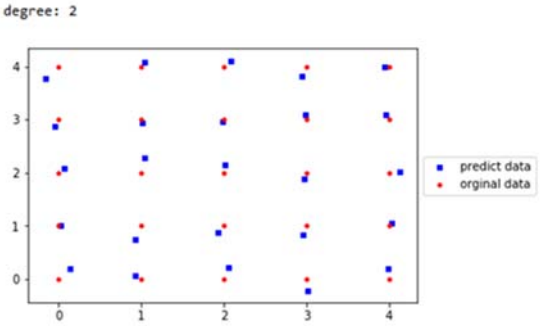

圖 4-6 二階線性方程式擬合

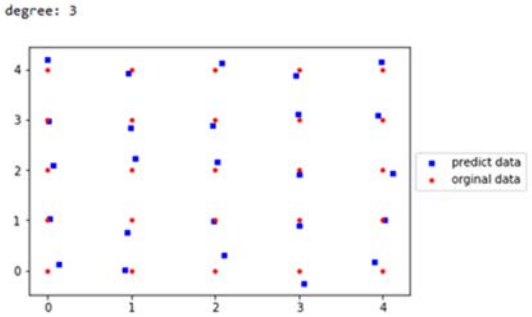

圖 4-7 三階線性方程式擬合

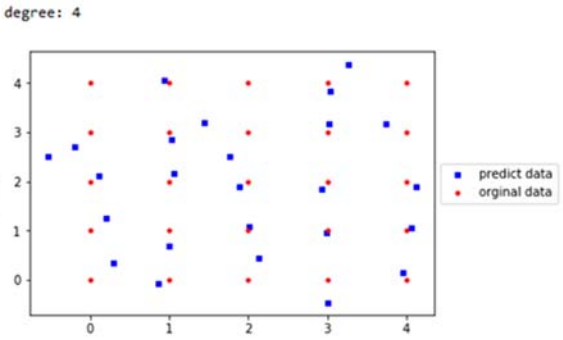

圖 4-8 四階線性方程式擬合

◆ **數據預處理**

在確定擬合度最好的方程式後，我們便可藉此得知目前攝影機畫面中每位用戶 n 的現實位置$M_n(t)$。但由於 YOLO v3 所偵測到的 bounding box 會因受試者走動而有不穩定大小與位置變化，而導致的原始影像坐標值會產生抖動與雜訊，如圖 4-9 所示。因此，我們使用移動平均法減少坐標值序列中的抖動[22]，定義為式子（4-11）。l為移動平均的個數，此算式中l =5。

$$M_n(t)=\left(\frac{x_n(t-1)+x_n(t-2)+x_n(t-3)+\cdots+x_n(t-l)}{l},\frac{y_n(t-1)+y_n(t-2)+y_n(t-3)+\cdots+y_n(t-l)}{l}\right) \quad (4\text{-}11)$$

圖 4-10 為經過移動平均後繪製出的結果，跟圖 4-9 相比，可看出線段平滑許多。

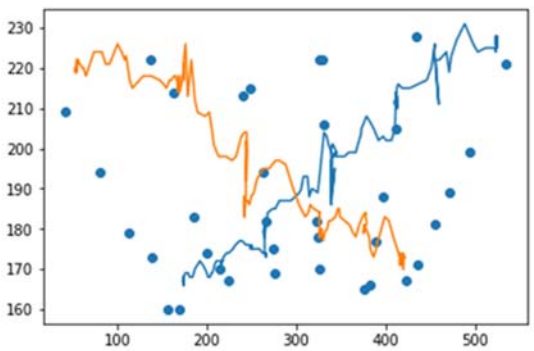

圖 4-9 原始坐標值

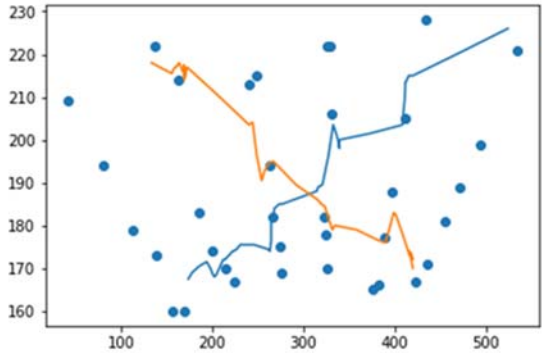

圖 4-10 移動平均後的坐標值

4.3 相似度分數

一般而言，兩個 n 維空間中的向量 x 和 y，它們之間的距離可以定義為兩點之間的直線距離，稱為歐基里得距離（Euclidean Distance），如式子 4-12 所示：

$$dist(x,y) = \|x-y\| = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots(x_n-y_n)^2} \quad (4\text{-}12)$$

但是如果向量的長度不同，那它們之間的距離，就無法使用上述的式子來計算。一般而言，假設這兩個向量的元素位置隨時間改變，由於必須容忍在時間軸的偏差，而且並不知道兩個向量的元素對應關係，因此我們必須靠著一套有效的運算方法，才可以找到最佳的對應關係。

DTW 是 Dynamic Time Warping 的簡稱，中文可以翻譯成「動態時間扭曲」或是「動態時間校正」，這是一套根基於「動態規劃」（Dynamic Programming，簡稱 DP）的方法，可以有效地將搜尋比對的時間大幅降低。透過動態扭曲兩者來計算序列的距離，距離越短則相似度則越高。

舉例來說，若要計算相似度的兩個時間序列 X 和 Y，長度分別為 $|X|$ 和 $|Y|$。DTW 的目的在於使歸整路徑(Warp Path)W 最小化，其表示為 $W = w_1, w_2, \ldots, w_k$，其中 $Max(|X|,|Y|) <= K <= |X|+|Y|$。$w_k$ 的形式為 (i,j)，其中 i 表示的是 X 中的 i 座標，j 表示的是 Y 中的 j 座標。歸整路徑 W 必須從 $w_1=(1,1)$ 開始，到 $w_k=(|X|,|Y|)$ 結尾，以保證 X 和 Y 中的每個座標都在 W 中出現。另外，W 中 $w(i,j)$ 的 i 和 j 必須是單調增加的，以保證圖 4-11 中的虛線不會相交，所謂單調增加如式子（4-13）所示：

$$w(i,j) = (i,j), w_{k+1}(i',j')$$
$$i \le i' \le i+1, j \le j' \le j+1 \quad (4\text{-}13)$$

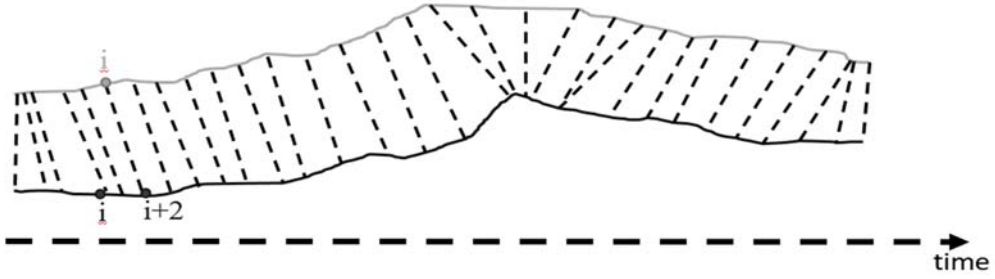

圖 4-11 兩個時間序列之間的動態扭曲

最後要得到的歸整路徑必須是一個距離最短的歸整路徑，如式子（4-14）所示：

$$DTW(i,j) = dist(i,j) + min[DTW(i-1,j), DTW(i,j-1), DTW(i-1,j-1)] \quad (4\text{-}14)$$

然而，由於軌跡資訊 $M_n(t)$ 和步伐數據 $I_m(t)$ 皆已經是由為 2 維座標構成，因此在上述 DTW 算法中造成時間這項特徵無法被納入計算，因此，我們進行改良，分別為 3D-DTW 演算法和 DTW＆State 演算法。

◆ 3D-DTW 演算法

3D-DTW 方法擴充原有 DTW (Dynamic Time Warping)軌跡比對方法，加入軌跡點的時間特徵，同時考慮影像與感測器軌跡間時間的相關性與空間中移動軌跡的相關性。此時需要進行 3 維的距離計算，在三維空間裡的兩個點之距離應為式子（4-15）：

$$3ddist(x_{12}, y_{12}, z_{12}) = \sqrt{(x_1 - x_2)^2 + (y - y_2)^2 + (z_1 - z_2)^2} \quad (4\text{-}15)$$

由於想加入軌跡點的時間特徵，因此我們將 z 軸的值以秒數替代。Timestamp 定義為自 1970 年 1 月 1 日（00:00:00 GMT）以來的秒數，我們藉由 Timestamp 將影像軌跡資料與感測器步伐資料上的日期時間全部換算為秒數，其定義為 *TS(t)*。接著將原本的軌跡資訊$M_n(t)$改寫為式子（4-16），步伐數據$I_m(t)$改寫為式子（4-17）：

$$M_n(t) = \left(\frac{x_n(t-1)+x_n(t-2)+x_n(t-3)+\cdots+x_n(t-l)}{l}, \frac{y_n(t-1)+y_n(t-2)+y_n(t-3)+\cdots+y_n(t-l)}{l}, TS_n(t)\right) \quad (4\text{-}16)$$

$$I_m(t) = (S_m \times cos\,\theta_i, S_m \times sin\,\theta_i, TS_m(t)) + \vec{d}_m \quad (4\text{-}17)$$

接著，我們將相似度得分定義為式子（4-18）：

$$D\left(M_n(t), I_m(t)\right) = 3ddist\left(M_n(t), I_m(t)\right) + min\left[DTW\left(M_n(t\text{-}1), I_m(t)\right), DTW\left(M_n(t), I_m(t\text{-}1)\right), DTW\left(M_n(t\text{-}1), I_m(t\text{-}1)\right)\right] \quad (4\text{-}18)$$

較大的 D 值意味著此影像與感測器軌跡之間距離更長且相似度較低。但由於 Timestamp 換算出來的值非常大，從圖 4-12.可知，因而導致座標與時間的權重不相等，因此透過標準化的程序將其值限定在一個範圍，來調整時間和座標的權重，使其相等，如圖 4-13 所示。由圖 4-14 可知其標準化對整體準確度的影響。

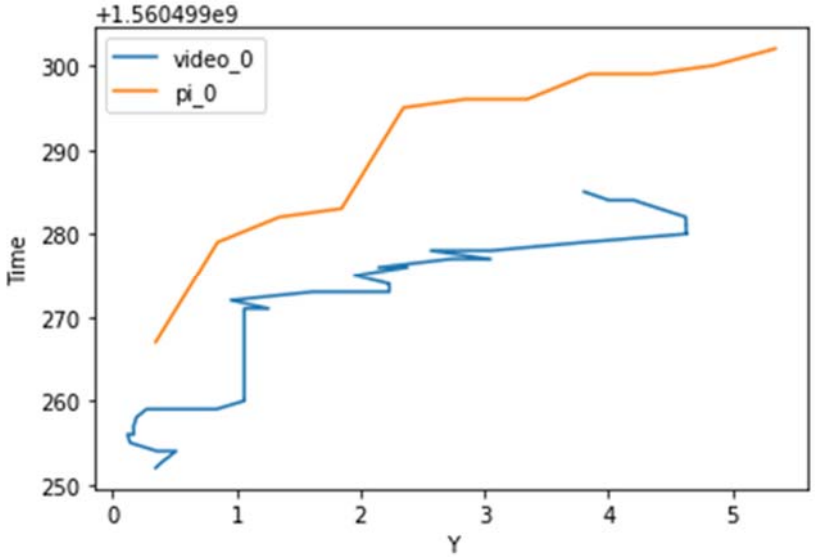

圖 4-12 標準化前座標與時間的可視化

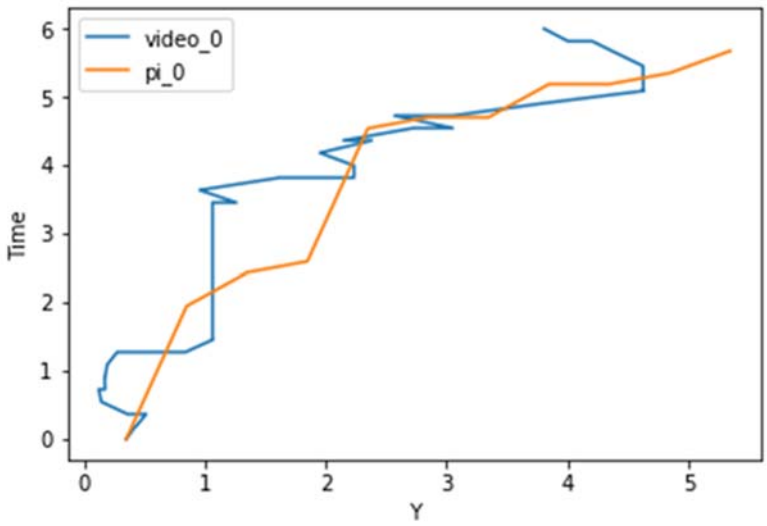

圖 4-13 標準化後座標與時間的可視化

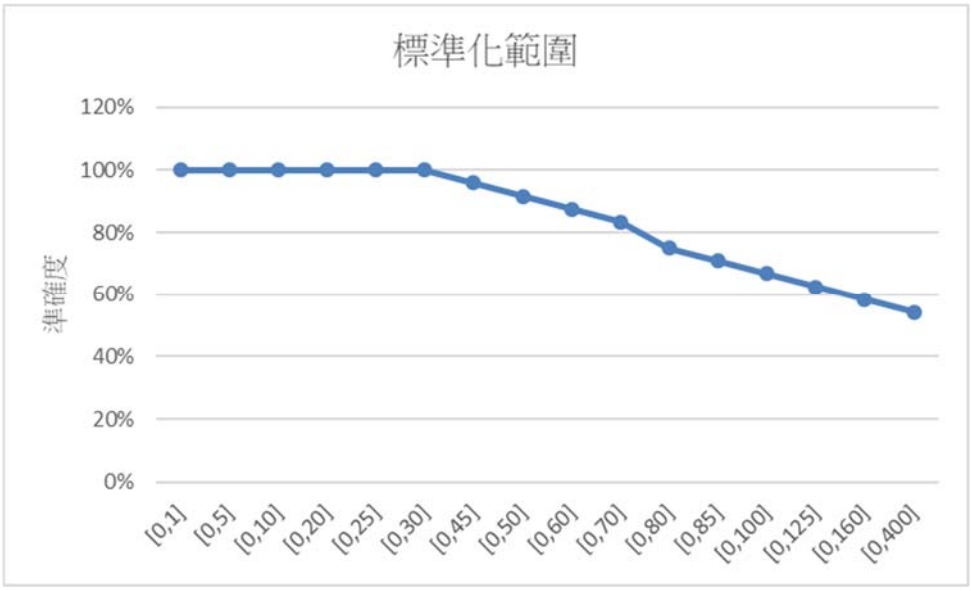

圖 4-14 標準化範圍與準確度

◆ DTW and State 演算法

DTW and State 算法採取兩階段匹配的策略簡化複雜度，第一個階段先使用影像及感測器於空間上的軌跡進行 DTW 相似性配對，若有疑義則進入第二階段進行時間的相似性比對。

第一階段

採用 DTW 算法，當軌跡資訊$M_n(t)$和步伐數據$I_m(t)$在路線上具有較高的相似度時，保留其分數，接著進入一對一配對。當其相似度在路線上無明顯區別時，則進入第二階段。

第二階段

當軌跡資訊$M_n(t)$和步伐數據$I_m(t)$在路線上無法判斷其相似度時，將兩條序列的時

間資料單獨取出，進行更進一步的分析和比較。首先若軌跡資訊$M_n(t)$其座標資訊發現用戶於一定範圍內停留超過一定時間即判定為停留，接著將其停留事件的開始時間與結束時間取出，將這段時間的狀態標註為"Stay State"。其他時間則標註為"Move State"，如圖 4-15.所示。

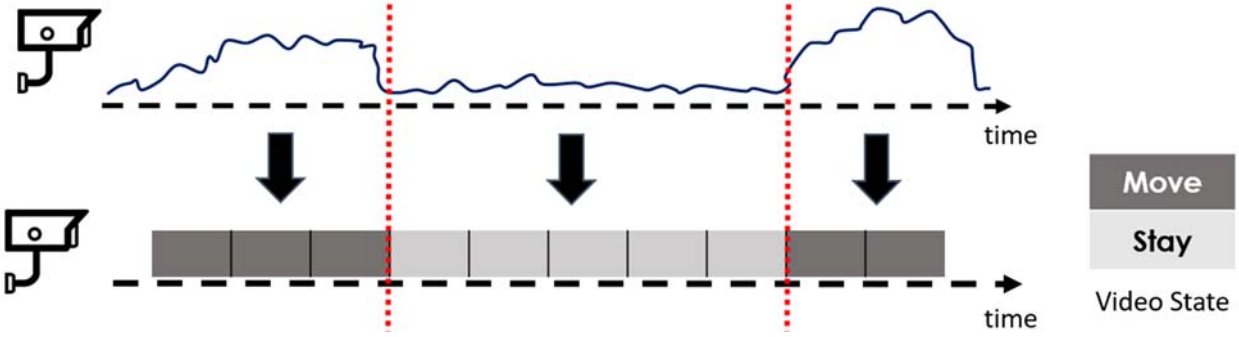

圖 4-15 影像狀態轉換對照圖

接著為了進一步比對軌跡資訊$M_n(t)$和步伐數據$I_m(t)$ 的狀態"State"，以$I_m(t)$為例。我們先標記每個步伐事件的開始和結束時間。由於移動是由一系列連續步伐組成的事件。因此，如果兩個步伐被一個小於門檻值ΔG 的時間間隔給隔開，我們會將兩個步伐合併為一個"Move State"。若兩個步伐被一個大於門檻值ΔG 的時間間隔給隔開，將兩個步伐合併為一個"Stay State"。該概念如圖 4-16.所示。最後將軌跡資訊$M_n(t)$和步伐數據$I_m(t)$的"State"用圖 4-17.呈現。

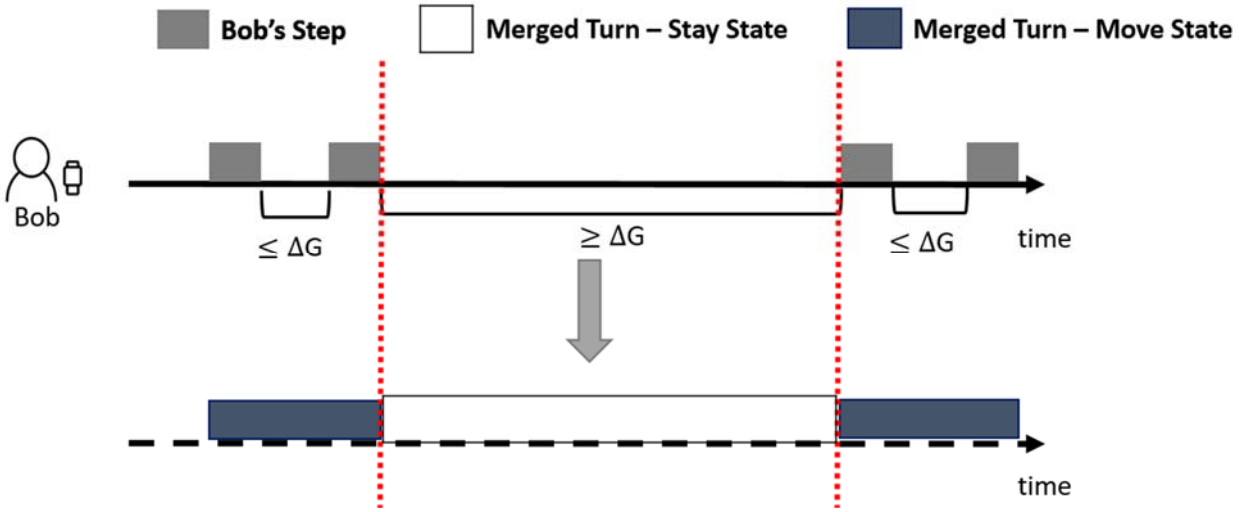

圖 4-16 步伐間的合併

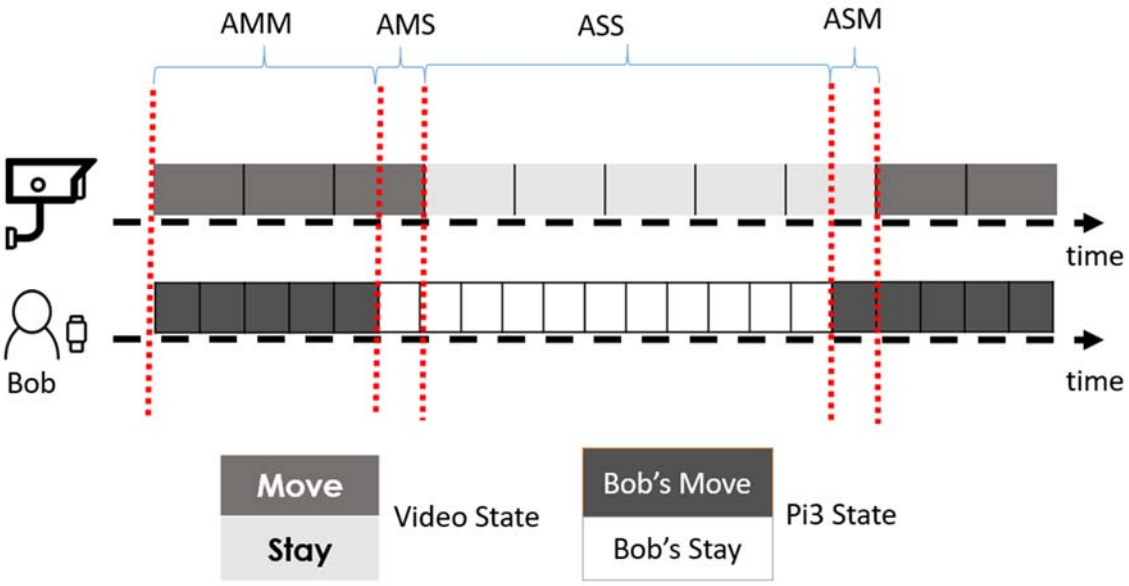

圖 4-17 步伐的停留時間與影像的停留時間比對圖

從圖 4-17.中，顯示在不同情況下這些重疊的例子，可從中識別出 4 種重疊。，如表 4-3 所示。這些重疊可以反映$I_m(t)$和$M_n(t)$的時間相關性。在這 4 種重疊中，AMM 和 ASS 定義為正相關重疊（PO）。相比之下，AMS 和 ASM 被認為是負相關重疊（NO）。

表 4-3 重疊的四種類型

重疊狀態	重疊狀態描述
AMM	$I_m(t)$和$M_n(t)$重疊部分皆為"Move"
ASS	$I_m(t)$和$M_n(t)$重疊部分皆為"Stay"
AMS	重疊部分，$M_n(t)$為"Move"而$I_m(t)$為"Stay"
ASM	重疊部分，$M_n(t)$為"Stay"而$I_m(t)$為"Move"

PO 由 AMM 和 ASS 組成，由於其重疊部分皆為相同狀態，因此在分數計算上皆定義為 0，而 NO 由 AMS 或 ASM 組成，則其重疊部分為不同狀態，在分數計算上，定義每隔 AMS 和 ASM 皆定義為 1。

在理想狀況下，步伐之間的停留時間與影像所顯示的停留時間應該兩個時間差差距應最小，其定義為式子（4-19）：

$$Sms\left(\mathrm{M}_n(t),\mathrm{I}_\mathrm{m}(\mathrm{t})\right)=\sum_{s\in PO}0\times(\mathrm{AMM}(|s|)+\mathrm{ASS}(|s|))+\sum_{s\in NO}1\times(\mathrm{AMS}(|s|)+\mathrm{ASM}(|s|)) \quad (4\text{-}19)$$

4.4 影像軌跡和步伐資料配對

經過上一節中的計算，我們可得到兩種演算法的相似度分數，分別為 3D- DTW 的 $D\left(M_n(t),I_m(t)\right)$和 DTW and State 算法的$Sms\left(M_n(t),I_m(t)\right)$為了得到$I_m(t)$和$M_n(t)$其所有組合的相似性得分，我們定義式子（4-20），來區別兩種算法的得分：

$$Sim\langle I_m^p(t),M_n^p(t)\rangle=\begin{cases}D\left(M_n(t),I_m(t)\right),if\ p=1\\Sms\left(M_n(t),I_m(t)\right),if\ p=2\end{cases} \quad (4\text{-}20)$$

在得到$I_m(t)$和$M_n(t)$其所有組合的相似性得分之後，我們可的到一個由相似度分數組成的矩陣，如圖 4-18.所示。

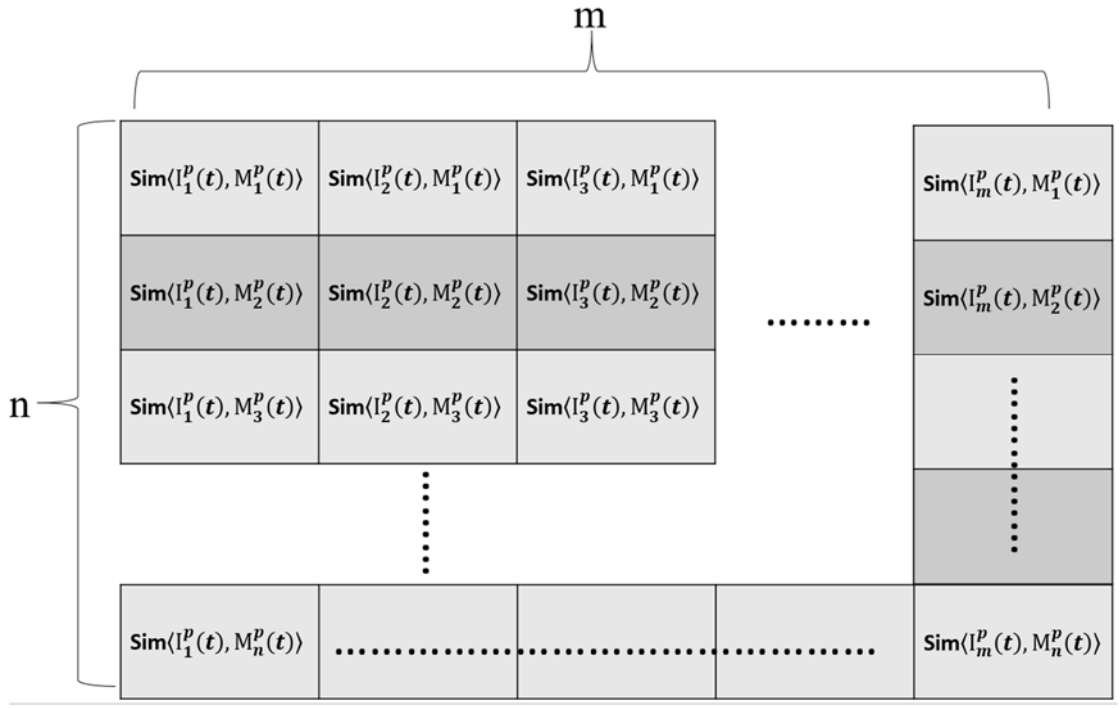

圖 4-18 相似性得分的所有組合

我們需要一種策略來將場景中的所有移動軌跡 M 與所有具有用戶身分 ID 的步伐數據I配對。最簡單的方法便是始終將最高相似度分數的組合配對在一起。然而，此方法有一個致命的缺點，就是當某一軌跡只與某一步伐相似，但因有其他軌跡和這步伐數據相似度分數更高，而造成此軌跡無法配對。因此除了相似度分數外，我們應當考慮其獨特性；也就是說當移動軌跡$M_n(t)$和步伐數據$I_m(t)$之間存在很大差異時，我們應該更優先配對此軌跡。透過這種調整方式，相似性得分已不再是我們唯一的考慮因素。其他組合的獨特性也是必須納入考量的。

基於這個想法，我們使用統計測量標準差（Standard deviation），它能夠量化一組數值的變化量。在我們的問題中，它意味著移動軌跡$M_n(t)$和步伐數據$I_m(t)$的所有組合之間有多大差異，獨特性方程式定義如式子（4-21）：

$$UL(Sim_{I,M_n}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(Sim(I_i, M_n) - \overline{Sim_{I,M_n}}\right)^2} \qquad (4\text{-}21)$$

其中$Sim(I_i, M_n)$是移動軌跡M_n和某個步伐數據I_i之間的相似性得分。$\overline{Sim_{I,M_n}}$是移動軌跡M_n及其與所有步伐數據 I 組合的相似度分數平均值，如圖 4-19.所示。

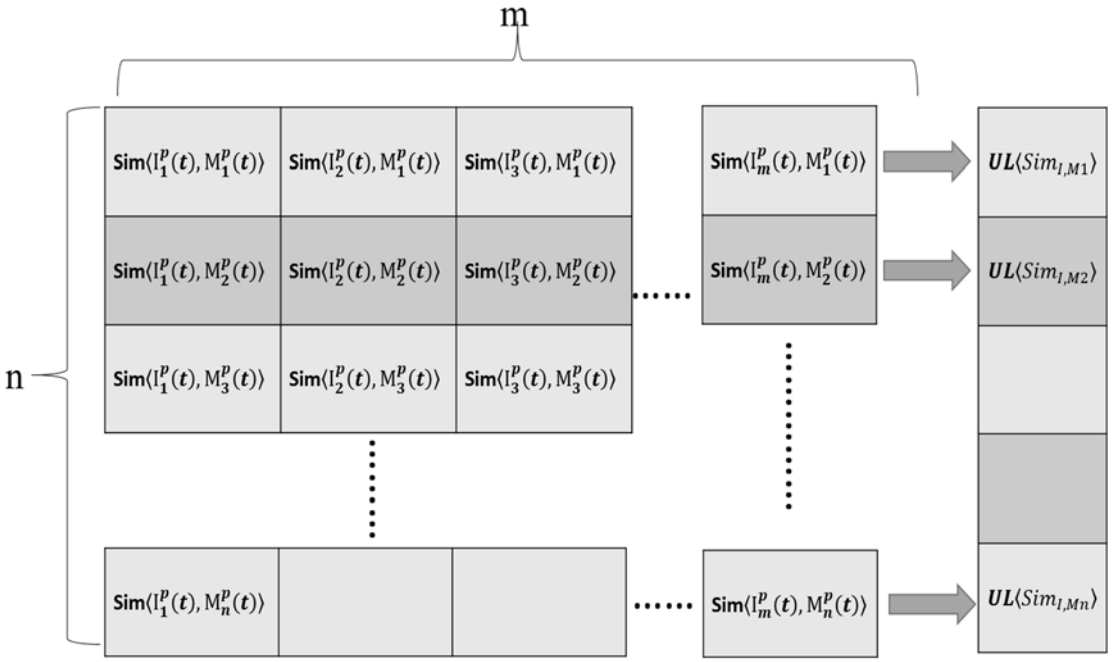

圖 4-19 相似性得分與獨特性

我們首先根據獨特性將軌跡與步伐數據進行配對排序，然後再依據相似性分數進行配對。在經過此配對過程之後，我們的系統便會生成每個移動軌跡和步伐數據的配對結果。藉由這種方式，我們便可以事先在穿戴設備中註冊唯一的身分 ID 來進一步識別攝影機畫面的人。

5. 實驗與評估

本章中，我們評估兩個融合算法在不同方面的性能，也增加了貼近現實的影響因素，例如人數和移動模式。

實驗中，每位受試者皆站在攝影機前並配戴穿戴式裝置，且要求受試者進行隨機的路線行走，透過我們的訪法來將這些組合配對並評估準確性。

5.1 路線種類

為了確認本系統的精準度是否不受路線影響，因此本論文中設計了兩個演算法，並參考了[23]的 Stop And Move 演算法。我們將路線分成同方向直線、不同方向直線與非直線等 3 種，分別進行實驗及比較不同演算法的效能，如圖 5-1.(a)(b)、圖 5-2.(a)(b)、圖 5-3.(a)。最後，將測試兩者的配對表現並接結果繪製在圖 5-4。

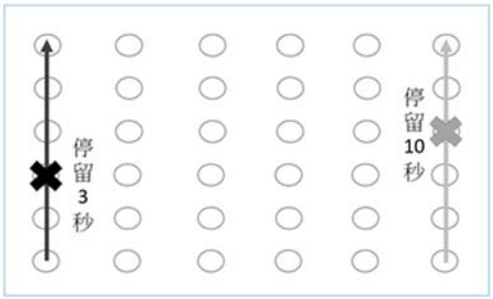

圖 5-1(a)停留時間不同

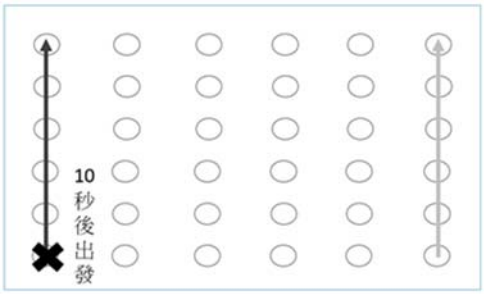

圖 5-1(b)出發時間不同

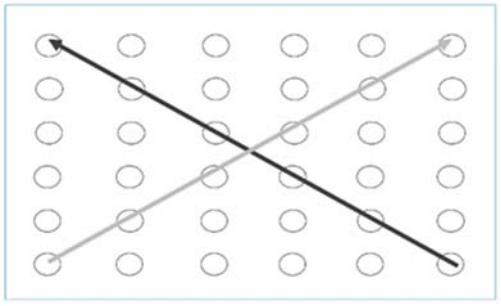

圖 5-2(a)不同方向直線

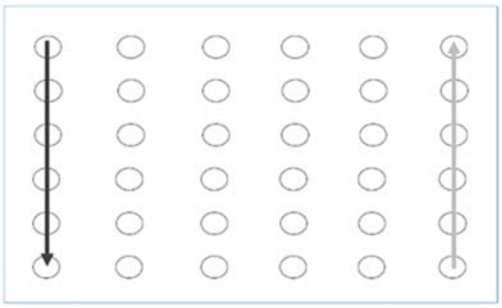

圖 5-2(b)不同方向直線

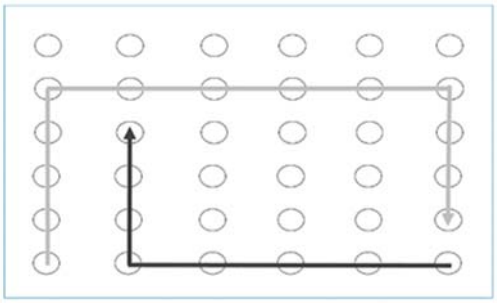

圖 5-3(a)非直線

在同方向直線路線中，兩位受試者行走過程皆具有不同的時間差距，例如：同時出發，但是過程中停留時間不同，如圖 5-1.(a)；或是倆受試者不同時間出發，中間不停留，如圖 5-1.(b)等等。若在同方向路線中，兩位受試者無時間特徵不同的情況，本系統則無法辨識。

從圖 5-4 可知，除了同方向直線其表現較差，3D-DTW 演算法的準確度約為 97%，DTW and State 約為 94%，同方向直線兩者準確度不同的原因在於 3D-DTW 是全盤性的考量座標與時間的相似性，而 DTW and State 則分為兩階段的匹配，若路線具有明顯差異，則直接輸出結果，反之，則進入第二階段。在同方向直線時 Stop and Move 和 DTW and State 兩個演算法非常相似，因此兩者準確度差距較小；而由於不同方向直線和非直線的路線及方向差異非常明顯，因此 3D-DTW 和 DTW and State 兩個演算法的辨識率均接近 100%，Stop and Move 的準確度則落在 60%至 70%。

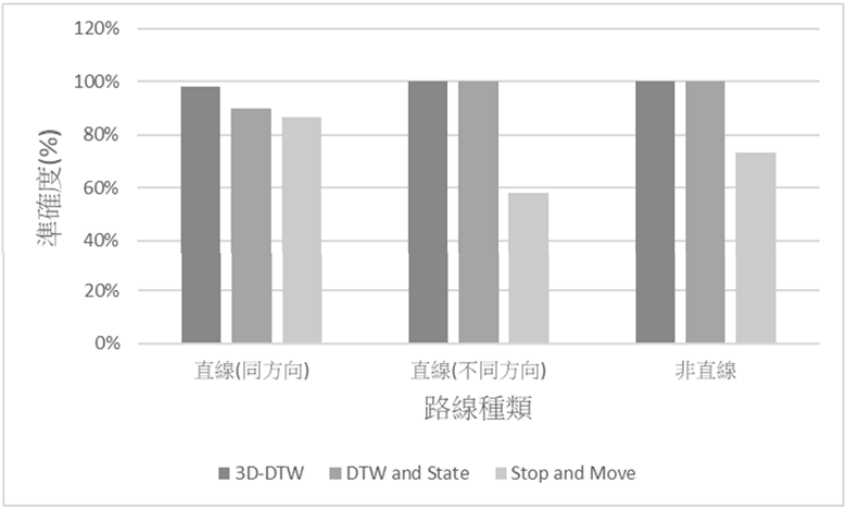

圖 5-4 各演算法的路線種類準確度

5.2 動作種類

為貼近現實生活中，本實驗將以活動種類區別，分別為線些行走和連續行走，其概念以圖 5-5.(a)(b)呈現。藉由這兩種活動種類來進行壓力測試，以評估我們系統的穩健性。最後，我們將測試兩者的配對表現並繪製結果在圖 5-6。

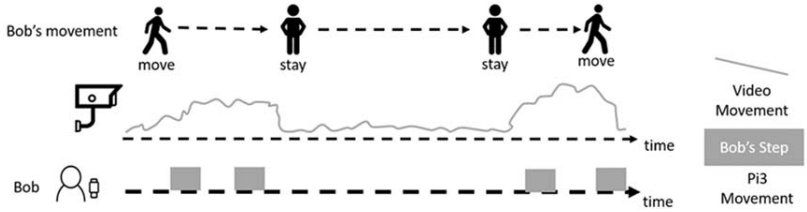

圖 5-5(a)間歇行走

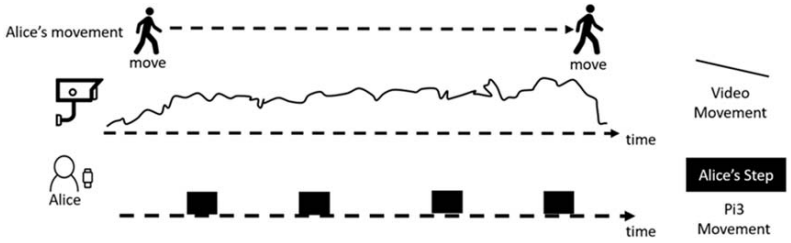

圖 5-5(b)連續行走

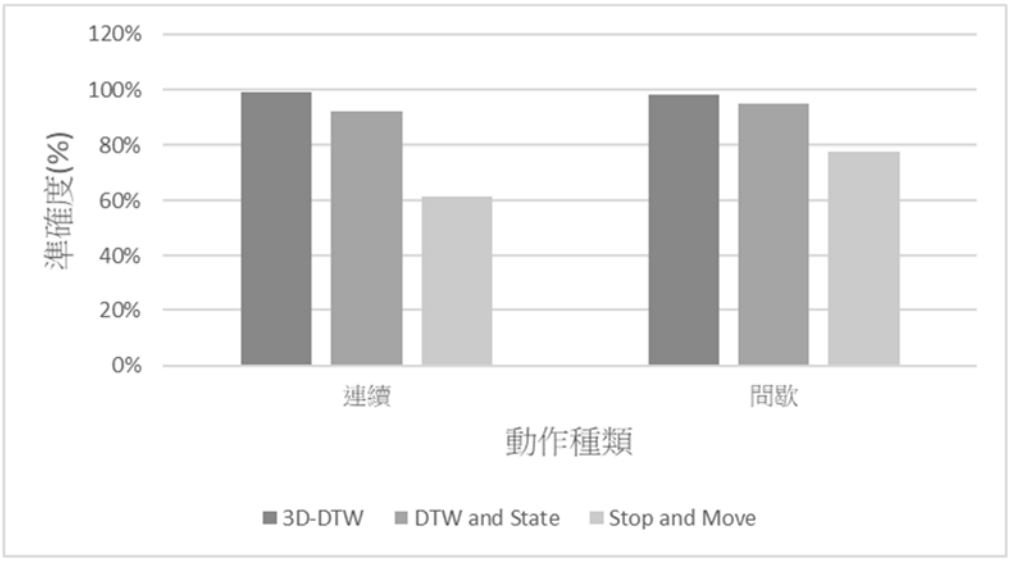

圖 5-6 各類演算法的動作種類精確度

5.3 實驗人數

接著我們模擬了人們在攝影機前面行動的場景，分別增加在攝影機面前行走的人數，然後分別將這些資料與三種融合算法配對。平均準確度繪製在圖 5-7 中。從圖 5-7 可明顯看出，人數對於 3D-DTW 的辨識準確度影響程度較小，則 DTW and State 和 Stop and Move 兩者算法非常相似，因此較易受人數影響準確度。儘管如此，DTW and State 仍然具有較高的穩定性，在五人的情況下仍達到約 85%的準確度，由此可確定本系統在辨識人群是非常準確且穩定的。

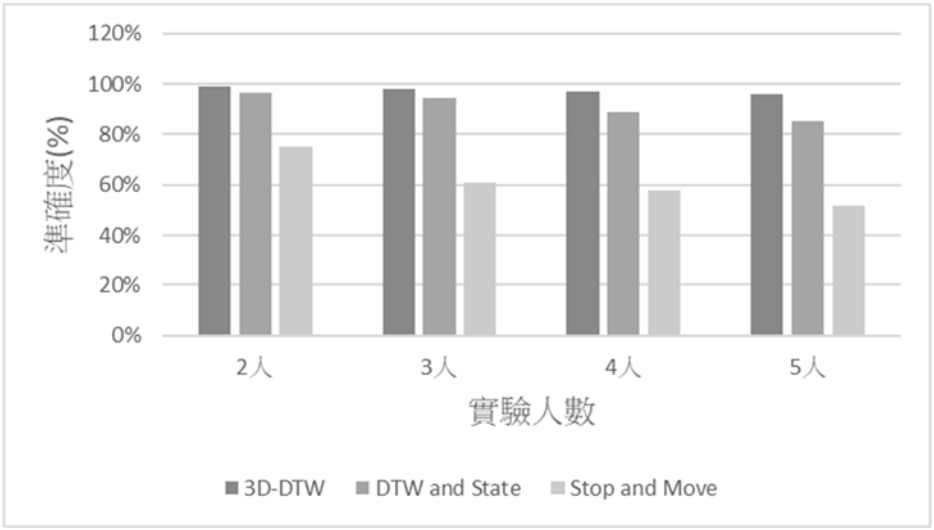

圖 5-7 各演算法的人數準確度

5.4 人數與裝置數目

此實驗項目目的在於實施更嚴格的模擬來娉孤本系統在實際情況下的可行性。依樣請受試者在他們手上配戴樹莓派 3 並在攝影機面前行走，但此時用兩種情況測試系統，第一種情況，將兩個影像軌跡和兩個具有唯一 ID 步伐資料配對，即為人數與裝置數目相等；另一種情況則是將三個影像軌跡和兩個具有唯一 ID 步伐資料配對，為人數與裝置數目不相等的情況。最後獲得的結果如圖 5-8 所示。

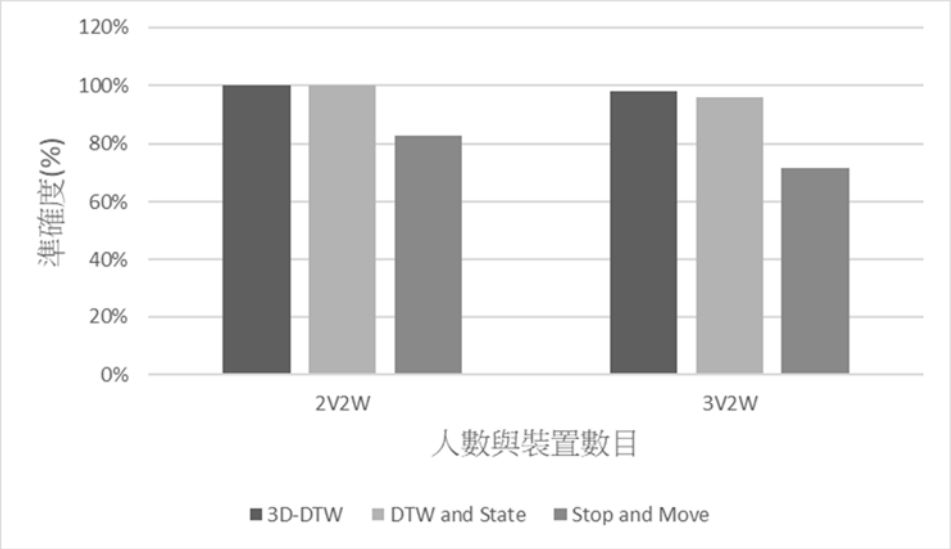

圖 5-8 各演算法人數與裝置數目的準確度

5.5 執行時間

為了更進一步分析 3D-DTW 算法和 DTW and State 算法的優缺點，我們將每次實驗數據的程式執行時間記錄下來，以利進行統計與分析，如表 5-1。DTW and State 算法的執行時間優於 3D-DTW 算法的主要原因為，執行時只需考慮時間特徵即可，則 3D-DTW 算法則需全盤性的考慮座標與時間之間的相關性，但這也是為何在各項實驗的準確度上，3D-DTW 算法比 DTW and State 算法優秀的原因。

表 5-1 各演算法程式執行時間

演算法	程式執行時間
DTW and State	0.0934 秒
3D-DTW	0.6899 秒

6. 結論與未來展望

在本論文中，我們提出了一個身分辨識系統，本系統融合了影像軌跡數據和慣性步伐數據，用以辨識人的身分。在配對影像軌跡數據和慣性步伐數據上，我們提出了一個融合算法，包含兩種演算法計算相關性，其優勢在於沒有任何繁瑣的數據標籤和費時的模型訓練過程，除此之外，我們的配對算法是利用統計中常用的度量，能非常有效率地配對兩種數據。本系統除了顯示了一群人的擴充性和穩定性，在人數與裝置數目不對等的情況下，也有非常亮眼且穩定的表現，因此相信本系統應用在現實生活中，應該也會有不錯的結果。

此系統目前尚在離線的狀況下進行配對工作，未來希望能擴充為及時的配對系統，為未來的物聯網或智慧工廠或大樓等等，提供更便利且安全的應用。

7. 參考文獻

[1] M. Rofouei, A. Wilson, A. Brush, and S. Tansley, "**Your phone or mine: fusing body, touch and device sensing for multi-user device-display interaction**," in Proc. ACM CHI,2012, pp. 1915–1918.

[2] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich,"**Common metrics for human-robot interaction**," in Proc. ACM HRI, 2006, pp. 33–40.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "**Face recognition: A literature survey,**" ACM computing surveys (CSUR), vol. 35, no. 4, pp. 399–458, 2003

[4] P. J. Grother, G. W. Quinn, and P. J. Phillips, "**Report on the evaluation of 2d still-image face recognition algorithms**," NIST interagency report, vol. 7709, p. 106, 2010

[5] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "**Deep face recognition**." in BMVC, vol. 1, no. 3, 2015, p. 6.

[6] F. Cafaro, A. Panella, L. Lyons, J. Roberts, and J. Radinsky, "**I see you there!: developing identity-preserving embodied interaction for museum exhibits**," in Proc. ACM CHI, 2013

[7] H. Li, P. Zhang, S. Al Moubayed, S. N. Patel, and A. P. Sample, "**ID-Match: A hybrid computer vision and RFID system for recognizing individuals in groups**," in Proc. ACM CHI, 2016, pp. 4933–4944

[8] T. Teixeira, D. Jung, and A. Savvides, "**Tasking networked CCTV cameras and mobile phones to identify and localize multiple people**," in Proc. ACM UbiComp, 2010.

[9] L. Xia and J. Aggarwal, "**Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera**," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2834–2841.

[10] O. D. Lara and M. A. Labrador, "**A survey on human activity recognition using wearable sensors**." IEEE Communications Surveys and Tutorials, vol. 15, no. 3, pp. 1192–1209, 2013.

[11] S. Lenman, L. Bretzner, and B. Thuresson, "**Using marking menus to develop command sets for computer vision based hand gesture interfaces**," in Proc. ACM NordiCHI, 2002.

[12] Y. Tao, H. Hu, and H. Zhou, "**Integration of vision and inertial sensors for 3d arm motion tracking in home-based rehabilitation**," The International Journal of Robotics Research, vol. 26, no. 6, pp. 607–624, 2007.

[13] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "**Fusion of inertial and depth sensor data for robust hand gesture recognition**," IEEE Sensors Journal, vol. 14, no. 6, pp. 1898–1903, 2014.

[14] C. Chen, R. Jafari, and N. Kehtarnavaz, "**Improving human action recognition using fusion of depth camera and inertial sensors**," IEEE Transactions on Human-Machine Systems, vol. 45, no. 1, pp. 51–61, 2015.

[15] J. Daugman, "**How iris recognition works**," IEEE Transactions on circuits and systems for video technology, vol. 14, no. 1, pp. 21–30, 2004.

[16] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "**Survey of wireless indoor positioning techniques and systems**," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 6, pp. 1067–1080, 2007.

[17] M. Alzantot and M. Youssef, "**UPTIME: Ubiquitous pedestrian tracking using mobile phones**," in IEEE Wireless Commun. and Networking Conf., 2012.

[18] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, and F. Zhao, "**A reliable and accurate indoor localization method using phone inertial sensors**," in ACM Conf. Ubiquitous Comput., 2012.

[19] P. Lawitzki and J. Charzinski, "**Application of dynamic binaural signals in acoustic games**," Master's thesis, Media University Stuttgart, Dec. 2011.

[20] D. J. Berndt and J. Clifford, "**Using dynamic time warping to find patterns in time Series**" in Proc. AAAIWS, 1994.

[21] W.-C. Chang, C.-W, Wu, R. Y.-C. Tsai, K. C.-J. Lin, and Y.-C. Tseng, "**Eye on you: Fusing gesture data from depth camera and inertial sensors for person identification,** "in IEEE 1CRA, 2018

[22] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "**Fusion of inertial and depth sensor data for robust hand gesture recognition**," IEEE Sensors Journal, vol. 14, no. 6, pp. 1898–1903, 2014

[23] W.-C. Chang, C.-W. Wu, Y.-C. Tsai, K. Lin, and Y.-C. Tseng, "**Eye on You: Fusing Gesture Data from Depth Camera and Inertial Sensors for Person Identification**", IEEE Int'l Conf. on Robotics and Automation (ICRA), 2018

[24] J.-W. Qiu and Y.-C. Tseng, "**M2M Encountering: Collaborative Localization via Instant Inter-Particle Filter Data Fusion**," IEEE Sensors Journal, Vol. 16

[25] Y.-C. Tsai, T.-Y. Ke, C.-J. Lin, and Y.-C. Tseng, "**Enabling Identification-Aware Tracking via Fusion of Visual and Inertial Features**," IEEE Int'l Conf. on Robotics and Automation (ICRA), 2019.