無標記式動作擷取技術運用於體感互動遊戲之研究

林哲緯(Chewei Lin)　盧天麒(Tainchi Lu)*
國立嘉義大學資訊工程學系
*e-mail: tclu@mail.ncyu.edu.tw

摘要

本文提出了通過使用低成本的靜態攝影機進行無標記人體動作捕捉。我們利用鉸鏈式人體骨架代表複雜的人體，並且不需要任何深度傳感器的輔助來取得深度資訊。首先將靜態照相機進行校準，取得適合的顏色和亮度。再根據靜態畫面與動態畫面的兩者差異，辨別出人體的輪廓。隨後將圖像劃分為多個大小一致的區域以增加辨識的效率。此外，姿勢決策樹已在初始化階段定義，目的是用於姿勢轉變的判斷，再根據不同的姿勢轉換不同的動作追蹤策略。最後，採用標準的 BVH 格式來存儲連續的運動數據，並提供其他系統進一步的使用。實驗結果顯示本論文所提出的方法，可以在成本大幅低於穿戴式動作擷取器的情況之下，擷取可接受準確度的人體連續動作。

關鍵詞：角色動畫、無標記動作捕捉、姿態估計、運動追蹤、動作感應遊戲

Abstract

The paper presents an easy-to-implement technique of performing a markerless humanoid motion capture by using a low-cost and stationary camera. We make use of articulated simplified skeleton to represent a complicated human body and do not require any auxiliary multiple depth sensors for obtaining depth information. First, a static camera is calibrated to obtain applicable color and luminance. A T-pose of a character is required to distinguish which part is the humanoid character within an image and to facilitate the motion tracking. Subsequently, we divide the image into a variety of uniform regions for rapidly estimating a human's pose. A pose decision tree has been defined in advance to benefit pose transitions while the character is changing a motion. Finally, a standard BVH color format is adopted to store the continuous motion data for further usage. Experimental results show that the proposed method is practical and feasible for capturing human motions. In particular, the cost of the proposed method is less expensive in comparison with the marker-based motion capture systems.

Keywords: Character animation, Markerless motion capture, Pose estimation, Motion tracking, Motion sensing game

1. 簡介

近年來電腦的圖形運算能力大幅提升，擬真的三維繪圖已不再遙不可及，在視覺效果的呈現與虛實互動的體驗都有更優異的表現，人們對於三維成像的擬真度要求相對更高，因此現今遊戲業者或是電影公司都會運用動作擷取來取得人體真實動作資料。而動作擷取技術可以區分成兩個部分，第一個部分為動作追蹤，首先將影像中的人體的關節做定義，並記錄關節的運動軌跡；第二個部分為動作辨識，此部分針對一連串的動作資料中，判斷一段時間內使用者所進行為哪一種動作，如走路、跑步、跳躍等。本研究將著重在動作追蹤的部分，並運用姿勢的辨識來輔助動作追蹤，以降低動作誤判的可能性與減少資料運算的時間成本。

動作追蹤區分為標記式與無標記式，本研究是基於動作追蹤與判斷的普及化為目標，無標記動作捕捉是一種沒有使用反射或電壓等標記點進行動作軌跡紀錄的技術，目前屬於一個快速發展且具高度挑戰的領域，應用方面如遊戲、監控、運動科學、臨床生物力學等領域，都運用了非常多無標記動作捕捉來進行人體姿勢的分析。使用有標記點的動作分析有以下幾項缺點，使用者穿戴標記點是很耗時的且有時具有被強迫穿戴的壓力，另外標記點的偏差可能使擷錄出的軌跡產生震動不平順的情形發生，另外錄製動作的環境也有所限制。常見的無標記動作捕捉系統為微軟公司出品的 Kinect 攝影機，雖然搭載 Kinect 的商業產品能提供有效率的即時無標記動作捕捉，但是仍有捕捉精確度不穩定與人體中心軸不能進行大幅度旋轉的問題[1]。因此本研究將以非常普遍的視訊攝影機進行動作錄製，以鉸鏈式人體模型合成骨架，再使用影像分割降低運算成本，並且運用馬可夫鏈的概念，設計了一套姿勢判別與轉換機制輔助動作追蹤，完成無標記動作擷取。本研究運用了動作追蹤技術並應用於體感互動舞蹈遊戲，玩家可以利用簡單且低成本的視訊設備，與虛擬角色世界的人物互動，當使用者跟隨著提示做出不同舞蹈動作時，使用者所選擇的角色將會跟隨使用者舞動身體；本研究希望提供使用者採用低成本的設備就可以體驗有趣的虛實互動遊戲。

本論文在第二章節中，將會敘述近年來無標記動作捕捉的相關技術發展與鉸鏈式骨架模型的種類，在第三章節描述無標記動作動態追蹤的技術，在此章節中我們先說明如何進行攝影機的校正與影像分析法的使用時機，接著完成人體骨架初始化等前置動作，並進行人體動作預測與關節位置的追蹤，最後說明如何匯出動作資料，在第四章節將展示動作追蹤的執行結果，最後在第五章節總結全文，並提到未來的研究發展方向。

2. 相關工作

2.1 無標記動作捕捉

近年來使用影像追蹤人體動作的演算法[2, 3, 4]有著令人驚艷的成長。部分的研究嘗試使用單一影像中推斷人體姿勢[5, 6]，或者從單一影像取得人體運動資訊[7]。對於姿勢評估可以利用額外的感應器，如慣性感應器[8]或深度感應器[9]，但是感應器等硬體裝備較為昂貴，使用者穿戴感應器後在操作上也較不方便，且錄製對象較為侷限。[10]提出了一種演算法與裝置 Kinect 可以追蹤使用者的運動深度資訊，他們同時優化了人體骨架的姿勢和特徵圖像的對應，與追蹤過程中使用者表面和幾何對應點的位置。然而 Kinect 重點在於捕捉少量使用者的動作，無法掃描大量人群的動作資訊[11]。因此本研究基於普及與未來發展性，將使用從單一影像擷取動作資訊的方式錄製人體動作。

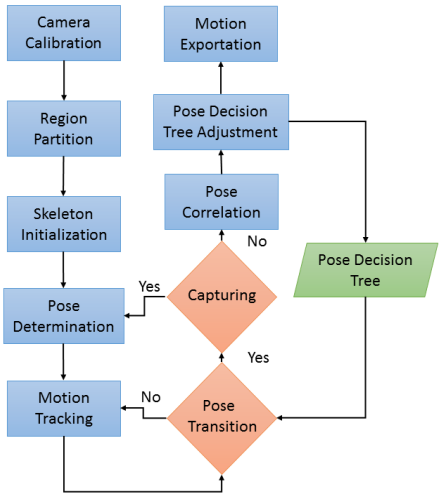
2.2 鉸鏈式骨架模型

鉸鏈式模型合成簡化的人體模型，提供較容易與快速的計算資料，[12]使用無網格的棒狀體重建模型，但出現了較多的偏差，[13]等人提出了另一種模型，由長方形構成軀幹。此種模型的關節共有 26 種自由度(DOFs)沒有被約束。再將此模型使用剪影分析與姿勢估計，並用高斯混和模型找尋在視圖中每個像素分布的位置，此種方法的精確度符合生物力學標準。更精準的特定主題模型由[14]提出，此關節模型使用錐形超二次曲面合成，並使用三維重建。[15]採用高斯圓塊(Gaussian blobs)定義球形的體積，並將球體附在骨架模型上。[16]使用的方法則是先用人工的方式定義出骨架模型，並使用三維網格與骨架關聯性的權重運算進行蒙皮。本研究將採用[17]所提出的鉸鏈式模型，定義各關節活動的限制，以防止模型網格不正常的變形。

3. 無標記人體動作動態追蹤

3.1 概述

本系統所提出的架構圖如圖一所示。本研究將使用無標記的方式追蹤人體動作，使用者可以在無穿戴任何裝備下，使用一般的視訊攝影機連續拍攝使用者的動作，而虛擬世界的角色就會跟隨使用者的動作作出相對應的動作或反應。本研究為了達到即時運算並降低來源資料量，使用了區域分析法，將攝影機拍攝的畫面分割成多個區塊，定義每個區塊為一個感應點，經由使用者擺出的不同姿勢，所感應到的區塊也有所不同，利用這些感應區塊所得到的資訊作為姿勢判斷與動作追蹤的依據。因人體的連續動作相對複雜，四肢與頭部彼此間的相對位置會隨著不同的動作而有所改變，因此本研究先判別使用者現在所處的姿勢狀態，再依據不同的姿勢進行動作追蹤。而判別姿勢的狀態與姿勢間的轉換則由姿勢決定樹來決定。人體動作追蹤的方法則為尋找畫面中四肢與頭部所在的位置，再經由人體骨架的建立而定義出各關節所在的位置。透過上述流程後就可以追蹤到每一個影格中使用者即時的動作，本研究將會提供不同的情境讓使用者跟隨指示作出舞蹈動作，虛擬角色則會跟隨著使用者一起跳舞。



圖一、 系統流程圖

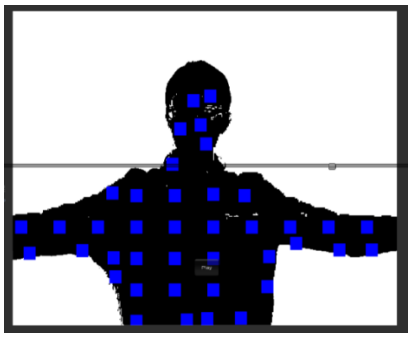
3.2 攝影機初始化

本人體動作錄製系統採用一般的視訊攝影機進行動作錄製，在錄製前須先將攝影機作色彩與亮度的校正以降低錄製期間的誤差。使用方法為先取得攝影機畫面的長寬比並做正規化，取得拍攝畫面的解析度作進行分區分析法的畫面切割。接著紀錄現場環境的影像資料，並請使用者站立至畫面中央，擺出校正姿勢(T-pose)，使用人體特徵偵測與二值化判斷出人體位置，並用高斯濾波器(Gaussian filters)進行卷積(convolved)，然後利用連續高斯模糊化影像差異來找出關鍵點，如式一。

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad \text{(式一)}$$

其中$I(x, y)$為原始影像，$L(x, y, \sigma)$為高斯模糊影像，$G(x, y, \sigma)$為高斯函數。將攝影機所拍攝的畫面做判斷，找出人體各部位的特徵，並在每台攝影機中做特徵標記，如四肢末端點及頭部位置，提供下一步建構人體骨架的基礎資料。

3.3 分區分析法



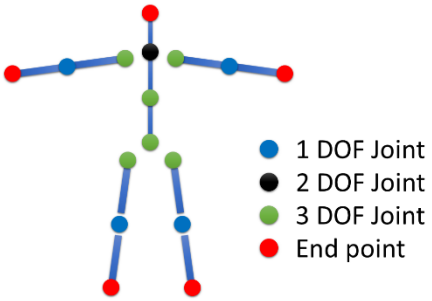
圖二、分區分析法示意圖

一張影像所儲存的資料相當龐大，要即時處理影像串流中的每個影格資訊，所使用的硬體設備相對要求較高，因此本研究將影像分割成數個區域，把每個區域視為一個感應點，當有物體進入感應點的範圍時送出偵測訊號，再將所有感應點的資訊作為動作判斷的依據，如圖二所示。而感應點的敏感度則根據權重值P_w的大

小來決定，不同的姿勢將定義不同的權重值，此權重值將影響姿勢判斷的準確度。較複雜且較難辨識的動作，如舉手加揮舞，權重值將會設的比較低，以助於系統快速的辨識出相對應的動作。如較簡單且須精準判定的動作，權重值將設的比較高，以降低誤判率。

3.4 人體骨架初始化

在此系統中將先取得視覺赫爾(Visual Hull, VH)的人體外型輪廓(Shape from Silhouette, SfS)，再將骨架正規化，定義一組符合使用者身型的剛體骨架。藉由初始化取得的特徵標記點資料，定義好四肢的末端點與頭部位置，再用人體結構學計算出各關節的位置、骨架的長度及骨架的階層關係，如圖三所示，此骨架已定義各關節的旋轉自由度及鉸鍊結構，並運用反向運動學的鏈結關係，使身體各部位移動符合人體運動學的規則。



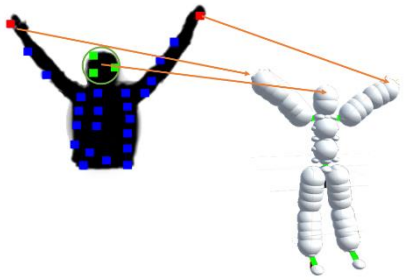
圖三、人體骨架基礎模型及其各關節的自由度做定義

3.5 人體骨架與影像對應

使用區域分析得到的感應點利用人體四肢都處於末端點的特性，定義出末端點，再由人體骨架初始化所得到的剛體骨架，將各人體關節透過式二

$$\mu = \begin{pmatrix} [\mu^p]_x/[\mu^p]_z \\ [\mu^p]_y/[\mu^p]_z \end{pmatrix} \qquad \text{(式二)}$$

投影至平面的影像中，與感應點作對應如圖四所示，便可經由剪影的變化驅動人體骨架。



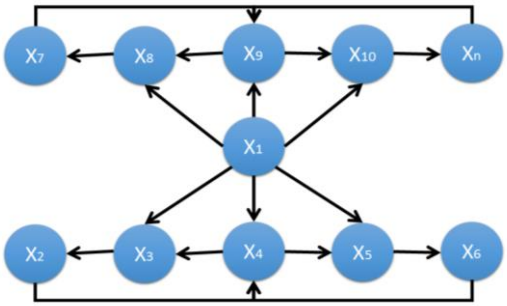
圖四、人體骨架與影像對應示意圖

3.6 人體動作預測

人體動作是由多個姿勢串接而成，因此本系統將採用不同姿勢用不同追蹤方法來錄製關節的軌跡。首先從動作關係資料庫中讀取現在處於的動作與下一個可能發生的動作，經由動作資料庫記錄的姿勢轉換機率模型與權重值，來預測進入下一個姿勢的可能性，所使用的演算法為條件隨機場(conditional random field, CRF)，設 S 為所有姿勢 X_0 、X_1 、X_2 ……X_n 的集合，

$$P(X_{n+1} = x | X_0, X_1, X_2 \ldots, X_n) \qquad \text{(式三)}$$

由式三可求出當在某姿勢狀態時，進入下一個狀態的機率。而進入每個姿勢都有權重值 W，當X_n的機率值 P_n 大於 W_{n+1} 時則姿勢狀態進入 X_{n+1}，而權重值的調整依據為系統的統計經驗值與使用者的錯誤回報，條件隨機場模型如圖五所示。



圖五、條件隨機場模型

3.7 人體關節位置追蹤

從人體動作預測中已得知目前可能的姿勢狀態，依照不同的姿勢狀態使用此姿勢最佳的關節追蹤法，演算法為先找出目前影像中最大範圍的感應區域，以此區域為圓心，建立動態半徑為 r 的圓，再逐漸縮小 r 的數值並判斷是否人體外型輪廓的範圍座標落在圓上。

再由各攝影機所拍攝的人體輪廓，依照初始定義好的關節規則與骨架長度重建骨架，在本系統中我們建立了懲罰函式$X(t,\psi_j)$ [18]，此函式的功能主要是外側的容許限制範圍$[l_l, l_h]$，$t\epsilon S$，因此我們可以定義懲罰函數$E_{lim}(j)$來檢測人體輪廓所覆蓋的感應點測量面積是否有超出限制，懲罰函數如下

$$E_{lim}(\psi_j) = \left(\int_{t\epsilon s \wedge x(t,\psi_j)<l_l} l_l - X(t,\psi_j)dt + \int_{t\epsilon s \wedge x(t,\psi_j)>l_h} X(t,\psi_j) - l_h dt\right)^2 \quad \text{(式四)}$$

最後再進行骨架長度與關節自由度檢查是否符合人體運動學限制，如果符合便紀錄目前影格的關節資訊，如果不符合，則轉換骨架重建策略直到符合人體運動學為止，如圖六所示。記錄關節的旋轉角度時，使用矩陣紀錄每影格的 x、y、z 軸的旋轉角度。最後將即時的動作資訊送至三維虛擬角色，使虛擬角色可以即時的做出相對應的反應

圖六、(a)取得人體輪廓的四肢位置與頭部位置；(b)將在初始化定義的肩膀寬度與人體高度建立人體骨架；(c)檢查關節自由度是否符合人體運動學，並依照人體輪廓重新修正關節位置

3.8 姿勢關聯性修正

本研究所利用的追蹤方法為先進行姿勢判斷，依據不同的姿勢再進行不同的追蹤策略，因此姿勢的判斷對錯十分重要。姿勢轉換機制需依照姿勢與姿勢間的連接關係與轉換權重所決定，當系統經由統計後判斷出某兩個有關聯的姿勢轉換機率很高，那麼系統將預測接下來錄製動作時可能在這兩個姿勢也較容易進行轉換，因此將這兩個姿勢轉換的權重值降低初始值的五個百分比，如果使用者持續在這兩個姿勢間轉換，系統將再調降初始權重值的十個百分比。另外當兩個姿勢本來無關聯，但使用者不斷地進行這兩個姿勢的轉換，系統將會建立這兩個姿勢的關聯性並給與初始權重值。

3.9 輸出動作資料

本研究中為了方便其他商業系統使用動作軌跡資料，因此採用層次模型(Biovison Hierarchy, BVH)的檔案格式來描述骨架移動的軌跡，在骨架初始化中就預先定義每個骨架的階層關係，將紀錄的骨架與關節資訊依照 BVH 格式輸出至檔案中。匯出動作資訊後，使用者可以針對錯誤的姿勢轉換影格進行標記，本系統將會依據使用者所標記的影格，進行姿勢間的關係權重與轉換機率模型進行修正，以提高下次錄製動作的正確率。

4. 執行結果

本章節將說明本研究所提出來整體流程的執行結果，首先在攝影機校正的過程中，我們將複雜的環境經由二值化與背景值相減濾掉不要的資訊。再由分區分析法將畫面切割，由圖七可以看到我們將 600*800 像素的畫面切割成 100 格，每格只要感應到有物體進入感

應區域就會傳送感應訊號，系統將可以判斷現在所處的姿勢狀態，並由姿勢決策樹決定何時進行姿勢的切換，在每個姿勢狀態，將有所對應的動作追蹤方式，由研究成果顯示透過影像資訊擷取出人體的動作，並且將影像做切割可以加速運算速度。區域分析法的測試環境中，當使用者做出圖八(a)雙手向上抬舉的姿勢，或是圖八(b)單手闊胸姿勢時，感應點感應到畫面有物體進入，則標記成不同顏色，當這些感應點構成圖像符合左方其中一種姿勢時，虛擬世界的角色就會做出該姿勢。本研究嘗試做出八種差異較大的動作做為追蹤測試，當使用者擺出關鍵姿勢後，虛擬角色將會跟著擺出相對應姿勢，接著使用者做出小幅度擺動，虛擬角色將會即時的跟隨運動。當使用者做出的動作已判定需切換姿勢狀態時，系統會進行姿勢轉換，並重新進行動作追蹤，如圖十所示。經過反覆測試得到如表一數據，當動作較為明顯，如站立雙手上舉等動作時，辨識成功率較高，當動作較為複雜且易於與其他動作混淆，其動作的辨識成功率較低。

經由實驗可得知同樣是舉手動作，雙手平舉的成功率相較其他舉手動作成功率較低，推測原因為做出雙手平舉與手插腰動作雙手都會往兩側移動，因此辨識時容易產生混淆，辨識率成功率較低；另外單手橫移動作有深度的變化，本系統較難利用剪影作明確深度區分，因此辨識率成功率較低。而如果動作的獨特性很高且前置動作也很明確，如雙手上舉或雙手打斜，該動作的辨識成功率將會高於其他動作。

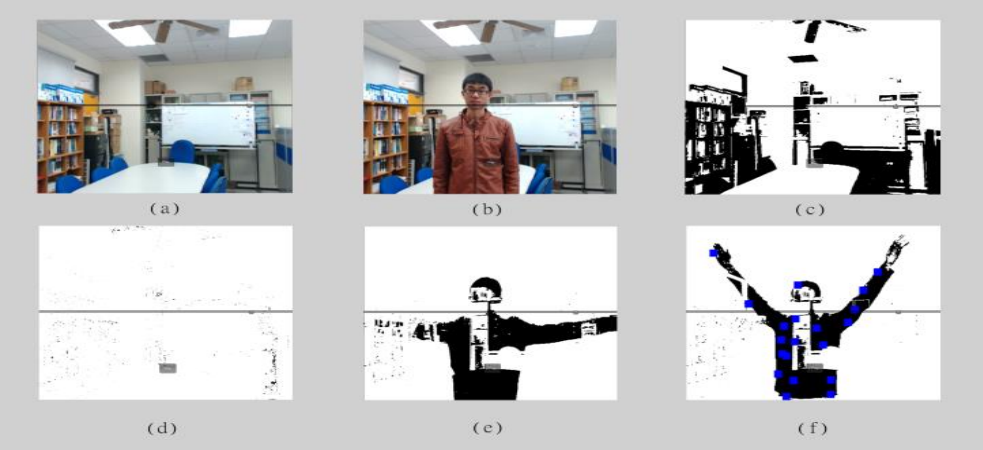
5. 結論與展望

本研究已成功的使用視訊攝影機擷取出人體動作，使用分區分析法達到即時運算的結果，但下半身的辨識效果較不理想，因此只有展示上半身的追蹤結果。分區分析法中畫面的分割數量與感應點的敏感度，將影響判斷姿勢是否轉換的準確率。經由姿勢轉換策略所進行人體姿勢預測，判斷出目前使用者的姿勢，再進行小範圍的動作追蹤，不但可以降低追蹤的誤判率，更可以增加系統運算的即時性。

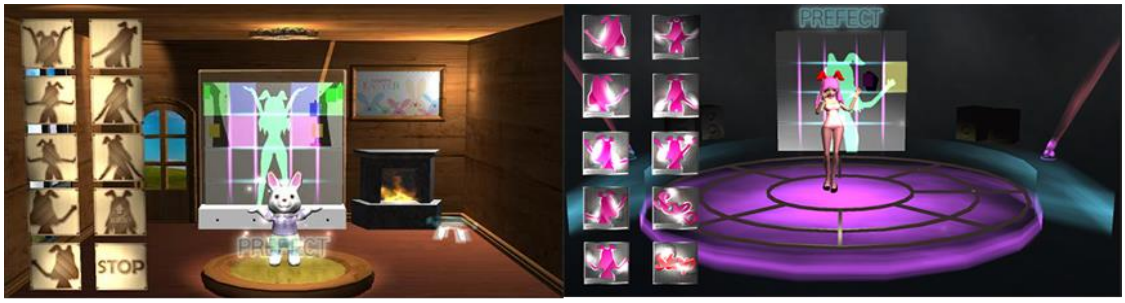
未來如果加上多台攝影機的輔助將可結合立體深度資料，本研究的成果將可以運用在如醫療行為監測、智慧家庭或是保全系統等需要監測人體行為的環境；以醫療監測為例，當攝影機拍到有病患不慎跌倒或不適等姿勢時，系統將發出警報，醫護人員可以即時的進行救護反應。本研究最大的貢獻在於運用非常普及的視訊攝影機取得的影像資料就可以擷取出人體動作，且可以即時運算，這解決了過去此領域在運算時相對耗時的問題。

致謝

本論文由科技部計劃編號 MOST 103-2221-E-415-012 補助辦理執行。

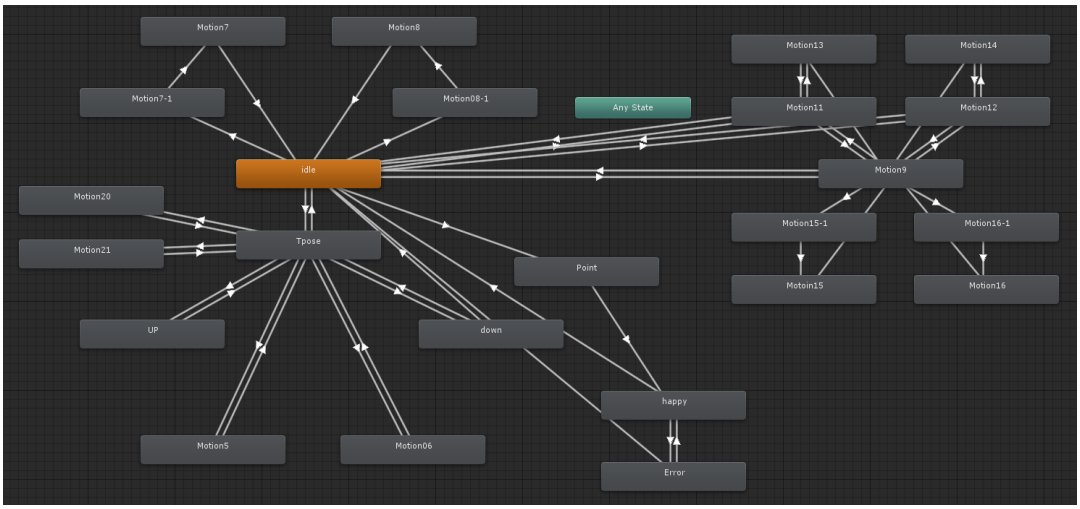


圖七、(a)攝影機拍攝的複雜靜態環境；(b)記錄靜態與動態畫面之間的差異；(c)進行影像二值化；(d)濾掉不要的資訊；(e)使用者回到畫面中查看效果；(f)分區分析法判斷感應區



(a) (b)

圖八 、區域分析法的測試環境。(a)雙手向上抬舉姿勢；(b)單手闔胸姿勢



圖九、姿勢轉換關係圖

圖十、動作追蹤實驗結果

表一、辨識次數數據

動作名稱	比出次數	成功次數	辨識成功率	備註
站立	100	82	82%	誤判成插腰
插腰	100	76	76%	誤判成站立
單手橫移	100	20	20%	系統無反應
單手向上舉	100	82	82%	誤判成單手橫移
雙手平舉	100	49	49%	系統無反應
雙手上舉	100	95	95%	
雙手打斜	100	92	92%	
側身	100	73	73%	系統無反應、判斷錯邊

參考文獻

[1] Choppin, S., "Engineering sport blog," *the Centre for Sports Engineering Research. Available from: http://engineeringsport.co.uk/2011/05/09/kinect-biomechanicspart-1/*, July, 2013.

[2] Bregler, C., and Malik, J., "Tracking people with twists and exponential maps," ***In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition***, pp. 8–15, 1998.

[3] Deutscher, J., and Reid, I., "Articulated body motion capture by stochastic search," ***International Journal of Computer Vision,*** 61, 2, pp. 185–205, 2005.

[4] Balan, A., Sigal, L., Black, M., Davis, J., and Haussecker, H., "Detailed human shape and pose from images," ***In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition***, pp. 1–8, 2007.

[5] Andriluka, M., Roth, S., and Schiele, B., "Monocular 3D pose estimation and tracking by detection," ***In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition***, pp. 623–630, 2011.

[6] Ionescu, C., Li, F., and Sminchisescu, C., "Latent structured models for human pose estimation," ***In Proceedings of International Conference on Computer Vision***, pp. 2220–2227, 2011.

[7] Wei, X., Chai, J., "VideoMocap: modeling physically realistic human motion from monocular video sequences," ***ACM Transactions on Graphics***, Vol. 29, No. 4, pp. 1–42, 2010.

[8] Pons-Moll, G., Baak, A., Gall, J., Leal-Taixe, L., Mueller, M., Seidel, H.-P., and Rosenhahn, B., "Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling," ***In Proceedings of International Conference on Computer***

Vision, pp. 1243–1250, 2011.

[9] Baak, A., Muller, M., Bharaj, G., Seidel, H.-P., and Theobalt, C., "A data-driven approach for real-time full body pose reconstruction from a depth camera," ***In Proceedings of International Conference on Computer Vision***, pp. 1092–1099, 2011.

[10] Ye, G., Liu Y., Hasler, N., Ji, X., Dai, Q., and Theobalt, C., "Performance capture of interacting characters with handheld Kinects," ***In Proceedings of European Conference on Computer Vision***, pp. 828–841, 2012.

[11] Ya-Li, H, and Pang, G.K.H., "People counting and human detection in a challenging situation," ***IEEE Transactions on Systems, Man, and Cybernetics***, Vol. 41, No. 1, pp. 24–33, 2011.

[12] Moschini, D., and Fusiello, A., "Tracking human motion with multiple cameras using an articulated model," ***In Proceedings of Computer vision/computer graphics collaboration techniques***, pp. 1–12, 2009.

[13] Kohli, P., Rihan, J., Bray, M., and Torr, P.H.S., "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts." ***International Journal of Computer Vision*** pp. 285–298, 2008.

[14] Sundaresan, A, and Chellappa, R., "Model driven segmentation of articulating humans in Laplacian Eigenspace," ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, pp. 1771–1785, 2008.

[15] Caillette, F., Galata A., and Howard T., "Real-time 3-D human body tracking using learnt models of behavior," ***Computer Vision and Image Understanding***, pp. 112–125, 2008.

[16] Gall, J., Stoll, C.D., Aguiar, E., Seidel, H-P., Theobalt, C., and Rosenhahn, B., "Motion capture using joint skeleton tracking and surface estimation," ***In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,*** pp. 1746–1753, 2009.

[17] Ogawara, K., Li, X., and Ikeuchi, K., "Marker-less human motion estimation using articulated deformable model," ***In Proceedings of IEEE International Conference on Robotics and Automation***, pp. 1-10, 46–51, 2007.

[18] Elhayek, C., Stoll, N., Hasler, K.I., Kim, H.-P., and Seidel. C., "Spatio-temporal motion tracking with unsynchronized cameras," ***In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition***, pp. 1870-1877, 2012.