國立臺灣師範大學
資訊工程研究所碩士論文

指導教授：　梁祐銘博士
　　陳世旺博士

Human Action Analysis Using RGB-D Sensors
利用 RGB-D Sensors 進行人類動作的分析

研究生：胡碩宸 撰

中華民國　一百零二 年　七　月

摘要

利用 RGB-D Sensors 進行人類動作的分析

國立臺灣師範大學資訊工程學系　胡碩宸

本研究利用視訊資料進行人類動作的分析，目的在於發展一套通用的人類動作分析技術可以應用在不同的領域上，如公共安全：機場、地鐵、體育館、購物中心等公共區域或是大樓中的自動化監控系統，偵測是否有人有異常行為（例如破壞公共區域）。或者居家照護系統：偵測家裡的孩童或老人是否有跌倒或爬到高處等危險的行為，若有危險行為發生便通報家屬及照護人員。

本研究利用微軟所開發的RGB-D Sensors（亦即Kinect）來擷取人體3D關節資訊，並計算關節夾角當作人類姿勢的特徵向量，由於這些特徵向量維度極高，因此我們利用流形學習（manifold learning）之等構映圖（Isomap）進行降維，並在低維度的等構映圖空間進行基本動作的切割與分群。接著將每一群的基本動作給定一個語意上的闡述並形成一個編碼書（codebook），最後此編碼書可以用來對測試者進行動作的辨識。

關鍵字：流形學習、等構映圖、動作辨識、Kinect

Abstract

human action analysis using RGB-D Sensors

Shuo-Chen Hu

In this study, we develop a human action analysis technology from video data, which can be applied to many different fields, such as the public safety monitoring system in subways, shopping malls, other public area or buildings to detect abnormal behavior; the home care system: to detect the danger behavior of children or elders, such as falling at home.

We use RGB-D Sensors (i.e. Kinect), developed by Microsoft, to retrieve the body joints 3D information and to calculate the joint angle as the feature vector of a human posture. Since the dimension of the feature vector is very high, we apply the Isomap algorithm, which is a manifold learning approach, to reduce the dimension. Then, each atomic action is segmented and clustered in the Isomap space, and all of the clusters from a codebook. Finally, we can use the codebook to recognize the tester’s actions.

Keywords: manifold learning, Isomap, action recognition, Kinect

誌謝

在念研究所的路上，我碰到了好的指導教授、同學以及提攜後輩的學長。我非常感謝陳世旺老師指導我在研究所這條路上的指導，教授不但指導我在學術上的精進，同時也交會我做研究的態度以及待人處事的原則，同時老師對於研究的熱忱也是我最佩服的，即使是假日，陳世旺老師還是不倦怠的來學校進行他的研究，非常慶幸給陳世旺老師指導。

非常感謝另一位指導教授梁祐銘學長帶給我非常多的經驗，梁祐銘教授每個禮拜都會撥一天抽空來和我做實務上的討論，對我幫助非常之大，同時也透過他學習到很多學術上的知識，以及如何撰寫論文及報告的邏輯，我真的學習了非常多的東西。在此也要感謝實驗室的王俊明學長、羅安鈞學長、許之凡學長、李俊億學長、郭彥佑學長和黃世豪學長，他們在做實驗及論文的經驗真的幫助了我非常多，省去我很多走冤枉路的時間，他們也時常在實務上給我許多寶貴的意見，沒有他們我在研究所的路上會辛苦很多。

接著要感謝和我同一屆的石翊辰同學、林育誠同學和呂佳儒同學，我們大家彼此扶持照顧，互相討論學術及實務上的經驗。也非常感謝他們三位配合我的實驗，沒有他們我的論文是做不出來的。很高興最後我們大家可以一起完成自己的論文，踏上人生旅途的下一站。

最後要感謝大家不斷的包容我,還不斷地給我鼓勵,這兩年我真的成長不少，也交到了不少朋友，也期望未來在職場上能夠和學校的前輩以及同學一起合作，相信會是非常愉快的經驗。

胡碩宸

2013/06/20

目錄

圖表目錄

第一章　　簡介

1.1、研究動機：

近來的高鐵炸彈案（圖 1-1），歹徒帶著四個行李箱炸彈放置在高鐵列車中，意圖在高鐵中引爆，最後雖無造成人員的傷亡，但對台灣社會卻已造成極大的震驚。由於高鐵的監視器及人員不足，才讓歹徒有機會進入高鐵站裡面做出危害他人安全的行為。在人員不足的情況下，若能夠在車站多架設幾台監視器，並利用監視器所取得的視訊資料進行人類動作的分析，一旦有人做出危險或不尋常的行為，便會被偵測出來，如此可避免一些不必要的意外發生，且在找尋嫌疑犯的程序上也可節省不少的時間及人力。

在 2005 年 7 月 7 日，英國倫敦利物浦街和阿爾蓋特車站間傳出爆炸聲，隨後，又有兩個地鐵車站相繼發生大爆炸，倫敦地鐵宣佈停運（圖 1-2）。同年 7 月 21 日，英國倫敦地鐵再次發生爆炸案，而作案方式和之前的案件如出一轍，將英國倫敦的恐怖氣氛推到最高點。最後倫敦地鐵爆炸案在英國警方的努力不懈下宣告偵破，多名犯人和嫌疑犯被逮捕歸案，而這一切都要歸功於監視攝影機的視訊資料證據。英國是世界上監控力度最大的國家之一，全國大約有 420 萬個監視器，超過 6000 個監視器覆蓋於 272 個地鐵站和車廂內，每位公民平均每天被拍攝的次數達

到 300 至 600 次之多，總時間超過三個小時，所以要交叉比對找出可疑的嫌疑犯得花上非常多的時間，執行上相當困難。因此我們若能利用這些大量的視訊資料來進行人類動作的分析，自動偵測可疑的視訊片段，將可以省下非常多的時間在做人工檢查。此外也能夠偵測出一個公共區域是否有入侵者破壞公共區域，或者炸彈客放置炸彈等異常行為。因此我們的研究目標便是如何利用視訊資料進行人類動作的分析。

圖 1-1：高鐵炸彈案（來源：TVBS）　　圖 1-2：倫敦地鐵爆炸案（來源：維基百科）

利用視訊資料進行人類動作的分析所涵蓋的應用面非常廣泛，是近幾年來重要的研究領域之一。例如可以被應用在機場、地鐵、體育館、購物中心等公共區域或者大樓中的自動化監控系統（圖 1-3），偵測是否有人異常行為（例如破壞公共區域）〔His 08〕；也可被應用在智慧家庭系統（圖 1-4），對獨居老人或肢體有殘障的人做居家照護，分析其是否有跌倒或異常的行為發生，若有異常行為或

危險行為發生，便發出警訊通知照護人員馬上進行處理〔Liu 10〕；或是應用到視訊註解（video annotation），可以讓使用者方便地從影片中搜尋某些特定動作的片段〔Su 07〕，例如從籃球比賽中找出投籃的片段；人類動作分析也可以被運用來分析運動員的動作來做為訓練時的範本〔Bid 10〕；或作為人機互動的界面或者互動式遊戲（圖 1–5）〔 Wei 09〕，而在微軟推出 RGB–D Sensors〔Zha 11〕後，此應用更蓬勃發展。

圖 1–3：監視系統（來源：台灣黃頁）

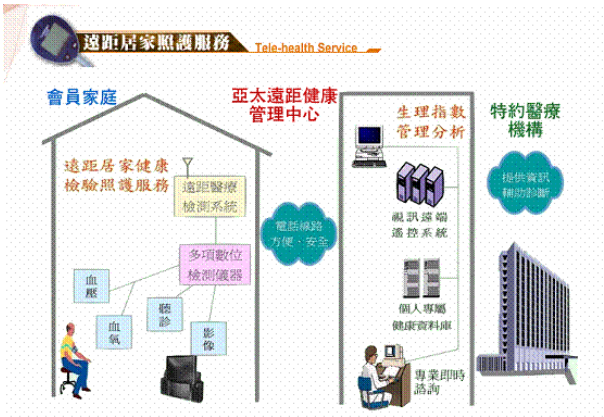

圖 1-4：智慧家庭系統（來源：老人福祉科技）

圖 1-5： 互動式遊戲（來源： http://www.youtube.com/watch?v=umJbeFUt8aI）

本研究目標是利用微軟推出的 RGB-D Sensors 來進行人類動作的分析，將此技術發展茁壯後可以應用到任何領域，只是動作資料庫對應不同領域去做變換。

1.2、研究目的：

本研究目的是要發展一個人類動作分析的技術，由於一段包含人類活動的視訊資料是由一連續包含人類姿勢（human posture）的影像序列組合而成，因此人類動作的分析首先必須找出影像序列中每一段連續的人類姿勢所形成之基本動作（atomic action），接著再由基本動作序列中找出連續的基本動作子序列形成之活動（activity），並用語意式描述來表示人類的活動。因此語意式地詮釋人類活動包含三個重要的步驟：1）如何將一段含有人類活動的影像序列切割成一段段的基本動作；2）如何辨識每個切割出來的基本動作；3）最後如何利用語意的方式來描述這些基本動作所形成的人類活動。因此在我們的研究中將人類動作的分析分成三個階層（圖 1-6）：姿勢、基本動作及活動，姿勢是人類動作最基本的元素，即一張影像中所呈現的人類姿態，基本動作是指一個在人類肢體等級上做基本移動（atomic movement）的描述，由一連續姿勢所組合而成。活動則包括一串連續的基本動作，並以高階語意（semantic）來描述。

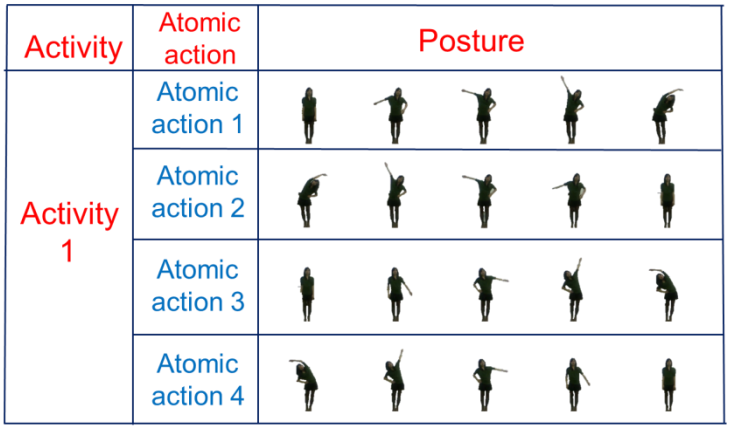

圖 1-6：人類行為的三個階層

1.3、文獻探討：

目前已經存在一些針對此研究作概述的論文〔Wan 03〕、〔Tur 08〕和〔Pop 10〕。Wang 等人〔Wan 03〕表示人類動作的分析通常可分成人類偵測與追蹤的低階處理，及詮釋人類動作的高階處理。低階的偵測與追蹤處理工作已研究了一段滿久的時間，已經有比較完整的方法與結果；而相對的，高階的動作詮釋仍然是一個有挑戰的研究議題。TREC Video Retrieval Evaluation（TRECVID） 在 2008 年後也將人類動作的分析列入一項非常重要的議題。

要從事人類動作分析的研究，首先必須從影像序列中切割並辨識出每一個基本動作，最後再用語意的方式來描述這些基本動作所形成的人類活動。目前大部份的研究都是假設所取得的影像片段是由一些連續基本動作組成的單一動作之視訊片段，如 KTH 動作資料〔Sch 04〕及 Weizmann 動作資料庫〔Bla 07〕中，

每一個視訊片段都是包含一個週期性的動作，如走路、跑步揮手等。因此大部份的研究都聚焦在如何辨識這些單一動作之視訊片段。

一般而言，現存的人類動作辨識方法大概可以依據其所採用的人類軀幹模型類別而分成 2D 及 3D 兩種類別〔Moe 01〕。在 2D 的方法中，Haritaoglue 等人〔Har 00〕提出 W4 系統，此系統為一即時監控系統，主要目的是在戶外環境中偵測及追蹤多人並且監控他們的活動。他們利用 2D 人體輪廓影像在垂直及水平方向的投影量來辨識人類的姿勢，包括站著、坐著、彎曲及躺著等不同的姿勢。在〔Bob 01〕中，Bobick 及 Davis 提出一個視覺的方法來描述及辨識人類的動作。首先，他們將一連續的影像框架（frame）堆積成一個 2D 的時間模板（temporal template）。他們利用動作能量影像（motion energy images, MEI）及動作歷史影像（motion history images, MHI）來描繪人類的動作。接著他們從 MEI 及 MHI 中擷取出運動的特徵，並且將這些特徵運用於模板比對（template matching）以便作人類動作的辨識。在〔Li 07〕中，Li 等人提出一個在跳水影片中自動分析複雜的運動員動作，主要目的在提供生理特徵（biometric）的測量及視覺工具來輔助運動員的訓練。他們利用 2D 人類關節相連的模型（articulated human body model）及外型分析技術來獲得運動員主要身體關節的角度。最後，他們為運動員的訓練提出兩項視覺分析的工具：motion panorama 及 overlay composition。在〔His 08〕中，Hsieh 等人提出一個姿勢辨識系統可以直接從視訊資料中分析

人類的動作。在他們的方法中，每一個動作的影像序列被轉換成一連串的姿勢序列。他們對人類姿勢作三角化成三角網格（triangular meshes），然後從中擷取出兩個特徵：骨骼特徵（skeleton feature）及中心上下文特徵（centroid context feature）。第一個特徵是當作人體姿勢粗劣的描述方法，而第二個特徵是用來取得較佳的描述。他們利用這兩個特徵從一個動作影像序列中擷取出一些代表姿勢的序列，因此一個動作序列可以被描述成一個符號字串。所以兩個不同動作序列的比對便可視為兩個不同字串比對的問題。Mukherjee 等人〔Muk 11〕利用關鍵姿勢（key pose）為基礎之 bag-of-word 方法來進行人類動作的辨識。首先，他們結合姿勢的資訊與運動的形態（motion pattern）來描述人類的姿勢。從每一種動作類別中找出所有的關鍵姿勢並形成一個編碼書（codebook），因此每一關鍵姿勢即代表編碼書中的每一個字彙。將每一個動作中的所有視訊框對映到某個關鍵姿勢，並對所有的姿勢字彙建立一個長方條統計圖（histogram）來描述每一個動作。最後再利用支持向量機（Support Vector Machine, SVM）來做動作的訓練與分類。

使用 2D 人類姿勢的方法雖然較為簡單，但受到視角的影響，可能無法處理太複雜的動作。因此有不少的研究學者便提出利用 3D 人類軀幹模型來分析人類的動作。例如，Ohya 〔Ohy 02〕 提出一個視覺方法來估測人類的姿勢及辨識人類與其它物體間的交互作用。他從多台攝影機所取得的影像序列中分別利用 heuristic 及非 heuristic 的方法來估測人類 3D 的姿勢。在 heuristic 的方法中，他

利用分析人類姿勢的輪廓來找出每張影像中人類肢體上較具代表性的點。而非 heuristic 的方法主要是利用一條曲線函數來分析姿勢輪廓。最後，他利用一個以函數為基礎的輪廓分析及一個以動作向量為基礎的分析來辨識相互作用，以至於系統可以判斷人體是否與物體有交互作用。在〔Bou 03〕中，Boulay 等人提出一個新的方法來對影像序列作人類姿勢的辨識。首先他們將移動像素映射在一個參考軸上，並且透過 PCA 來學習 2D 的肢體外形。然後，他們利用一個人類軀幹的 3D 模組來描述人類姿勢,使得以映射為基礎的方法與攝影機位置互不相關。在〔Wer 05〕中，Werghi 提出一個以模組為基礎的方法從 3D 掃描器取得的資料中辨識人類姿勢。為了找到一個對人類姿勢具有高區別性的表示方法，他發展一個名為小波轉換係數（wavelet transform coefficients, WC）的 3D 外形描述方法。這些人類姿勢的特徵可以視為是〔She 99〕中所提的 2D 小波外形描述方法的延伸。最後，他利用一個 Bayesian 分類架構來比較 WC 與其它 3D 外形描述的方法。Gu 等人〔Gu 10〕 利用取得的 3D 人類關節資訊來進行人類動作與走路步伐的辨識。首先， 他們利用非標誌姿勢重建的方法，從多台攝影機所重建的 3D 體積資料（volume data）中取得 3D 人類關節資訊。接著再從 3D 人類關節資訊中擷取出移動資訊特徵（位置、方向及身體的高度）與形態特徵（configuration feature）。然後分別再利用隱藏式馬可夫模組（hidden Markov model, HMM）與以範例為基礎（exemplar-based）之隱藏式馬可夫模組（EHMM）

來模組化所取得的移動資訊特徵與形態特徵。最後，一個階層式的最大後驗法（maximum a posteriori, MAP）被用來進行人類動作與走路步伐的辨識。在上述 3D 方法中，取得人類 3D 的資訊大部份都是利用動作捕捉技術（motion capture）或多台經校正後的攝影機來取得，所以成本或複雜度較高。最近微軟推出平價的 Kinect 後，讓研究者可以利用較低的成本便可取得人類關節的 3D 資訊，因此最近有不少研究學者便利用 Kinect 來進行人類動作的辨識。例如在〔Lin 11〕，作者利用 3D 的人類關節資訊來進行人類動作的辨識。首先，他們利用 Kinect 取得 3D 的人類關節資訊，接著再利用所取得 3D 人類關節位置的分佈來產生動作圖表（action graph），目的是要對每個動作進行編碼。在動作圖表中，每個節點代表一個顯著的姿勢，節點間的權重則代表兩個姿勢間的轉換機率。最後再利用每個動作的動作圖表來辨識所觀察到的動作。

在人類動作分析中較高層的人類活動辨識部份，目前也存在一篇概述的論文〔Tur 08〕。要進行複雜活動的內在結構與語意的分析，較高等級的表示法與推論的方法是必須的，這些方法可被分成三類：圖形方法（graphical models）〔 Ham 07〕、句法的方法（syntactic approach）〔 Ryo 06〕及以知識或推理為基礎的方法（knowledge and logic-based approach）〔Tra 08〕。在〔Ham 07〕中，作者利用字尾樹（suffix tree）演算法來進行人類活動的分析。人類的活動可以被視為一連續的事件序列組合而成。由於事件前後因果關係與人類活動特殊的限制，這些連

續的事件序列中會存在著一些特定先後順序關係的事件子序列。利用這些限制，一個活動可以被表示成一群事件子序列，而這些事件子序列可利用字尾樹的方法取得。最後，他們再利用字尾樹的特性，來偵測活動中不正常的事件子序列。Ryoo 與 Aggarwal〔Ryo 06〕利用上下文無關語言（context-free grammar, CFG）的表示法來辨識混合的複雜動作及人與人之間的互動動作。他們利用階層式的辨識方法，分成身體軀幹擷取層（the body-part extraction layer）、姿勢辨識層（the pose layer）、手勢辨識層（the gesture layer）及動作與互動辨識層（the action and interaction layer）。首先從影像序列中取出身體軀幹的特徵出來，接著再利用這些特徵來估測姿勢與手勢動作，最後利用語意表示方法來辨識動作與互動。Tran 與 Davis〔Tra 08〕 利用馬可夫邏輯網路（Markov Logic Network, MLN）來對視訊監控系統進行事件的模組化與辨識。他們結合邏輯與機率的模組來解決觀察資料不確定性的問題，其中每一個邏輯的規則皆是利用一階邏輯式子。而所有規則由其對應的權重結合在一起，每個權重所代表即為該規則正確性的信賴度。最後再利用 MLN 對輸入的查詢事件進行推論。

1.4、實驗器材：

本研究利用 Kinect（圖 1-7）來捕捉人體骨架資訊〔Li 12〕。Kinect 為微軟（Microsoft）公司所開發給 Xbox 360 遊戲主機的一個 RGB-D Sensors，讓玩家

可以利用手勢去操作系統介面，而不用透過搖桿。Kinect 甚至可以捕捉玩家骨架資訊（圖 1–8）的功能，使玩家可以利用肢體來體驗遊戲（圖 1–9），進而達到運動的效果，所有的關節位置與名稱如（圖 1–10）（微軟提供的 SDK）。

圖 1–7：Kinect（來源： Microsoft Developer Network ）

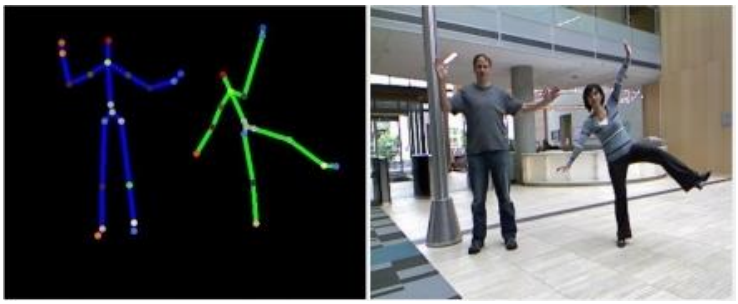

圖 1–8： Kinect 擷取的骨架資訊（來源： 小 7 俱樂部 ）

圖 1-9： 玩家利用肢體來體驗遊戲（來源：Xbox）

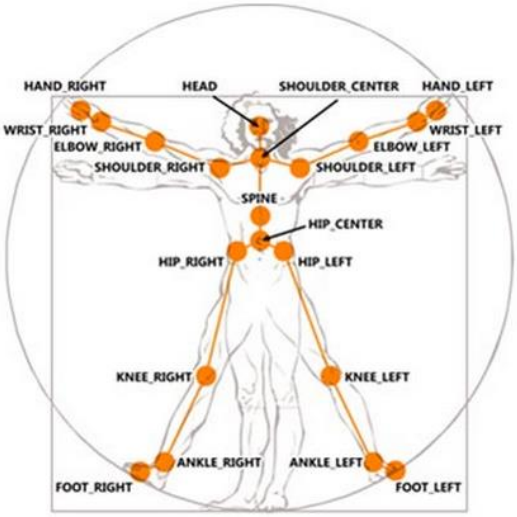

圖 1-10： 關節位置與名稱（來源：msdn）

Kinect 感應器可以得到三種資訊：彩色影像（圖 1-11）、3D 深度影像（圖 1-11）及聲音。彩色影像透由中間的 RGB 鏡頭取得、3D 深度影像由左右兩邊的紅外線發射器和紅外線 CMOS 攝影機產生，而聲音是透過陣列式麥克風取得（圖 1-12）。

圖 1–11： （左）深度影像、（中）骨架資訊、（右）彩色影像（來源： Kinect for Windoes）

圖 1–12：Kinect 外部構造（來源：msdn）

Kinect 的深度影像是利用近紅外線對所要測量的空間進行編碼，然後再透過感應器讀取編碼的光線，最後交給晶片解碼後產生一張深度影像，稱之為光編碼（Light Coding）。而雷射光散斑（Laser Speckle）是光編碼其中的關鍵。當雷射光照射到物體或穿透之後，便會形成隨機的反射斑點（稱之為散斑）。空間中任

意兩處的散斑圖案並不會相同，是因為散斑的圖案會隨著距離的不同而改變，也就是將整個空間註解上標記，因此任何物體進入此空間時或者在此空間移動時，都可以紀錄其位置。

Kinect 利用紅外線發出 class 1 雷射光（肉眼看不見），將雷射光透過鏡頭前的散光片均勻的投射在測量的空間（圖 1-13），而後透過 CMOS 紅外線攝影機記錄空間中全部的散班，最後再經由晶片運算後形成三維的深度影像。

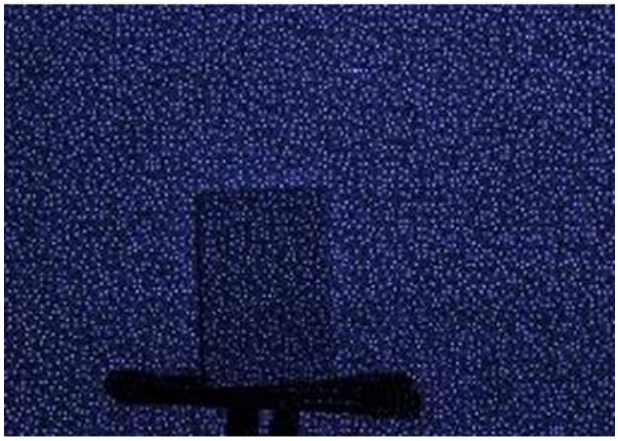

圖 1-13： Kinect 發出的散斑（來源：javaforge. com）

1. 5、文章架構：

接下來將在第二章介紹系統架構以及實驗環境。而三、四、五章將會介紹系統中的各個細節。如第三章將介紹如何利用微軟所開發的 Kinect 來擷取特徵；第四章會介紹如何對基本動作進行分割及分群；第五章會介紹如何將分群完的基本

動作進行動作的分類。第六章將展示本研究的實驗結果，而在第七章為總結並討論未來可能的研究方向。

第二章　　系統架構

本章節將敘述本研究的系統環境及系統架構與流程。內容主要分成兩部分：2. 1 節將描述系統的環境架設，亦即 Kinect 的硬體特性；2. 2 節將針對本系統的流程做簡單的說明。

2.1、系統環境：

如圖 2–1 所示，Kinect 垂直視角為 43 度，上下各有 27 度的傾斜角度，因此 Kinect 在垂直方向上，可視範圍共有 27+43+27=97 度。Kinect 的水平視角為 57 度，即以感應器為中心，左右各 28. 5 度的可視範圍，如圖 2–2 所示。

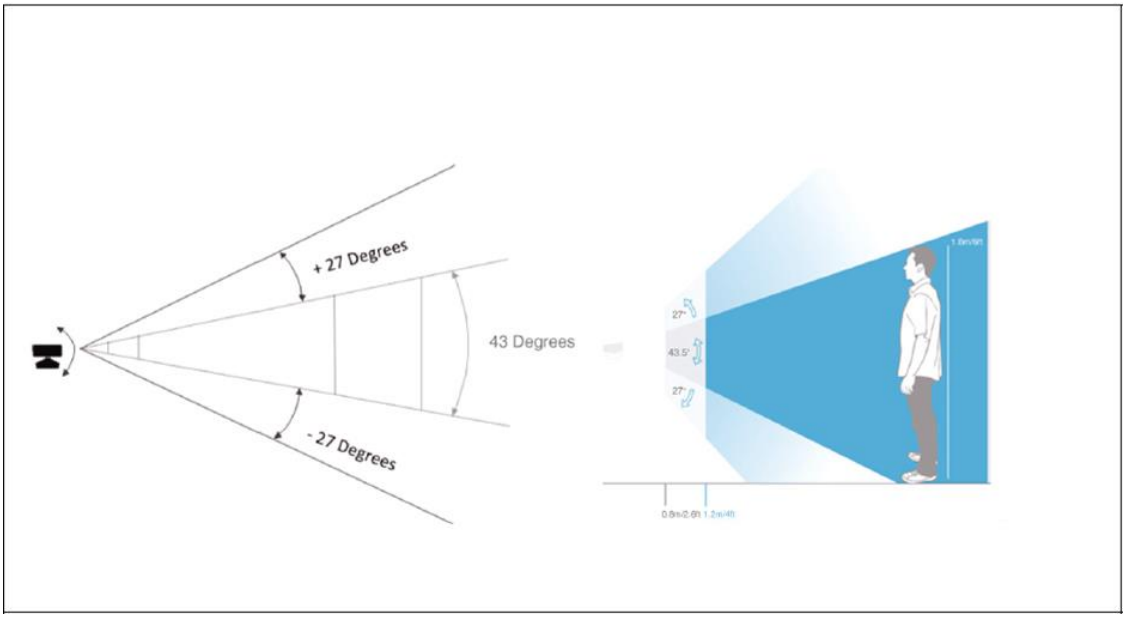

圖 2-1:Kinect 垂直視角(來源:Microsoft)

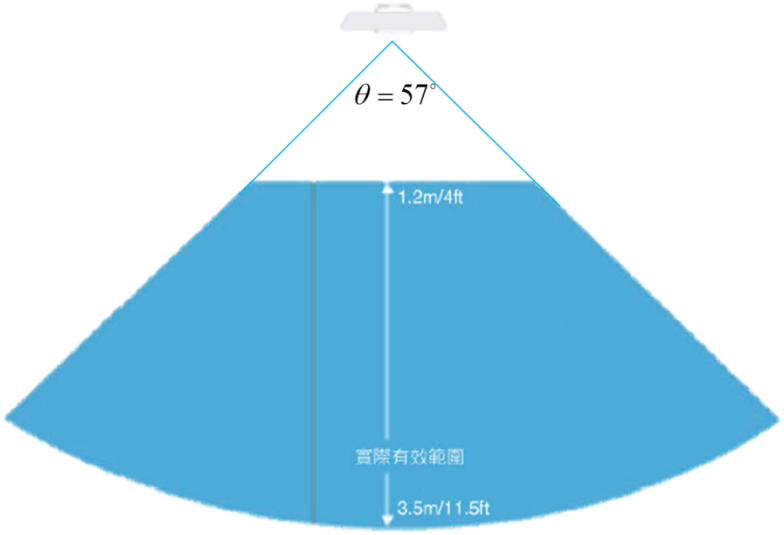

圖 2-2:Kinect 水平視角

2.2、系統流程：

本系統主要分成二個階段，如圖2-3所示，分別為訓練過程及測試過程。第一個階段為訓練過程，首先請一位訓練者演示所有訓練的動作，利用Kinect取像並取得訓練者姿勢的三維關節資訊，如圖2-4所示，因此我們可以得到所有動作的三維關節資訊的特徵序列。接著我們利用等構映圖(isometric feature mapping, Isomap)的技術將所得到三維關節資訊的特徵序列轉換至低維度的等構映圖空間(Isomap space)，在此空間中不同的動作將形成不同的流形軌跡(manifold trajectory)，如圖2-5所示。為了有效地找出連續兩個基本動作軌跡中的中斷點，我們將Isomap空間中的流形軌跡描述為低維度的時間序列。接著再利用時間分割的技術將此時間序列分割成許多子序列，每個子序列都代表一個基本動作。然後我們再利用分群(clustering)的技術對這些基本動作序列分群。每一群的基本動作將被給予一個符號表示，並形成一個編碼書，如圖2-6所示。

第二階段為測試過程，測試者同樣利用Kinect取像並取得測試者姿勢的三維關節資訊後，為了節省重新生成新Isomap空間的時間成本，我們利用Isomap mapping的方法，將所有測試者的動作姿勢映射到現有的Isomap空間中，接著同樣利用時間分割的技術將所有測試者動作切割成一個個的基本動作，最後便可利用基於動態時間扭曲技術(Dynamic Time Warping, DTW)之最近鄰(nearest-neighbor)的比對方法進行基本動作的分類。

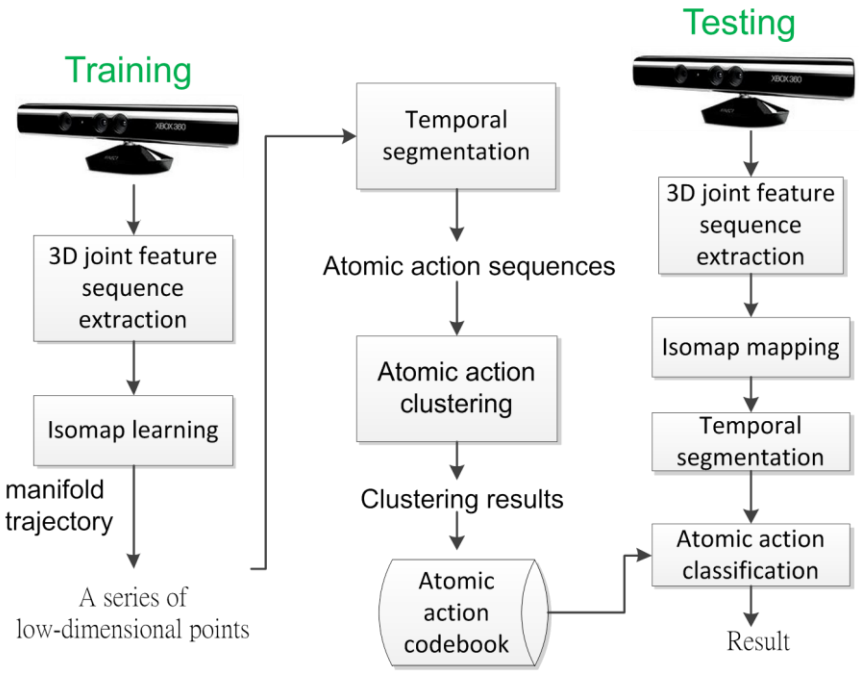

圖 2-3: 系統流程圖

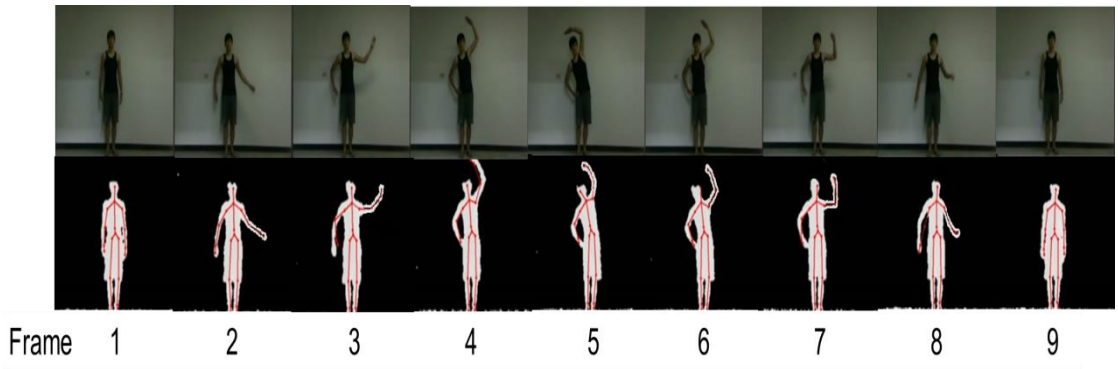

圖 2-4: 利用 Kinect 取得訓練者姿勢的三維關節資訊

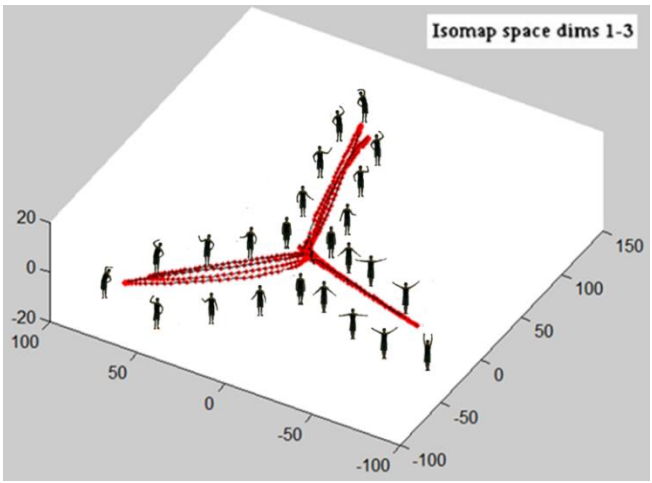

圖 2-5:6 種不同基本動作的流形軌跡

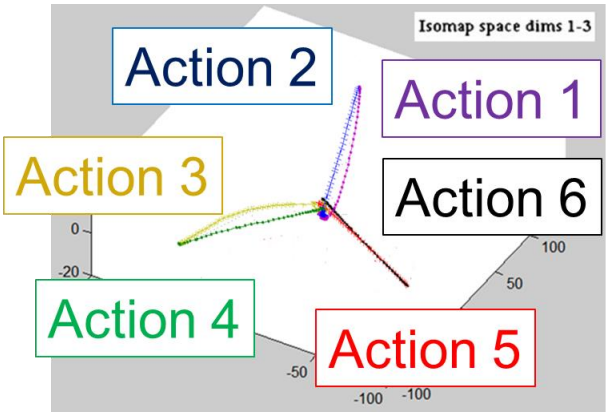

圖 2-6:6 種基本動作的分群

第三章　三維關節特徵序列的擷取

此步驟主要目的是要從輸入的影像序列中，取出人類所有關節的三維特徵資訊。在此我們利用微軟所開發的 RGB–D sensor（亦即 Kinect）可以取得一張彩色影像及一張深度影像，再利用微軟提供的 SDK 便可取得人類 20 個關節的三維座標，如圖 1-10 所示。因此，對每一張輸入的影像，皆可取得 20 個關節座標來表示的人類姿勢。

為了不受身體所在的位置、方向、及使用者姿態差異的影響，我們從 20 個人類關節座標中計算出 19 個相鄰關節的向量（圖 3–1），再由這 19 個向量計算出關節角度向量 x_t（圖 3–2），而此關節角度向量x_t包含了人體 20 個關節夾角。兩個三維向量的夾角計算式如下：給定兩向量$\mathrm{u} = (u_1, u_2, u_3)$、$\mathrm{v} = (v_1, v_2, v_3)$，兩向量內積為$\mathrm{u}\cdot\mathrm{v} = (u_1v_1 + u_2v_2 + u_3v_3)$。假設 θ 為兩個非零向量的夾角，則 $\theta=\cos^{-1}(\frac{\mathrm{u}\cdot\mathrm{v}}{\|\mathrm{u}\|\,\|\mathrm{v}\|})$。

由於 Kinect 所擷取的三維人體骨架資訊相當敏感，所得的深度值並不是很穩定，即使人體靜止不動，深度值也會不斷的大幅度跳動，這對後續計算關節之間的夾角會有很大的影響。在正常情形下，連續畫面之間的深度值之差值並不會太大，即使人體做出深度值會改變的動作時，連續畫面間的深度值差值也會落在一定的範圍內，加上我們的應用不需要太精細的深度值，所以我們先將深度值分成

數個相等的區段（bin），座落於同一個區段的深度值皆給予相同的值，以達到降低深度值的敏感度。另外，我們再設一個門檻值，若當下畫面的深度值和先前畫面深度值之間的差值超過此門檻值時，則使用前 4 個畫面之間的差值之平均值來預測下一個深度資訊。

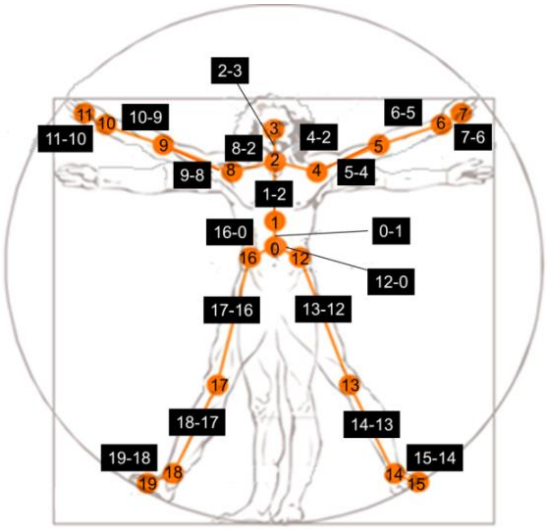

圖 3-1：根據骨架位置計算出來的 19 個向量

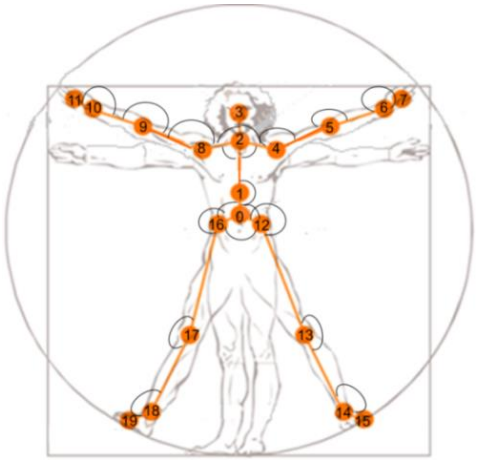

圖 3-2：根據 19 個向量所計算出來的 20 個夾角

對於每一個人類姿勢，我們定義一個對應的特徵向量$X_t = (x_t, \dot{x}_t)$，包含人類關節的角度向量x_t 及其一次微分值$\dot{x}_t$。我們同時以關節角度向量及一次微分值來代表人類姿勢的特徵，除了可以避免上述的身體全域位置、方向與使用者姿態差異影響外，也可以避免姿態模稜兩可（ambiguity）的情形。因此人類活動可以被視為一個在上述特徵空間中的運動軌跡（motion trajecotry），經過取樣後得到特徵向量X_t 序列。這個特徵向量序列便是代表人類姿勢在時間上的變化，亦即人類的動作，我們將它表示成$\{X_1, X_2, \ldots, X_t\}$（圖 3-3）。

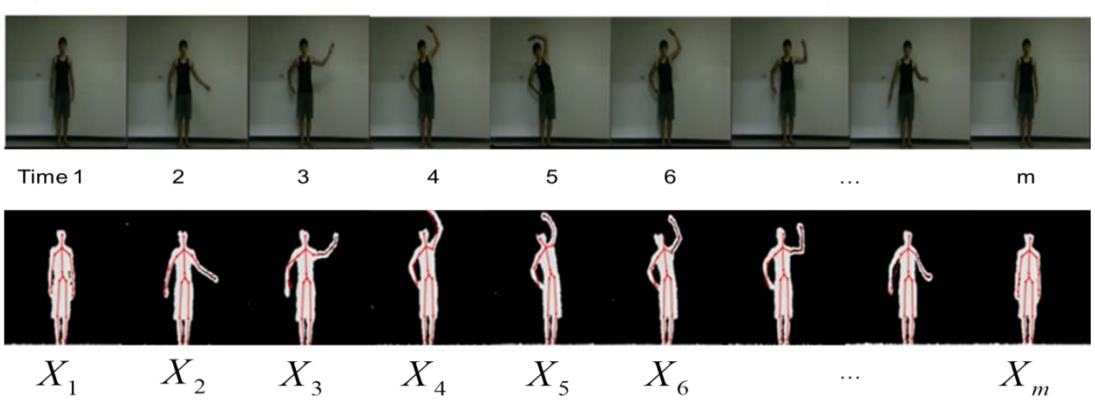

圖 3-3： 人類姿勢變化由特徵向量序列表示

第四章　基本動作的切割與分群

由於我們利用人類關節的角度向量及其一次微分值來表示人類的姿勢，其維度相當的高。因此要在高維度的特徵空間中進行動作的切割、分群及辨識是相當地複雜。一般而言，人類動作中相臨連續姿勢的變化不太，其連續姿勢序列則可視為是被嵌入在高維度空間中之低維度流形上的變化。因此本研究將利用流形學習（Manifold learning）方法中的 Isomap 技術來進行人類基本動作的切割與分群。

4.1、Isomap 演算法：

Isomap 目的是要找到一個低維度的歐氏空間，使其最能保持原有的高維空間中任意兩個資料點之間的測地線距離（geodesic distance），而測地線則為流形上連接兩個點的最短曲線，測地線距離則為此測地線的長度。由於流形學習假設所有資料點中存在一個低維的結構，沿曲線之低維結構的測地線距離是比較合適測量任何兩個資料點之間的距離，而不是在高維空間中的歐幾里得距離（Euclidean distance）。因此，Isomap 的方法是利用連接相鄰點的鄰域圖中的最短路徑來估算測地線距離。這種用測地線距離代替傳統的歐幾里得距離的算法，可更有效地在

低維空間表達高維空間的資料，減少降維後損失掉的資料訊息，保留在高維度空間裡所隱含的重要性質。Isomap 演算法包括三個步驟：

(1) 建構鄰域圖（neighborhood graph）：

Isomap 的輸入是許多高維度的資料，並把它們當作一個圖(graph)，只要兩個點（vertex）是鄰居，就會有一條邊連結（edge），至於判定鄰居的方法有兩種，一種是找 k 個最鄰近的點（K-nearest neighbors），或者讓資料點座落在一個半徑為 ε 的超球面上。

(2) 計算出任兩點之間的最短路徑：

估算鄰域圖任意兩個節點之間的測地線距離是透過計算它們在鄰近圖上的最短路徑。

(3) 建立一個維度為 d 的空間：

利用 multidimensional scaling（MDS）技術把資料降到維度為 d 的空間。

在〔Ten 12〕中，圖 4-1 中資料分佈在瑞士捲（swiss roll）上，兩點間的歐幾里得距離(虛線)並不能代表兩點的實際距離。分佈在流形面上的曲線長度代表兩點間的測地線距離。若流形的結構事先不知道時，Isomap 可以透過將資料點連接起來構成一個鄰接 Graph 來離散地近似原來的流形，而測地距離也可透過 Graph 上的最短路徑來近似。圖 4-2 及圖 4-3 顯示 Isomap 降維度後的測地線（紅線）及近似測地線的直線(藍線)的映射結果。

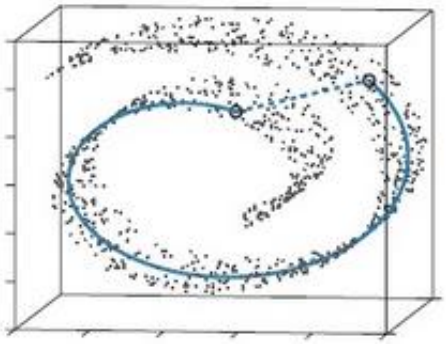

圖 4-1：資料分佈在瑞士捲上，兩點間的歐幾里得距離(虛線)並不能代表兩點的實際距離（*資料來源：*〔Ten 12〕）

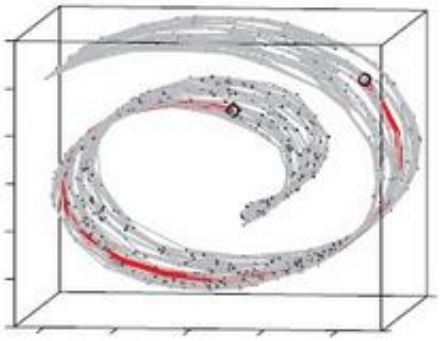

圖 4-2：Isomap 降維度後的測地線（紅線）（*資料來源：*〔Ten 12〕）

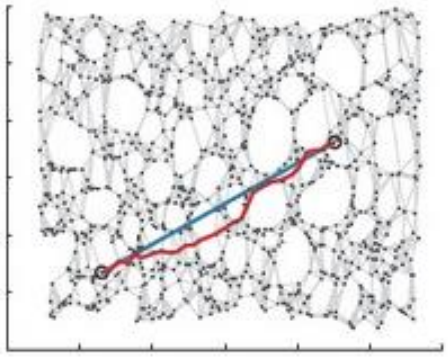

圖 4-3：近似測地線的直線(藍線)（*資料來源：*〔Ten 12〕）

在 Isomap 演算法中有件很重要的事情，就是如何決定 d 為多少維度的 Isomap 空間。剩餘方差（residual variance），R_d，可被定義用來評估降維後的資料和原本高維度資料的誤差：

$$R_d = 1 - r^2(G, D_d) \qquad (4.1)$$

G 代表測地線距離矩陣（*geodesic distance matrix*）；D_d 代表在 d 維空間上的歐幾里得距離矩陣（*Euclidean distance matrix*）；$r(G, D_d)$ 代表 G 和 D_d 的相關係數（correlation coefficient）。當剩餘方差達到收斂時，此時維度即為我們要取的 d 值。如圖 4-4 所示，維度為 4 時，剩餘方差已經達到收斂，圖 4-5 和圖 4-6 則顯示人類的動作序列經 Isomap 降維後的 4 維空間圖。

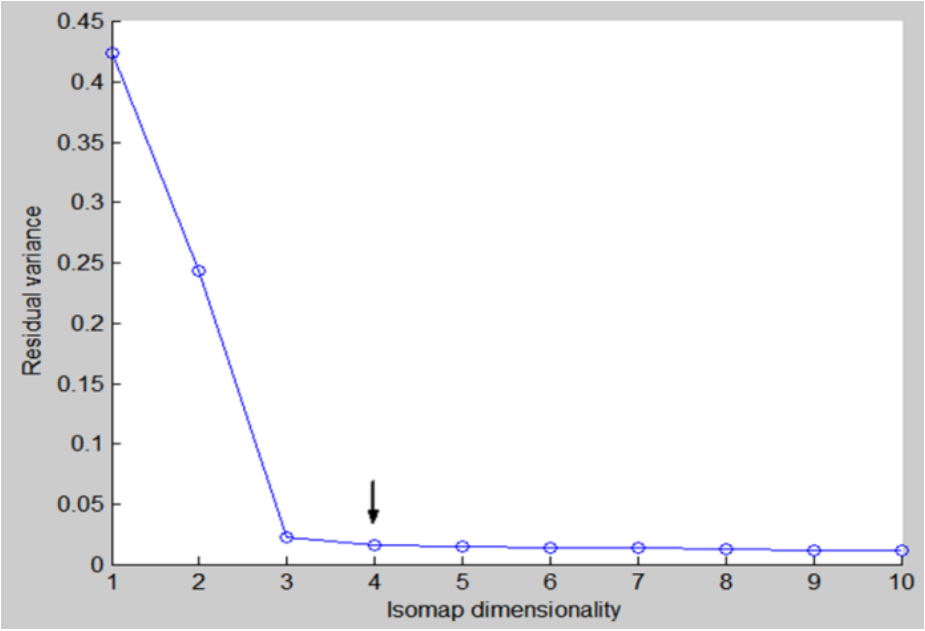

圖 4-4：Isomap 訓練資料的剩餘方差

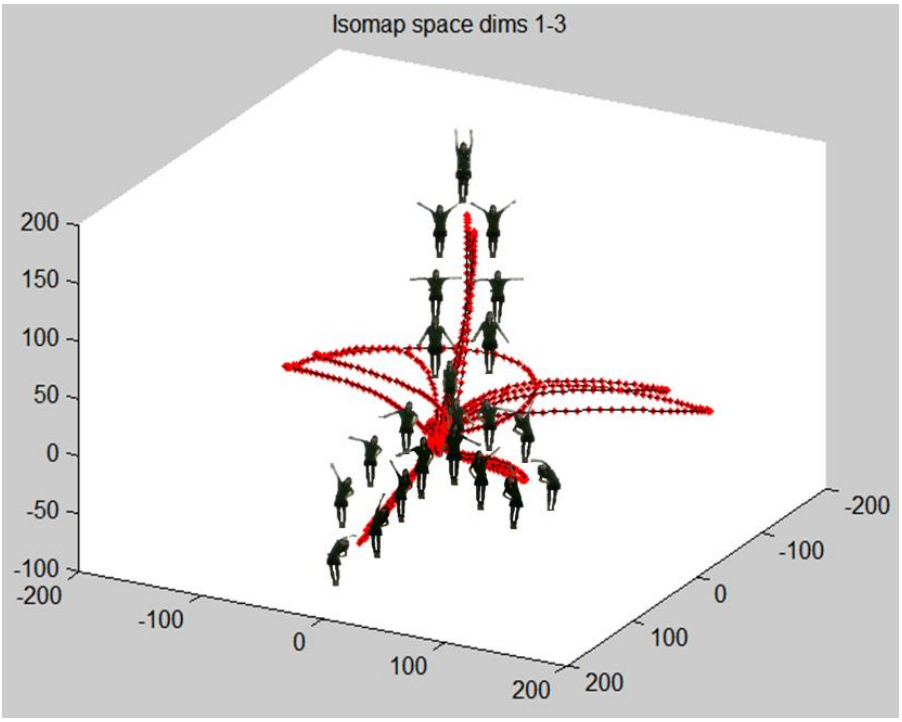

圖 4-5：動作序列（10 種基本動作）manifold 軌跡映射在 1 到 3 維的 Isomap 空間

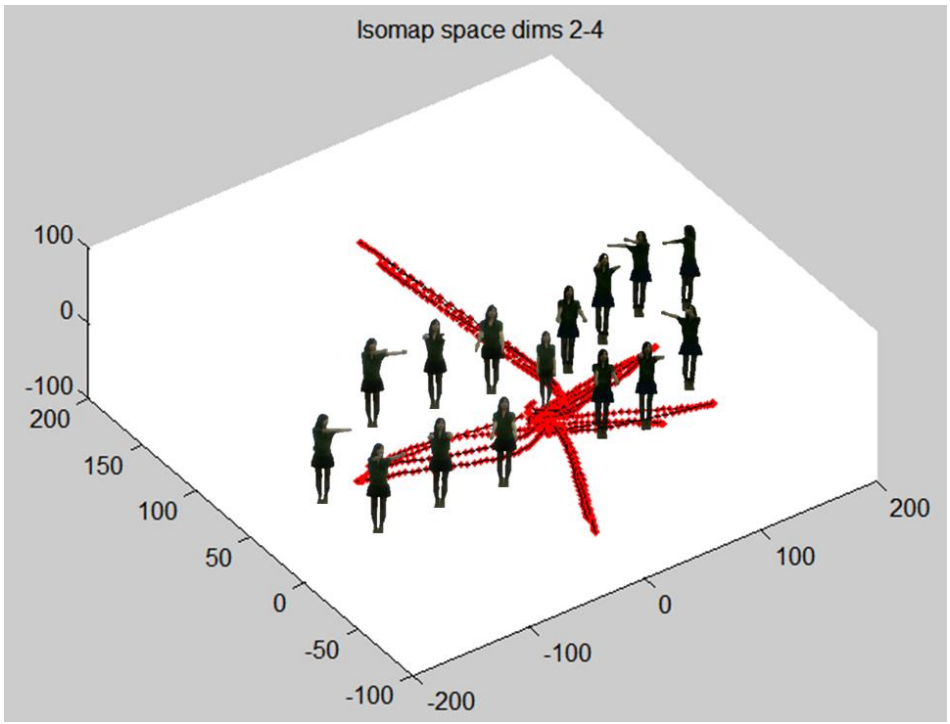

圖 4-6：動作序列（10 種基本動作）manifold 軌跡映射在 2 到 4 維的 Isomap 空間

接下來將探討 MDS 如何去建構一個 d 維度的空間。

4.1.1、Multidimensional scaling 演算法：

MDS 是一種降維方法，它的目的是使降維之後的點兩兩之間的歐幾里得距離盡量保持不變（也就是和在原本高維空間中對應的兩點之間距離差不多）。G $= [g_{ij}] \in R^{n\times n}$，$g_{ij}$代表 i和j的距離。MDS 建構了在 d 維的歐幾里得空間中 n 個點使得 n 個點兩兩之間的距離與G中的距離最接近。令$x_i = (x_{i1}, \dots, x_{id})^T$代表 Isomap 的歐幾里得空間中第$i$個點的座標。第$i$個和第$j$個點之間的歐幾里得距離，可以計算如下：

$$d_{ij}{}^2 = (x_i - x_j)^T(x_i - x_j) = x_i{}^T x_i + x_j{}^T x_j - 2x_i{}^T x_j \qquad (4.2)$$

為了解決任意位移所造結果的不確定性，我們假設：

$$\sum_{i=1}^{n} x_i = 0 \qquad (4.3)$$

由方程（4.2）和（4.3）中，x_i和x_j之間的內積可以推導如下：

$$b_{ij} = x_i{}^T x_j = -\frac{1}{2}(d_{ij}{}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}{}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}{}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}{}^2) \qquad (4.4)$$

令D $= [d_{ij}]$代表在 Isomap 空間中計算的距離矩陣，由於 Isomap 空間是使 D 接近 G 來決定，內積矩陣B $= [b_{ij}]$可以透過以下方式獲得

$$B = -\frac{1}{2}HGH \qquad (4.5)$$

其中$\mathrm{H} = \mathrm{I} - \frac{1}{\mathrm{n}}11^{T}$，而$1 = [1, \ldots, 1]^{T}$為一個包含 n 個 1 的向量。令$\mathrm{X} = [x_1, \ldots, x_n]^{T}$為 n 個點在 Isomap 空間中未知座標的$n \times d$矩陣。內積矩陣可以表示為$\mathrm{B} = \mathrm{X}X^{T}$。由於 B 可被計算出來，所以我們可以利用 B 去計算 X。首先將 B 分解成$\mathrm{V}\Lambda\mathrm{V}^{\mathrm{T}}$，而 $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ 為對角矩陣的特徵值，而 $\mathrm{V} = [v_1, v_2, \ldots, v_n]$ 是對應的特徵向量矩陣。矩陣 X 的坐標可以如下：

$$\mathrm{X} = \mathrm{V}'\Lambda'^{\frac{1}{2}} \qquad (4.6)$$

其中$\Lambda'^{\frac{1}{2}} = diag(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \ldots, \lambda_d^{\frac{1}{2}})$，$V' = [v_1, v_2, \ldots, v_d]$。

4.2、基本動作分割：

基本動作分割的目的是要找出合適的斷點將連續動作序列分割成許多基本動作。從圖 4-5 和圖 4-6 可知。在 Isomap 空間中我們可以很明顯的區分出不同的基本動作有著各自的流形軌跡，所以在分割過程中如何去判斷兩個連續基本動作軌跡的斷點便非常的重要。由於 Isomap 空間並沒有顯示出時間的資訊，所以很難找出斷點並分割出連續的基本動作，為了方便找出斷點，我們將流形軌跡轉換成 d 維資料點的時間序列，再計算每個資料點的 magnitude（即與原點的 L_2 norm 值），如圖 4-7 所示。

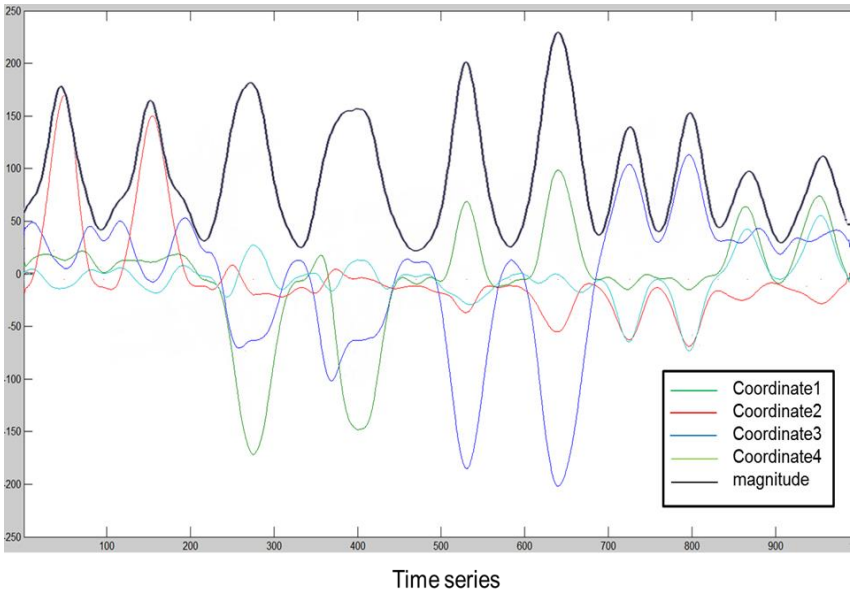

圖 4-7：利用 Gaussian smoothing 得到的資料點時間序列及相對應的 magnitude（10 種基本動作）

在一般情況下，一個人的動作會在基本動作的邊界處緩慢下來，因此，在 magnitude序列中的局部最小值（local minima）和局部極大值（local maxima）可以被視為候選的斷點，如圖4-8所示。

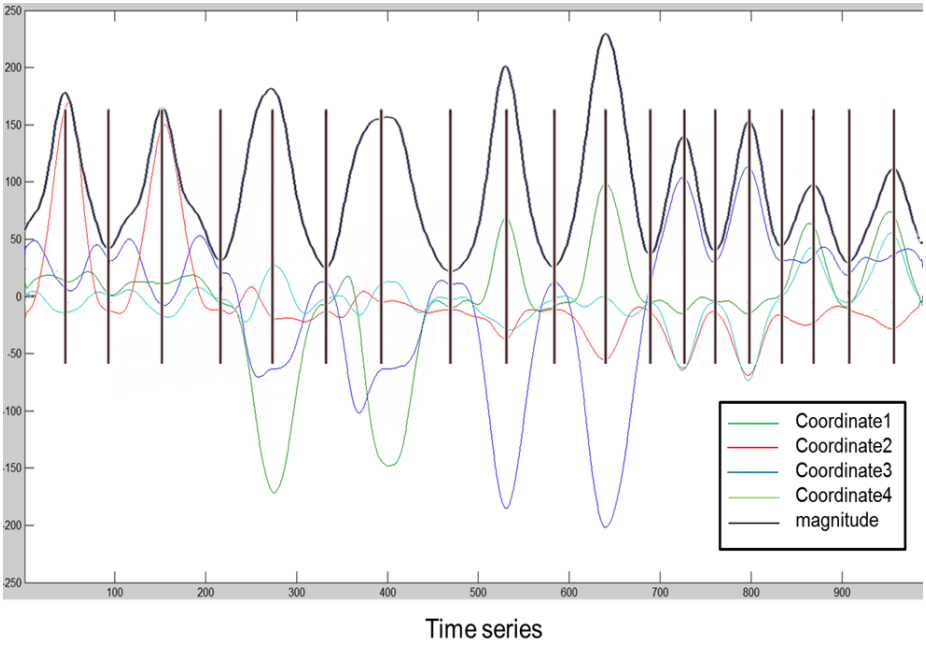

圖 4-8：在 local minima 和 local maxima 處分割基本動作（10 種基本動作）

4.3、基本動作分群：

在分割訓練的動作序列後，我們可以對被分割完的基本動作進行分群，並利用分群結果來模組化每個基本動作群。由於各分段動作的持續時間不一定相同，所以DTW可以提供一個近似的相似性度量（similarity measurement）。DTW利用在兩個資料序列間找出最佳的warping path來對齊並比較兩個資料序列。舉例來說，我們有兩個序列A $= [a_1, a_2, \dots, a_m]$和B $= [b_1, b_2, \dots, b_n]$，而a_i和b_j是在Isomap空間中的兩個d維向量。W $= (w(1), w(2), \dots, w(m))$是一個用來對齊A和B的warping path，而$j = w(i)$代表a_i和b_j對齊。計算warping path有以下三個條件：$1 = w(1)$，$n = w(m)$和$w(k+1) \geq w(k)$。沿著warping path W，A和B之間的距離可以被計

算如下：

$$D_W(A,B) = \frac{1}{m}\sum_{i=1}^{m}\left\|a_i - b_{w(i)}\right\| \quad (4.6)$$

DTW的目的是要找到warping path去最小化距離$D_W(A,B)$。因此，A和B之間的DTW匹配分數可以由下式計算：

$$DTW(A,B) = min_w\{D_W(A,B)\} \quad (4.7)$$

由於DTW 匹配分數在定義上並不是對稱的，所以我們定義兩個分割好的動作序列間之DTW動作距離（pairwise action distances）為：

$$AD(A,B) = \frac{1}{2}(DTW(A,B) + DTW(B,A)) \quad (4.8)$$

計算完所有分割後動作之間的動作距離後，我們用階層式分群法（hierarchical clustering algorithm）把相似的基本動作分在同一群。階層式分群法由樹狀結構的底部開始層層聚合（圖4-9），一開始將每一筆資料視為一個群集（cluster），假設我們現在擁有n筆資料，則將這n筆資料視為n個群集，亦即每個群集包含一筆資料，演算法如下：

(1) 將每筆資料視為一個群集 C_i, i=1 to n。

(2) 找出所有群聚間，距離最接近的想個群集C_i、C_j。

(3) 合併 C_i、C_j成為一個新的群集。

我們設定一個距離門檻值，當群內的距離超過某一個門檻值就不再合併成新的群集，而門檻值是透過多次的實驗所得到的結果。

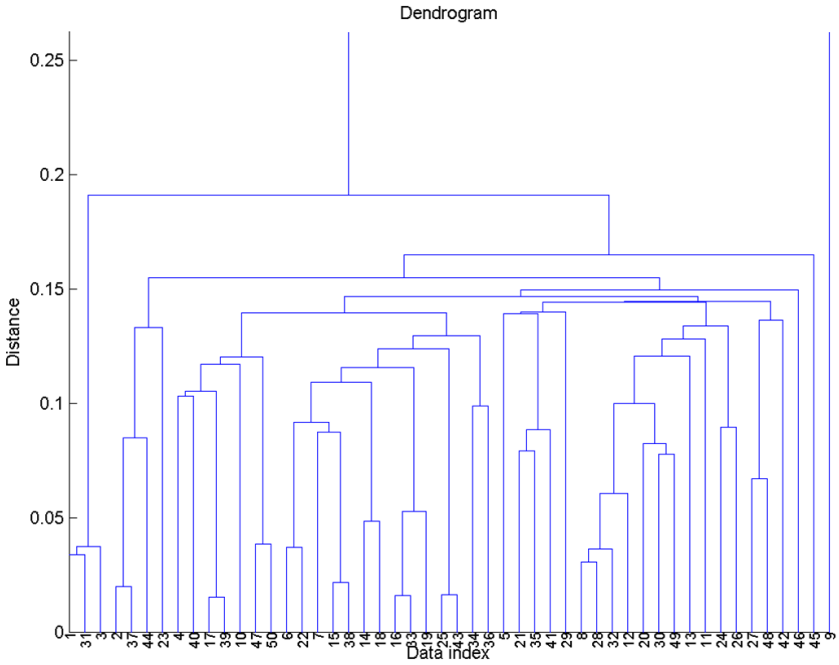

圖 4-9：階層式分群法示意圖（資料來源：http://neural.cs.nthu.edu.tw）

第五章　基本動作的學習及分類

在此步驟中，我們利用每個群集的平均軌跡（mean trajectory）當作模板(template)來代表每一群集，如圖 5-1 及 5-2 所示。在此，平均軌跡是指在同一個群集中的所有分割動作的平均時間序列。由於各分段動作序列的持續時間不同，我們採用內插法（interpolation method）正規化它們的持續時間。最後我們利用基本動作的模板來分類一個新輸入的動作，此分類的方法是利用最近鄰方法（nearest-neighbor approach）及 DTW 距離。

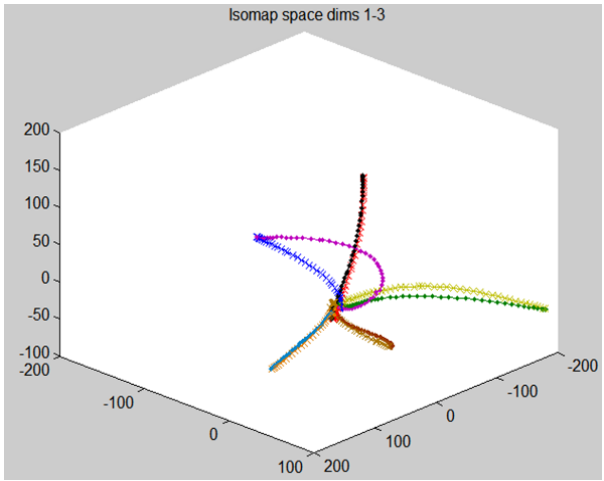

圖 5-1：10 個基本動作群集（10 種顏色）在 1 到 3 維 Isomap 空間

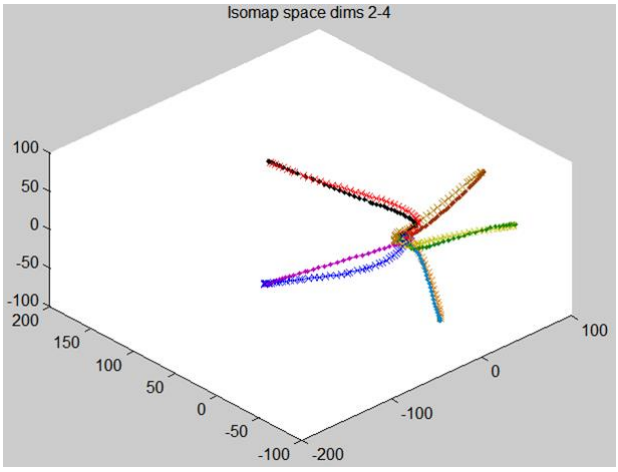

圖 5-2：10 個基本動作群集（10 種顏色）在 2 到 4 維 Isomap 空間

為了辨識出一個新輸入的動作序列，我們需要把新的序列映射到 Isomap 的空間。由於 Isomap 空間是由訓練的資料所構成，為了將新的資料點映射到 Isomap 空間上，鄰域圖和測地線距離都必須重新計算，然後再利用 MDS 生成一個新的 Isomap 的空間。然而加入新的資料去重建 Isomap 是非常耗時間的，特別是當新輸入的資料序列非常大的時候。為了解決這個問題，Law 和 Jain〔Jai 06〕提出增量 Isomap 的學習方法（incremental Isomap learning method），這種方法僅更新部分和新樣本有關的點的相鄰圖（adjacent graph）和測地距離，以便確定新樣本的座標，然後再更新所有點的座標。

雖然這種方法可以讓計算時間減少，但它仍是一個耗時的過程。在這項工作中，我們提出了另一種方式來處理這個問題。我們發現，除非有新的動作模組加入，不然沒有必要為了基本動作的辨識建立一個新的 Isomap 空間。因此在〔Jai 06〕

中，簡單地估算一個新的樣本映射到現有 Isomap 空間之新資料點的座標。假設訓練的資料點的座標為$[x_1, x_2, \dots, x_n]$，新的資料點的座標向量表示成x_{n+1}，在方程式（4.2）的d_{ij}中i = 1, … , n + 1，j = 1, … , n + 1可導出：

$$x_{n+1}{}^T x_{n+1} = \frac{1}{n}\sum_{i=1}^{n} d_{i(n+1)}^2 - \frac{1}{n}\sum_{i=1}^{n} x_i{}^T x_i \quad (5.1)$$

x_{n+1}和x_j的內積可以用（4.2）計算出如下：

$$x_{n+1}{}^T x_j = -\frac{1}{2}\left(d_{(n+1)}^2 - x_{n+1}{}^T x_{n+1} - x_j{}^T x_j\right) \quad (5.2)$$

從方程式（4.2）、（5.1）和（5.2），$x_{n+1}{}^T x_j$可以被計算出來：

$$x_{n+1}{}^T x_j = -\frac{1}{2}\left(d_{(n+1)j}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{(n+1)}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) = f_j \quad (5.3)$$

因此，

$$[x_1, \dots, x_n]^T x_{n+1} = \mathrm{f} = [f_1, \dots, f_n]^T \quad (5.4)$$

將方程式（4.6）代入（5.4）可得：

$$\left[\sqrt{\lambda_1}v_1, \dots, \sqrt{\lambda_d}v_d\right]^T x_{n+1} = f \quad (5.5)$$

然後x_{n+1}的 least square solution 可以被推導如下：

$$x_{n+1} = \left(\frac{1}{\sqrt{\lambda_1}} x_1{}^T f, \dots, \frac{1}{\sqrt{\lambda_d}} x_d{}^T f\right)^T \quad (5.6)$$

在測試動作序列的每一個人類姿勢都被映射到 Isomap 空間後，可以利用基於 DTW 距離的最近鄰方法把測試序列分類到某個群集。我們利用了一個門檻值去判斷一個未知的動作序列屬於某一個已經學習過的群集，如果未知的動作序列和最接近的平均序列(mean sequence)之間的距離大於門檻值，則未知的動作會被認

為是一個新的動作。而門檻值的設定為$\mu_k + 2\sigma_k$，μ_k, σ_k分別為訓練動作序列和第k個群集模範之間距離的平均值與標準差(standard deviation)。

雖然不去建立一個新的 Isomap 空間有達到省時的效果，但方程式(5.3)中，我們還是要去計算所有測試的人類姿勢和已經訓練好的人類姿勢特徵向量(第三章所提到的X_t)的距離。然而，當已經訓練好的姿勢數量太龐大時，計算特徵向量間的距離便非常耗時，所以為了解決這個問題，我們提出了一種加速映射的方法。首先我們利用簡單的取樣方法從訓練集中找出一些代表性的關鍵資料點，而這些關鍵點將取代原本的訓練集成為新的訓練集，如圖 5-3 和圖 5-4，而方程式（5.1）、（5.2）、（5.3）、（5.4）中，我們用關鍵資料點來估算一個新資料點的座標向量，由於降低了訓練集的數目，所以也省去了不少時間。而方程式的推導基於假設訓練集資料點的零均值(zero-mean)，因此我們必須位移關鍵資料點的座標如下：

$$x'_{ai} = x_{ai} - \overline{x}_a \qquad (5.7)$$

其中$x_{ai}, i = 1, \dots, k$為關鍵資料點，而$\overline{x}_a = \frac{1}{k}\sum_{i=1}^{k} x_{ai}$。所以$x'_{n+1}$和所有關鍵資料點的內積用方程式（5.4）計算：

$$[x'_{a1}, \dots, x'_{ak}]^T x'_{n+1} = \mathrm{X}_a^T x'_{n+1} = f_a = [f_{a1}, \dots, f_{ak}]^T \qquad (5.8)$$

接著，x'_{n+1}的 least-square solution 被推導如下：

$$x'_{n+1} = (X_a^T X_a)^{-1} X_a^T f_a \qquad (5.9)$$

最後，$x'_{n+1} + \overline{x}_a$可以計算出新資料點的座標向量$x_{n+1}$。在方程式（5.9）中，$(X_a^T X_a)^{-1} X_a^T$是可以事先計算出來的。因此，估計新的姿勢座標在嵌入式的歐幾里得空間中是非常有效的。

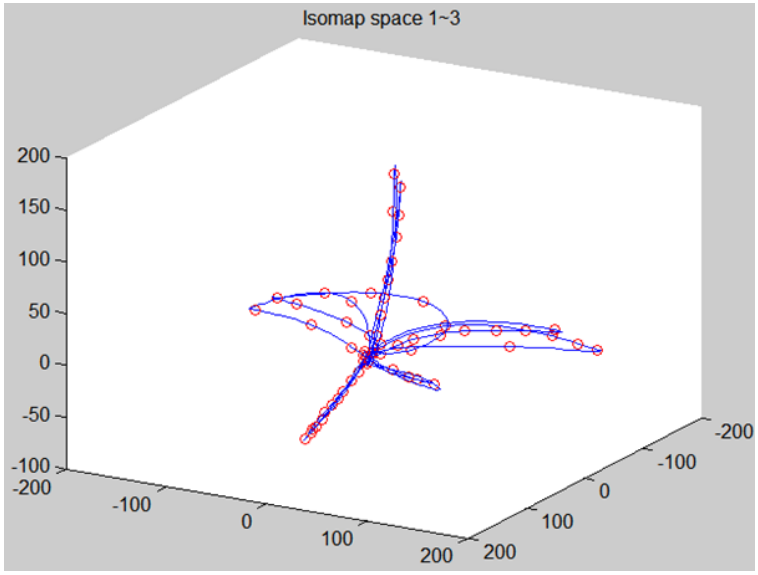

圖 5-3：利用簡單取樣的方法找出訓練集在 1 到 3 維 Isomap 空間的關鍵點

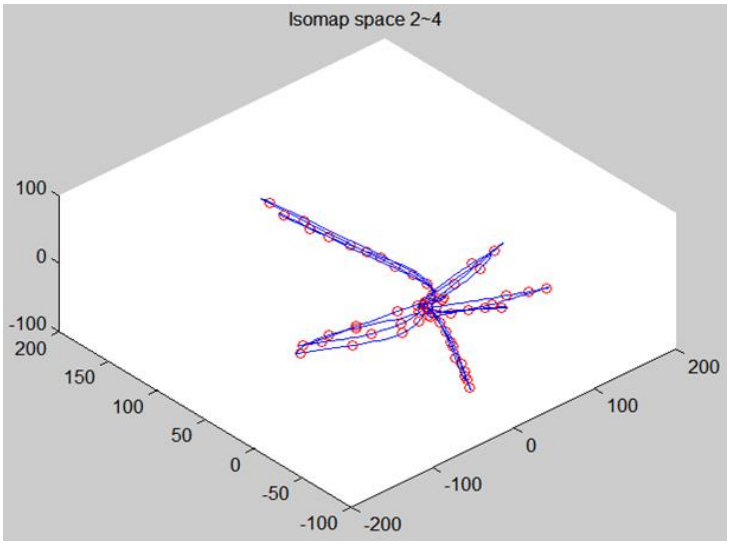

圖 5-4：利用簡單取樣的方法找出訓練集在 2 到 4 維 Isomap 空間的關鍵點

第六章　實驗結果

本研究包含兩個實驗。實驗一：找一位動作標準的人當作訓練者，並提供訓練用的動作序列，利用 Kinect 取得姿勢的三維關節資訊。訓練者的訓練動作序列中包括 32 個基本動作屬於 16 種不同的基本動作類別，共有 2220 個畫面(frame)，16 種基本動作如圖 6-1 所示。訓練動作序列利用 Isomap 演算法建構出維度為 5 的 Isomap 空間如圖 6-2、圖 6-3 及圖 6-4。我們將 Isomap 空間中的流形軌跡描述為低維度的時間序列，接著再利用時間分割的技術將此時間序列分割成許多子序列，如圖 6–5 所示，每個維度表示在 Isomap 空間中的座標，黑色垂直線則表示每一個斷點。根據觀察，某些座標在不同的基本動作中會有不同的時間序列。我們將被分割的基本動作分類成 16 個群集，而且 16 個平均軌跡（mean trajectorie）則用來代表 16 個不同群集的模範，如圖 6–6、圖 6–7 及圖 6–8，一種顏色代表一種基本動作群集。雖然有些基本動作在 Isomap 空間中部份重疊在一起，但利用 DTW 還是可以分在不同的基本動作群集中，因為動作的軌跡方向是不同的。

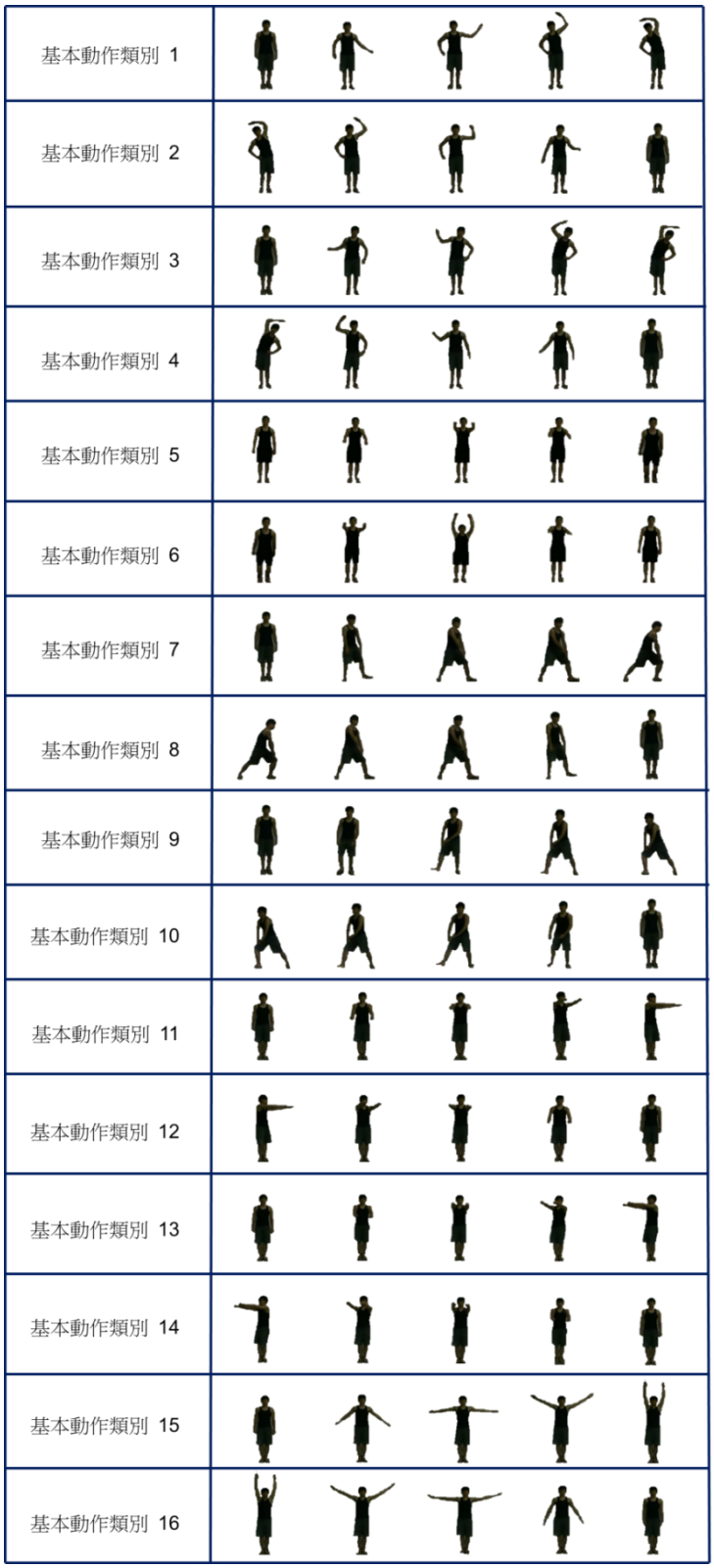

圖 6-1：16 種基本動作類別

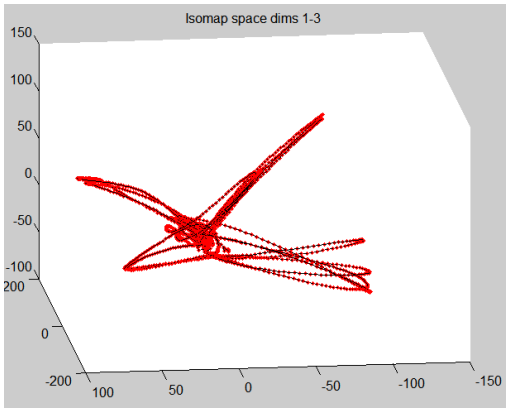

圖 6–2：5 維的流形軌跡映射在 1 到 3 維 Isomap 空間

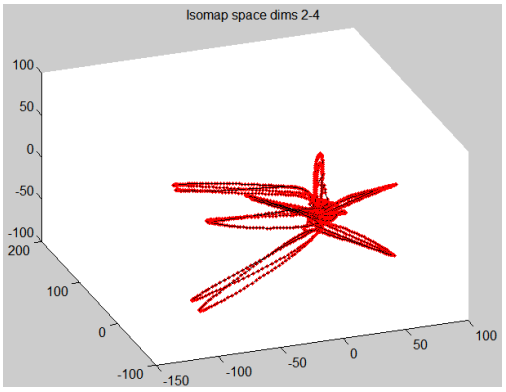

圖 6–3：5 維的流形軌跡映射在 2 到 4 維 Isomap 空間

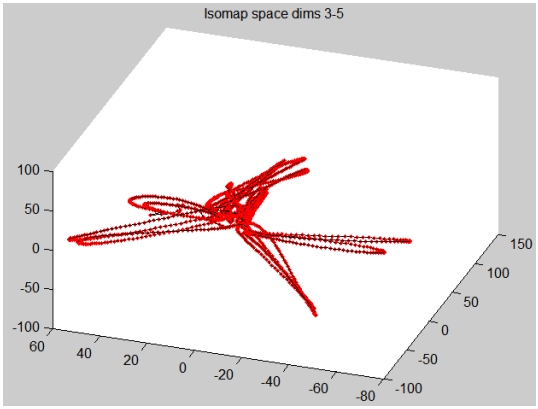

圖 6–4：5 維的流形軌跡映射在 3 到 5 維 Isomap 空間

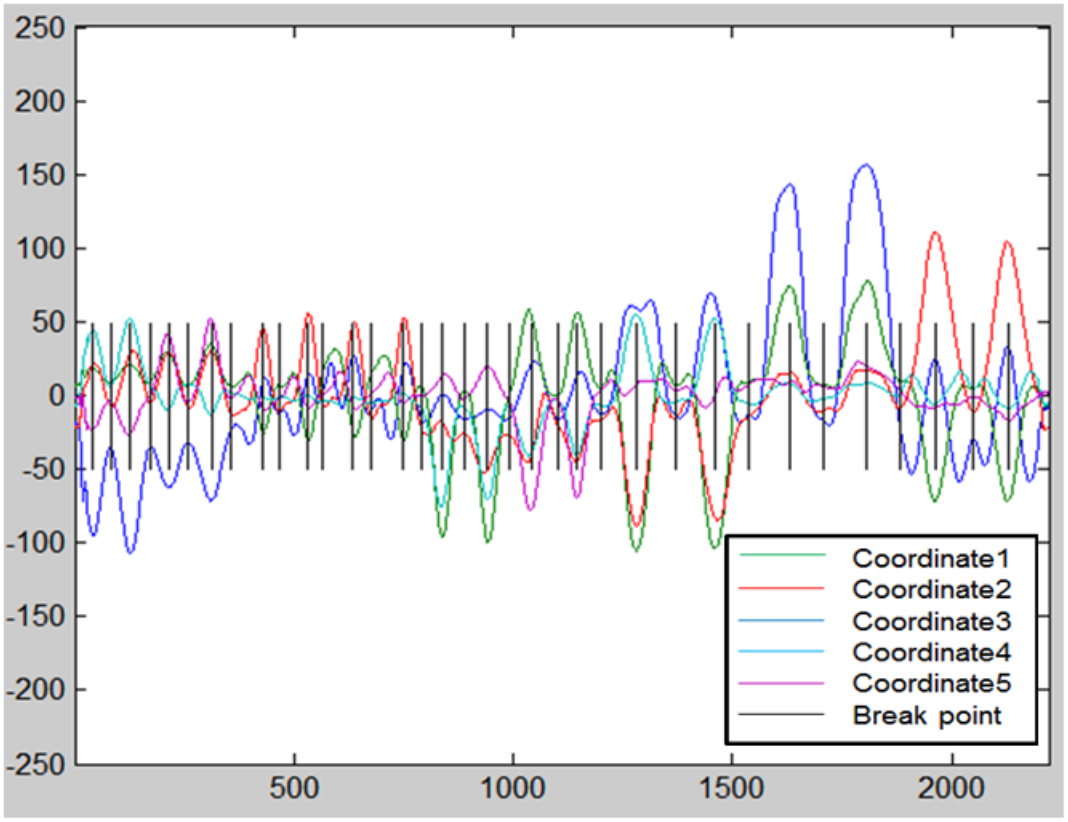

圖 6-5：時間分割結果

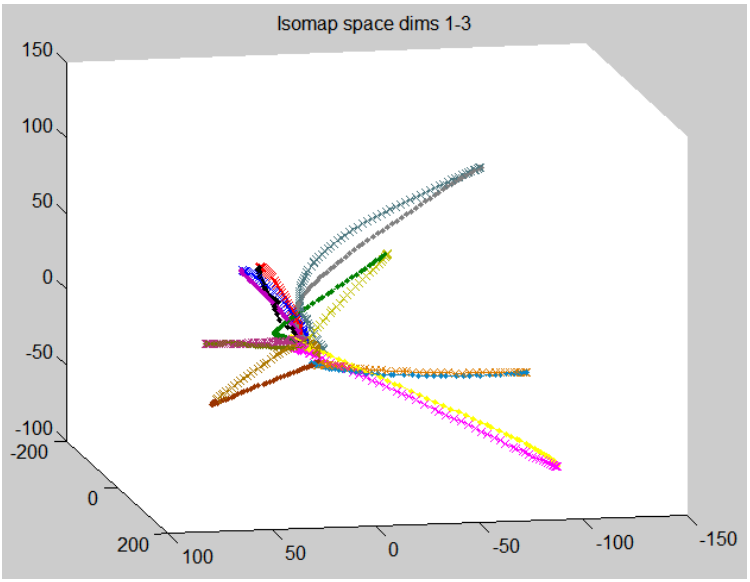

圖 6-6：平均軌跡在 1 到 3 維 Isomap 空間呈現 16 種基本動作（16 種顏色）

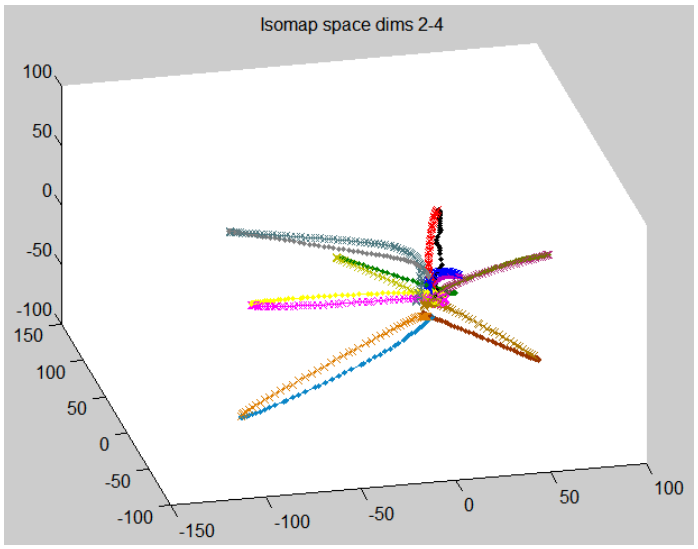

圖 6-7：平均軌跡在 2 到 4 維 Isomap 空間呈現 16 種基本動作（16 種顏色）

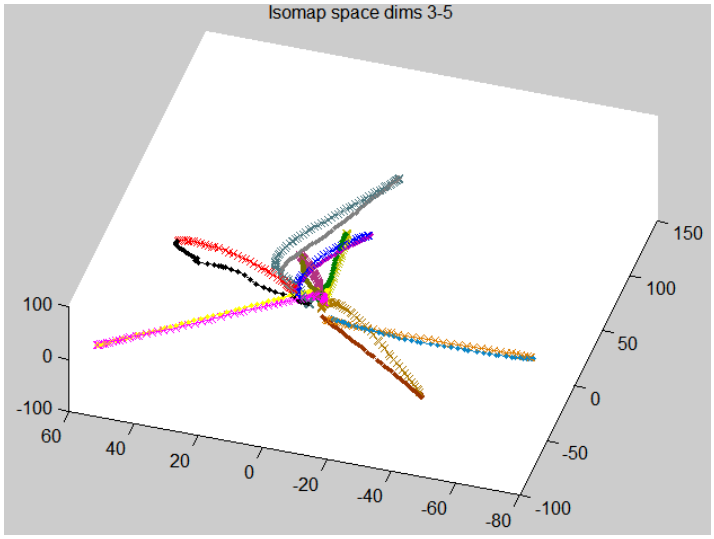

圖 6-8：平均軌跡在 3 到 5 維 Isomap 空間呈現 16 種基本動作（16 種顏色）

接下來我們將原來的訓練動作序列映射到 Isomap 空間，來觀察原來的 Isomap 與只利用關鍵動作點所建的 Isomap 之間的差異。圖 6-9、圖 6-10 及圖 6-11 顯示利用關鍵動作點（在此關鍵動作點數目為 60）所重建的 Isomap，由圖可以發現，重建後的 Isomap 和原本的（圖 6-2 到圖 6-4）非常相似。和原本的映射方法相比，本來要評估 2220 個特徵向量距離，在取完關鍵動作點後，只需要評估 60 個特徵向量距離即可，降低了計算複雜度。

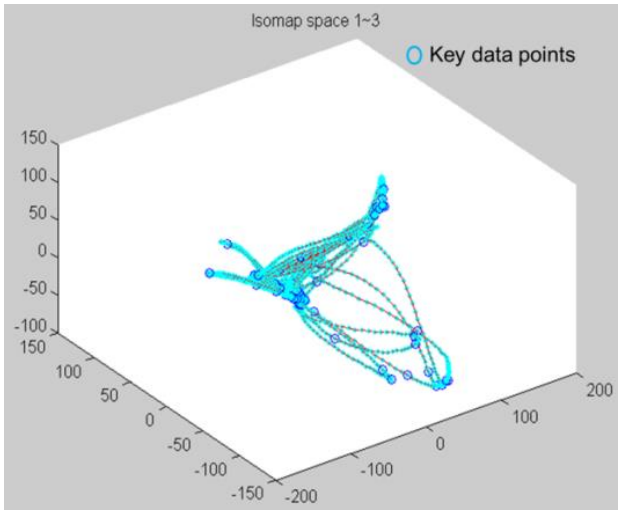

圖 6-9：所選擇的關鍵動作點和所重建的 Isomap 空間（1 到 3 維）

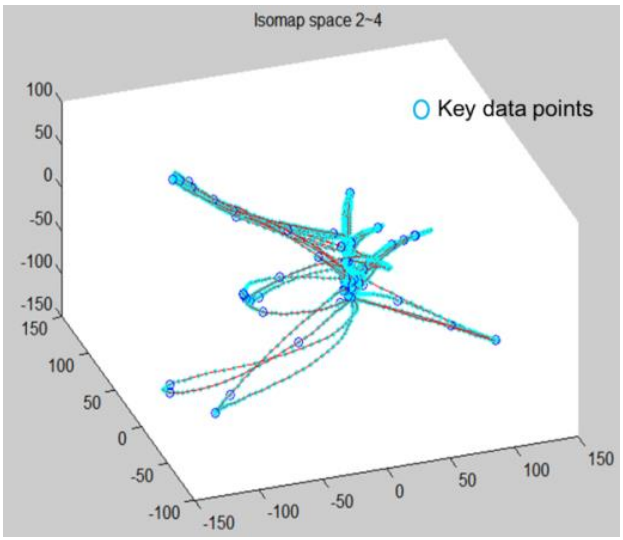

圖 6-10：所選擇的關鍵動作點和所重建的 Isomap 空間（2 到 4 維）

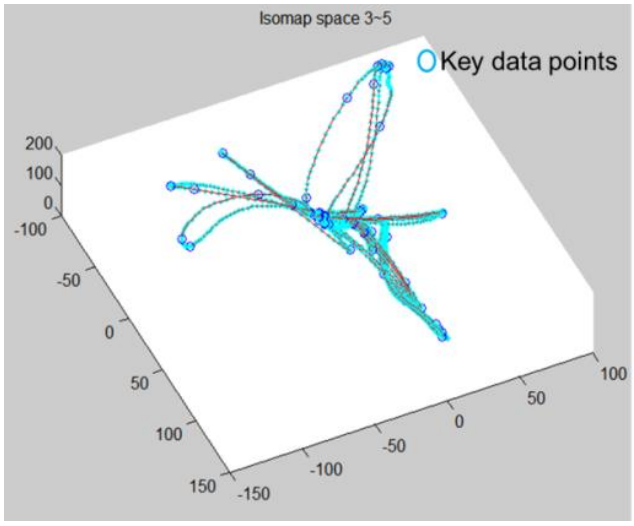

圖 6-11：所選擇的關鍵動作點和所重建的 Isomap 空間（3 到 5 維）

另外我們再找五名測試者提供測試動作序列，同樣利用 Kinect 取得姿勢的三維關節資訊。每一位測試者皆提供 64 個基本動作的動作序列，而 64 個基本動作皆包含在 16 種訓練的基本動作類別中。表 6-1 顯示各類別在 5 位測試結果後的

辨識率。由表 6–1 可觀察出基本動作 1 有 5%的機率辨識成基本動作 15，因為兩個基本動作有部分重複的軌跡如圖 6–12；有兩位測試者所作的基本動作 2 被辨識成基本動作 6，是因為腋下夾角在這兩個基本動作中皆有大幅度變化，所以可能被辨識成同一類基本動作；基本動作 13 和基本動作 5 皆有手往上移動至胸口，基本動作 14 和基本動作 5 皆有手從胸口往下移動至大腿，所以基本動作 13 及 14 皆有 5%及 10%的機率錯誤辨識成基本動作 5；有些測試者做基本動作 9 和基本動作 10 時，膝蓋並不會太彎曲，且手部的移動非常類似基本動作 5，因此皆有 20%的錯誤辨識率。

表 6-1：5 位測試者所執行的 16 種基本動作類別對應的辨識率

實際動作 \ 辨識動作	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.9				0.05										0.05	
2		0.9				0.1										
3			1													
4				1												
5					0.95				0.05							
6						1										
7							1									
8								1								
9					0.2				0.8							
10					0.2					0.8						
11											1					
12												1				
13					0.05	0.1		0.05					0.8			
14					0.1			0.05		0.05				0.8		
15															1	
16																1

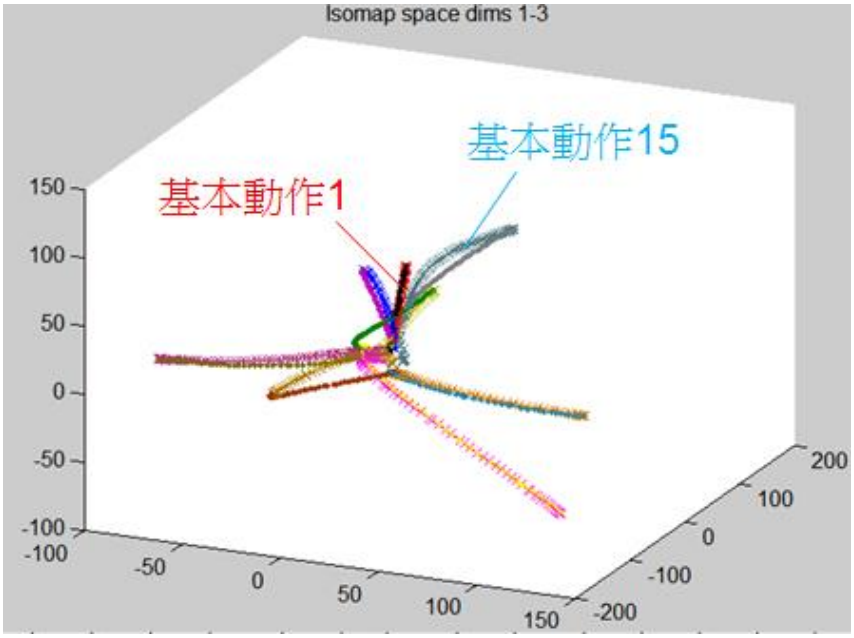

圖 6-12：部分軌跡重複的基本動作 1 和基本動作 15

實驗二：本實驗目的是要測試不同的 Kinect 視角對我們的動作辨識率影響程度。在此，我們同樣利用實驗一中訓練者的訓練動作序列來當作基準，而測試者（和訓練者同一人）則以不同的 Kinect 視角偏差來拍攝，如圖 6-13 所示。在此，我們以 5 度為間隔，分別測試 5 度到 45 度的辨識結果。表 6-2 顯示測試者和 Kinect 在不同的視角偏差所得到的辨識率。在 20 度之後的分類效果便往下降低了，因為有些動作(如圖 6-1 中的基本動作類別 7 到 10)和 Kinect 偏差角度過大時，不容易判斷出膝蓋彎曲，因此造成分類錯誤；又如基本動作類別 11 到 14，視角偏差過大會造成雙手往外拉時，手臂被遮蔽或者無法和背景的深度作區分而影響到辨識率，而動作類別 11 到 14 和基本動作類別 5 有部分移動軌跡重疊，所以有錯誤辨識成基本動作 5 的機會。

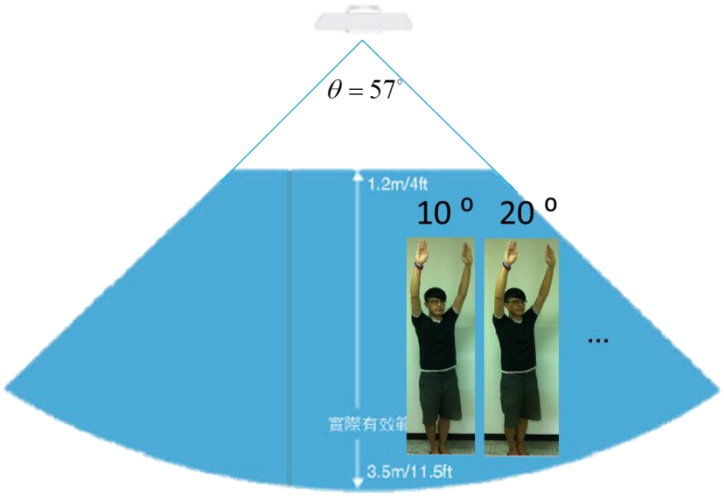

圖 6–13：利用 Kinect 做有角度偏差的拍攝

表 6–2：測試者和 Kinect 視角偏差角度所對應的辨識率

偏差角度	5 度	10 度	15 度	20 度	25 度	30 度	35 度	40 度	45 度
辨識率	100%	100%	97%	89%	72%	68%	60%	39%	35%

第七章　結論及未來方向

本章將對本研究進行總結，並且提出未來可以改進的部分，希望未來此系統可以更穩定並且更人性化。

7.1、結論：

本研究利用視訊資料進行人類動作的分析，目的在於發展一套通用的人類動作分析技術可以應用在不同的領域上。本研究分成兩階段，分別為訓練過程及測試過程。第一階段利用微軟開發的 RGB-D sensor（即 Kinect）取像，取得訓練者訓練姿勢的三維關節資訊並得到三維關節資訊的特徵序列，再利用 Isomap 的技術將所得到三維關節資訊的特徵序列轉換至低維度的 Isomap 空間，將 Isomap 空間中的流形軌跡描述為低維度的時間序列後，利用時間分割的技術將此時間序列分割成許多的基本動作，然後再利用分群的技術，包括動態時間扭曲技術及階層式分群法來對這些基本動作序列分群。第二階段同樣利用 Kinect 取得測試者三維關節資訊的特徵序列後，我們利用 Isomap mapping 的方法映射新座標點到現有的 Isomap 空間中，再利用時間分割的技術將此時間序列分割成許多子序列後，利用基於動態時間扭曲技術的最近鄰比對方法辨識測試者的動作。

使用 Kinect 可以解決身體全域位置、方向與使用者姿態差異影響。因為三維關節資訊的擷取並不會受到上述三者所影響，所得到三維關節資訊的特徵序列較為穩定，亦可使本研究技術穩定許多。

7.2、未來工作：

由於人類的動作太過於複雜，所以本研究先進行低階動作的描述、切割與辨識，未來將基本動作結合自然語言做語意式的描述高階的人類活動將是下一個階段的工作。

參考文獻

- 〔His 08〕 J. W. Hsieh, Y. T. Hsu, H. Y. Mark Liao and C. C. Chen, "Video-based human movement analysis and its application to surveillance systems," *IEEE Transactions on Multimedia*, Vol. 10, pp. 372- 384, 2008.
- 〔Liu 10〕 C. D. Liu, Y. N. Chung, and P. C. Chung, "An interaction-embedded HMM framework for human behavior understanding: with nursing environments as examples," *IEEE Transactions on Information Technology in Biomedicine*, Vol. 14, No. 5, pp. 1236- 1246, 2010.
- 〔Su 07〕 C. W. Su, H. Y. Mark Liao, H. R. Tyan, C. W. Lin, D. Y. Chen, and K. C. Fan, "Motion flow-based video retrieval," *IEEE Transactions on Multimedia*, Vol. 9, No. 6, pp. 1193- 1201, 2007.
- 〔Bid 10〕 B. Bideau, R. Kulpa, N. Vignais, S. Brault, F. Multon, and C. Craig, "Using virtual reality to analyze sports performance," *IEEE Computer Graphics and Applications*, Vol. 30, No. 2, 2010.
- 〔Wei 09〕 J. Wei, H. Qin, J. Guo, and Y. Chen, "The hand shape recognition of human computer interaction with artificial neural network," *IEEE International Conference on Virtual Environments, Human-Computer*

- *Interfaces and Measurements Systems*, pp. 350–354, 2009.
- 〔Zha 11〕 Z. Zhang, "Human body language understanding with 3D sensors," (Panels on 3D Media Analysis and Retrieval), *IEEE International Conference on Multimedia and Expo*, 2011.
- 〔Wan 03〕 L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, Vol. 36, No. 3, pp. 585–601, 2003.
- 〔Tur 08〕 P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 11, 2008.
- 〔Pop 10〕R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, Vol. 28, pp. 976–990, 2010.
- 〔Moe 06〕 T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, Vol. 104, No. 2, pp. 90–126, 2006.
- 〔Lia 13〕 Y. M. Liang, S. W. Shih, and A. C. C. Shih, "Human action segmentation and classification based on the Isomap algorithm," To Appear in *Multimedia Tools and Applications*.

- 〔Sch 04〕 C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," *Proceedings of the 17th IEEE International Conference on Pattern Recognition*, pp. 32-36, 2004.
- 〔Bla 07〕 B. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29, No. 12, 2007.
- 〔Moe 01〕 T. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, Vol. 81, No. 3, pp. 231-268, 2001.
- 〔Har 00〕 I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, 2000.
- 〔Bob 01〕 A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267, 2001.
- 〔Li 07〕 H. Li, S. Lin, Y. Zhang, and K. Tao, "Automatic video-based analysis of athlete action," *Proceedings of IEEE International Conference on Image Analysis and Processing*, pp. 205-210, Modena, Italy, Sep. 2007.

- 〔Muk 11〕 S. Mukherjee, S. K. Biswas, and D. P. Mukerjee, "Recognizing human action at a distance in video by key poses," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 21, No. 9, 2011.
- 〔Ohy 02〕 J. Ohya, "Analysis of human behaviors by computer vision based approaches," *Proceedings of IEEE International Conference on Multimedia and Expo*, Vol. 1, pp. 913–916, Lusanne, Switzerland, Aug. 2002.
- 〔Bou 03〕 B. Boulay, F. Bremond, and M. Thonnat, "Human posture recognition in video sequence," *Proceedings IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 23–29, 2003.
- 〔Wer 05〕 N. Werghi, "A discriminative 3D wavelet-based descriptors：application to the recognition of human body postures," *Pattern Recognition Letters*, Vol. 26, pp. 663–677, 2005.
- 〔She 99〕 D. Shen and H. H. S. Ip, "Discriminative wavelet shape descriptors for recognition of 2-D patterns," *Pattern Recognition*, Vol. 32, pp. 151–165, 1999.

- 〔Gu 10〕 J. Gu, X. Ding, S. Wang, and Y. Wu, "Action and gait recognition from recovered 3-D human joints," *IEEE Transactions on System, Man, and Cybernetics – Part B: Cybernetics*, Vol. 40, No. 4, 2010.
- 〔Lin 11〕 S. Y. Lin, Z. H. You, and Y. P. Hung, "A real-time action recognition approach with 3D tracked body joints and its applications," *Proceedings of the 24th IPPR Conference on Computer Vision, Graphics, and Image Processing*, 2011.
- 〔Ham 07〕 R. Hamid, S. Maddi, A. Bobick, and I. Essa, "Structure from statistics, unsupervised activity analysis using suffix tree," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-8, 2007.
- 〔Ryo 06〕 M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," *Proceedings of IEEE Computer Society Conference on Computer Vision*
- *and Pattern Recognition*, 2006.
- 〔Tra 08〕 S. D. Tran and L. S. Davis, "Event modeling and recognition using Markov logic networks," *Proceedings of the 10th European Conference on Computer Vision: PartII*, 2008.

- 〔Cai 08〕 F. Caillette, A. Galata, and T. Howard, "Real-time 3-D human body tracking using learnt models of behaviour," *Computer Vision and Image Understanding*, Vol. 109, pp. 112-125, 2008.
- 〔Lia 09〕 Y. M. Liang, S. W. Shih, A. C. C. Shih, H. Y. M. Liao, and C. C. Lin, "Learning atomic human actions using variable-length Markov models," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, Vol. 39, No. 1, 2009.
- 〔Nev 00〕 C. G. Nevill-Manning and I. H. Witten, "On-line and off-line heuristics for inferring hierarchies of repetitions in sequence," *Proceedings of the IEEE*, Vol. 88, No. 11, 2000.
- 〔Li 12〕 J. Y. Li, A. C. Luo , and S.W.Chen , "Automated Lecture Recording System-Virtual Cameraman", The 25th IPPR Conference on Computer Vision, Graphics, and Image Processing, session F2-8, 2012.
- 〔Jai 06〕Law MHC, Jain AK (2006) "Incremental nonlinear dimensionality reduction by manifold learning," IEEE Trans Pattern Anal Mach Intell 28(3):377–391

- 〔Ten 12〕Joshua B. Tenenbaum, Vin de Silva and John C. Langford “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” 450 Serra Mall Stanford University Stanford, CA 94305-2130 (650) 724-4676
- 網路資料來源：

  TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/

  Kinect for Windoes, http：//kinectforwindows.org/download/