在複雜背景作基於模版之手追蹤

唐政元
華梵大學資管系
cytang@cc.hfu.edu.tw
吳怡樂
台科大資工系
ywu@csie.ntust.edu.tw
吳松航
華梵大學資管系
m9525020@cat.hfu.edu.tw
趙士賓
工研院電光所
chaoshihpi@itri.org.tw
陳彥伶
華梵大學資管系
m9525004@cat.hfu.edu.tw
李翊綸
華梵大學資管系
m9525026@cat.hfu.edu.tw

摘要

本論文主要是開發一個可用於複雜背景、不需貼 marker(markerless，無標記)的手部追蹤系統。利用手的輪廓模型，目前的系統已可以追蹤不同遠近、不同角度的手，且也可以作簡易的手勢辨識。本研究包含手部的辨識與追蹤。我們的系統包含建立模版與手部追蹤兩個主要部份。

在建立模板部分，我們根據角度建立了 72 個模板(每 5 度建一個模版)；根據影像 edge 的 orientation，分成數個 channels，以利後續的比對使用。而在手部追蹤部分，我們對輸入的影像，依序做 skin color detection、orientation、以及 distance transform (DT) ，使 template 與影像中手的部分能 match，成功的追蹤到手部。初始的第一張影像偵測人手掌是採用所有角度的模版找，之後的追蹤是根據上個時間影像的追蹤結果作些許角度變化的搜尋。初步的實驗結果相當不錯，未來將進一步應用到實際的應用上。

關鍵詞：追蹤、手部輪廓、基於模版比對、無標記手部追蹤、複雜環境

一、緒論

最近一、二十年來，由於電腦計算能力的大幅提昇與電腦視覺（computer vision）的理論不斷的突破，電腦視覺在實際應用上有相當大的進展。例如：基於視訊的人機介面（video-based human-computer interface)、基於視訊的智慧型保全系統（video-based intelligent security system)、基於模式的視訊壓縮（model-based video compression，如 MPEG-4)、以影像或視訊內涵為基礎的資料庫搜尋（content-based image/video retrieval）等。而利用視訊資料作監控（surveillance）則是電腦視覺研究領域上一個非常主要的應用。此項監控技術可用在社會安全、工業與國防等，對整個社會有相當大的重要性。視訊監控是藉由分析從一個或多個攝影機所取得的影像，以用來對大範圍、複雜的區域或分散在好幾個地方的區域作監督與控制，諸如：市中心、飛機場、公路與鐵路網路系統、個別的建築物等等。在本研究計畫中，我們希望能利用視訊監控的相關技術來建構一個智慧型數位學習環境。

基於視訊的人機介面與基於視訊的保全系統最主要的相同點即在於都是使用電腦視覺的技術來處理（包括追蹤與識別等）一系列影像中所包含的人的資訊。此項研究課題大致可以分成兩大類：一類是追蹤整個人的身體，一類是追蹤人的局部部位（如頭部運、臉的表情、手的動作和眼球的轉動等等)。基於應用價值與商機的關係，探討有關以人（如人頭追蹤、人臉辨識、手勢辨識）方面的研究，在最近幾年非常受到重視且非常盛行。其中，對偵測、追蹤人頭或人臉（臉部表情）的研究非常地多，甚至連資訊界的巨擘—Microsoft 不僅在本部與中國大陸成立 Vision 部門且與多個研發單位進行多向合作，積極的投入此項研究。追蹤人的不同部位有其不同的應用，本研究計劃主要專注於偵測與追蹤動態影像中人臉的運動，在偵測與追蹤當中再做人臉辨識或行為辨識。

近年來，電腦視覺的研究學者對人臉、手勢的追蹤辨識與分析的研究課題越來越重視，目前已有一個國際會議（International Workshop on Automatic Face- and Gesture- Recognition）專門探討此一研究課題（1995 年 6 月於瑞士 Zurich 舉辦第一屆，1996 年 10 月於美國舉辦第二屆，並於 1998 在日本舉辦第三屆），可見本研究課題的重要性與熱度。IEEE Transaction on PAMI 在 1997 有個專刊（special issue）專門探討有關人臉與手勢辨識上的研究。而且在一些有關電腦視覺的主要國際會議上（如 ICCV, ACCV 和 ECCV)，偵測與辨識人臉、手勢的相關研究都佔有相當大的比例。

現今社會中，使人們與電腦的互動更人性化

為重要議題。目前大多需要透過滑鼠等相關的硬體設備來操控電腦，希望將來趨勢能以非接觸的方式來做相關互動。若以現今非接觸方式的互動設備，以 Wii 為最具熱門之商品，但是目前此相關商品仍然受限於過多的硬體，能做出的變化也有限。因此若透過手部的追蹤，加以應用，便能以更人性化的方式做出多種變化。本研究是參考 University of Cambridge B. D. R. Stenger 在 2004 的博士論文[1]與在 2006 IEEE Trans. Pattern Analysis and Machine Intelligence 所發表的論文。

二、文獻探討

目前，在世界上電腦視覺研究領域當中，使用攝影機從事相關人臉偵測與追蹤[3]的研究單位相當多[2][4][5]。當今世界上從事有關人臉追蹤研究的機構相當多，因篇幅有限，在此只列出一些有關的研究：麻省理工學院媒體實驗室的 A. Pentland 與人工智慧實驗室的 Poggio、伊利諾大學（UIUC）的 Thomas S. Huang、劍橋大學（Cambridge University）的 Cipolla、馬里南大學（University of Maryland）的 Davis、英國 Queen Mary and Westfield 學院的 S. G. Gong 和台灣中研院的 Hung。Pentland 發展了一個 IVE（Interactive Video Environment）測試系統、即時的 Pfinder（Person Finder）追蹤系統、臉的辨識等系統，其應用包括：利用手勢（gesture）來作控制的人機介面、利用 Pfinder 所偵測到的位置結合圖學以產生虛擬的化身（avatars）。Thomas Huang 比較著眼於人臉、臉的表情和手勢的自動偵測與辨識；並結合人臉表情模型，追蹤人臉上的特徵點，以獲得人臉表情的演變。Cipolla 一方面使用基於手勢的介面（gesture-based interface）來對機器人作導引；另一方面也利用最黑影像點為特徵，結合 RANSAC 技術，以建立一個快速的人頭追蹤系統。美國 MIT 的 Basu 與 Pentland 使用橢球模型來做基於模型的頭部追蹤；美國 CMU 的 Stiefelhagen 與 Yang 提出一個基於模型的人頭追蹤系統來估測人頭的凝視點。Metaxas 整合光流與變形（deformable）模型來對人臉的表情與運動作追蹤。Harashima 則利用人頭與手臂的追蹤來作基於模型的影像壓縮。而在美國的 University of Maryland 相當聞名的電腦視覺實驗室（computer vision laboratory）是其中做得較好的研究單位之一。他們發展了一套相當有名的系統：W4[5][6][7]。W4 是一個在戶外用來偵測、追蹤人們與監控這些人的活動之即時（real-time）視訊監控系統。W4 的意思是“Who？ When？ Where？ What？”，即是系統能夠知道被監視的人，他們在做什麼（What？）、他們在哪裡、在何時動作（When？ Where？）與他們到底是誰（Who？）。他們不僅僅是學理上的探討，也成功地把他們所發展出的系統用在監控他們的大學校園。

(一) 使用膚色

無論是在人臉的偵測或是手部追蹤的領域中，膚色偵測是首要的議題。人的膚色和背景顏色通常有相當的差異，因此使用 RGB 為主要的色彩空間，其偵測出的膚色會因光線的不同而不穩，並不適合用於描述膚色的分佈範圍。然而根據研究顯示（Garcia & Tziritas, 1999；Chai & Ngan, 1999），HSV 色彩空間可對皮膚顏色及背景顏色有效的分離。所以本研究將影像由 RGB 色彩空間轉換為 HSV 色彩空間，兩者間的轉換公式如下所示：

$$H1 = \cos^{-1}\left\{\frac{0.5\left[(R-G)+(R-B)\right]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\} \quad ...(1)$$

$$H = \begin{cases} H1 & \text{if } B \le G \\ 360^\circ - H1 & \text{if } B > G \end{cases} \quad ....(2)$$

$$S = \frac{\text{Max}(R, G, B) - \text{Min}(R, G, B)}{\text{Max}(R, G, B)} \quad ....(3)$$

$$V = \frac{\text{Max}(R, G, B)}{255} \quad ....(4)$$

其中色調 H(Hue)為膚色的依據，主因為 H 較不容易受到光的強弱影響；S 代表顏色中的飽和度(Saturation)，其值介於 0 到 1 之間；V 代表顏色的明暗度(Value)，也是介於 0 到 1 之間。本研究去除 V 的部份，且將 H 值定義為 0 到 0.08，S 值定義為 0.23 到 0.63。

(二) 使用 Edge orientation

許多研究中，得出 edge orientation 的 filter 有多種。本研究利用 sobel filter 的特性算出水平和垂直的角度，再利用 gradient 算出整張影像的每個 edge 的 orientation。

Sobel filter，包含兩個基本運算子如下：

$$hx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad hy = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad ...(5)$$

hx 與 hy 分別對 x 和 y 方向微分，∂f 表示整張影像，則整張圖對垂直與水平斜率定義為：

$$\begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix} .....(6)$$

$$\partial x = hx * \partial f \quad .....(7)$$

$$\partial y = hy * \partial f \quad .....(8)$$

其表示 edge 方向角度 θ 為:

$$\theta = \tan^{-1}\left(\frac{\partial f/\partial y}{\partial f/\partial x}\right) \text{.....(9)}$$

(三) Distance transform

Distance transform 是一種應用在二元影像的運算子，其運算結果則為一灰階影像。與一般灰階影像不同，其強度並非表示亮度值，而是表示物件內部每一點與物件邊緣的距離。

如果以 1 表示物件像素，0 是背景像素，則 distance transform 定義為對於每一個物件區域的像素，計算其與最近的背景像素的距離，並以此距離值取代原像素值。

設有物件內部的兩點 p1=(x1, y1), 和 p2=(x2, y2)，其距離可以用以下三種 distance metrics 表示：

Euclidean distance：

$$D_e(p1,p2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \text{ .....(10)}$$

相鄰的 pixel 為 1 分相鄰斜角的 pixel 以上面公式計算。結果如圖 2.1。

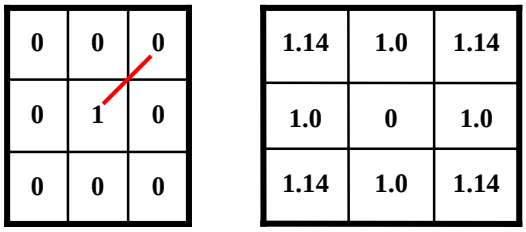


圖 2.1 左圖為原圖，右圖為 Euclidean distance

常用 Distance Transform 計算 distance 的方法主要有兩種：City block distance 與 Chessboard distance。分別敘述如下。

City block distance：

$$D_4(p1,p2) = |x1 - x2| + |y1 - y2| \text{ .....(11)}$$

相鄰的 pixel 為 1 分相鄰斜角的 pixel 以上面公式計算。結果如圖 2.2。

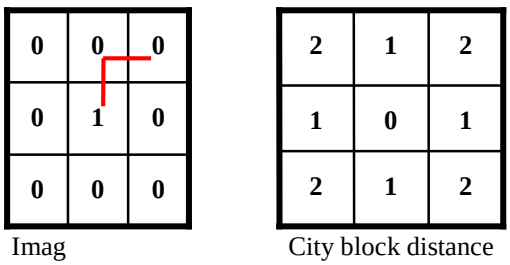


圖 2.2 左圖為原圖，右圖為 City block distance

Chessboard distance：

$$D_8(p1,p2) = \max(|x1 - x2|, |y1 - y2|) \text{ .....(12)}$$

相鄰的八個pixel均為1分。結果如圖2.3。

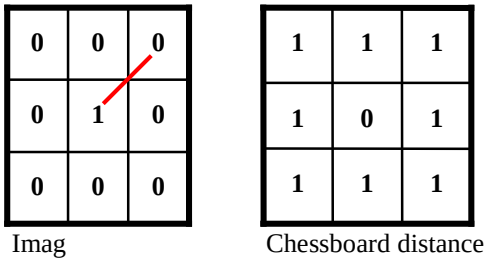


圖 2.3 左圖為原圖，右圖為 Chessboard distance

在本論文中，用來作手部追蹤的 Distance Transform 計算方式，是採用 Chessboard distance。

三、追蹤流程與相關演算法

本文以 skin color 作為主要的搜尋範圍，利用此範圍加以套用所建立的手部 template 來辨識出正確的手部位置和手勢。手部追蹤流程包含兩個部份：1.建立 templates、2. templates 比對、3.tracking。

（一）建立 template

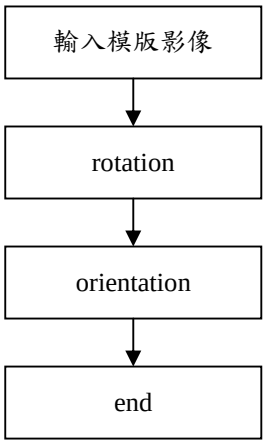


圖 3.1 建立 template 流程圖

此手部追蹤主要有兩種手勢之判別，因此所建立的 template 分別為圖 3.2 以及圖 3.3。

圖 3.2 手勢一　　圖 3.3 手勢二

首先，分別對圖 3.2 和圖 3.3 以順時針為方向，每轉 10 度建立一個 template，共建立 36+36=72 個 template。

其次，分別對每個 template 轉灰階後做 canny edge 處理，接著針對 edge 上的亮點做 orientation[$\frac{\pi}{2}$,-$\frac{\pi}{2}$]。將[$\frac{\pi}{2}$,-$\frac{\pi}{2}$]分成 6 種 channel：[0,30],[31,60],[61,90],[0,-30],[-31,-60],[-61,-90]。如圖 3.4 表示。

			
	0˚~ 30˚	31˚~ 60˚	61˚~ 90˚
[$\frac{\pi}{2}$,-$\frac{\pi}{2}$]			
	0˚~ -30˚	-31˚~-60˚	-61˚~ -90˚

圖 3.4 6-channel 以顏色表示

最後把處理 216 個 template 所產生的資訊儲存起來，分別為 6-channel 的亮點座標以及整張 template 圖。

(二) Templates 比對

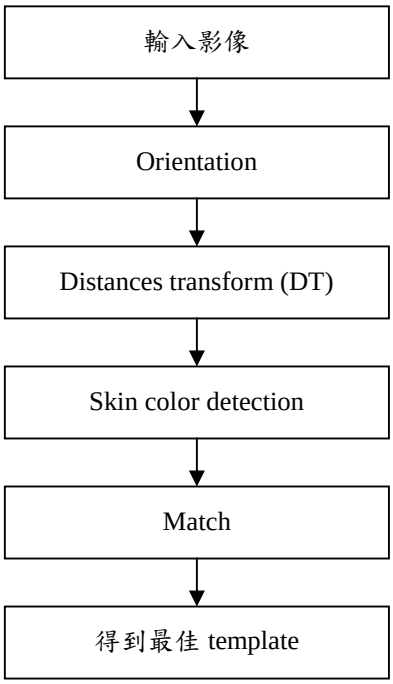


圖 3.5 Templates 比對流程圖

在 template 比對之前，我們必須要對所輸入的影像做一些處理，在比對前所需要的資訊要有 6-channel 和 DT 的資訊。

首先，將整張影像轉灰階後做 canny edge 處理，接著針對 edge 上的亮點做 orientation[$\frac{\pi}{2}$,-$\frac{\pi}{2}$]，分為 6 channel。此 6 channel 的處理程序必須與 template 相同，兩者在 match 時，手部的 6 channel 分佈才會一致。

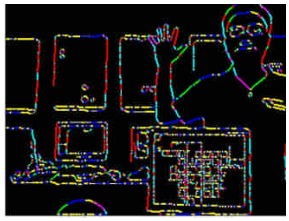
	
原始影像	orientation
	
distance transform (DT)	skin color

圖 3.5 將原始影像以 orientation、DT、skin color 方式處理

我們利用影像 DT 值和 6 個 templates 上 edge 的角度等相關資訊求出最小的分數。首先，對原始影像的 edge 分成六種 orientation，然後再對六種角度分別做 DT(如圖 3.6)。

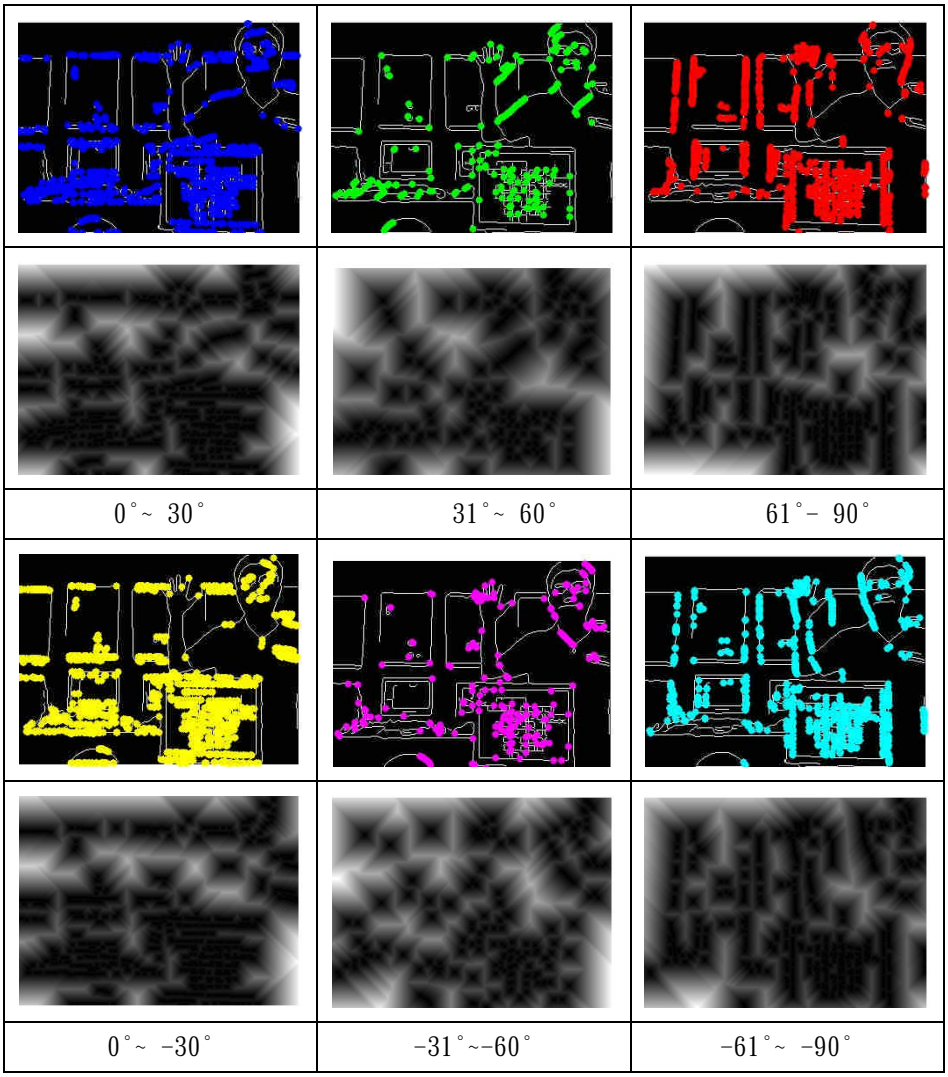


圖 3.6 對一張原始圖分成 6 channel 分別作 DT

template 比對時，先將第一個 template 拿到原始圖上，從左上依序收尋到右下，為了加快速度每隔 5 pixel 搜尋，對搜尋的座標點再 crop 一個區域(10*10)計算裡面 skin color 的數量點，如果大於 50 pixel 點時再進行 template 套用。

選定可執行的座標點後，將 template 的 6 channel 再對原始圖上 6 channel 作比對。依此類推，216 個模版依序套用，然後選取最佳的模版(如圖 3.8)。

圖 3.8 分別得出的最佳套用的兩種 template。

我們在 template 比對的過程，為了克服手部在影像中變化的比例，將 template 大小比例分成 3 種 scale。

以連續的影像為主，手部在大小變化的過程

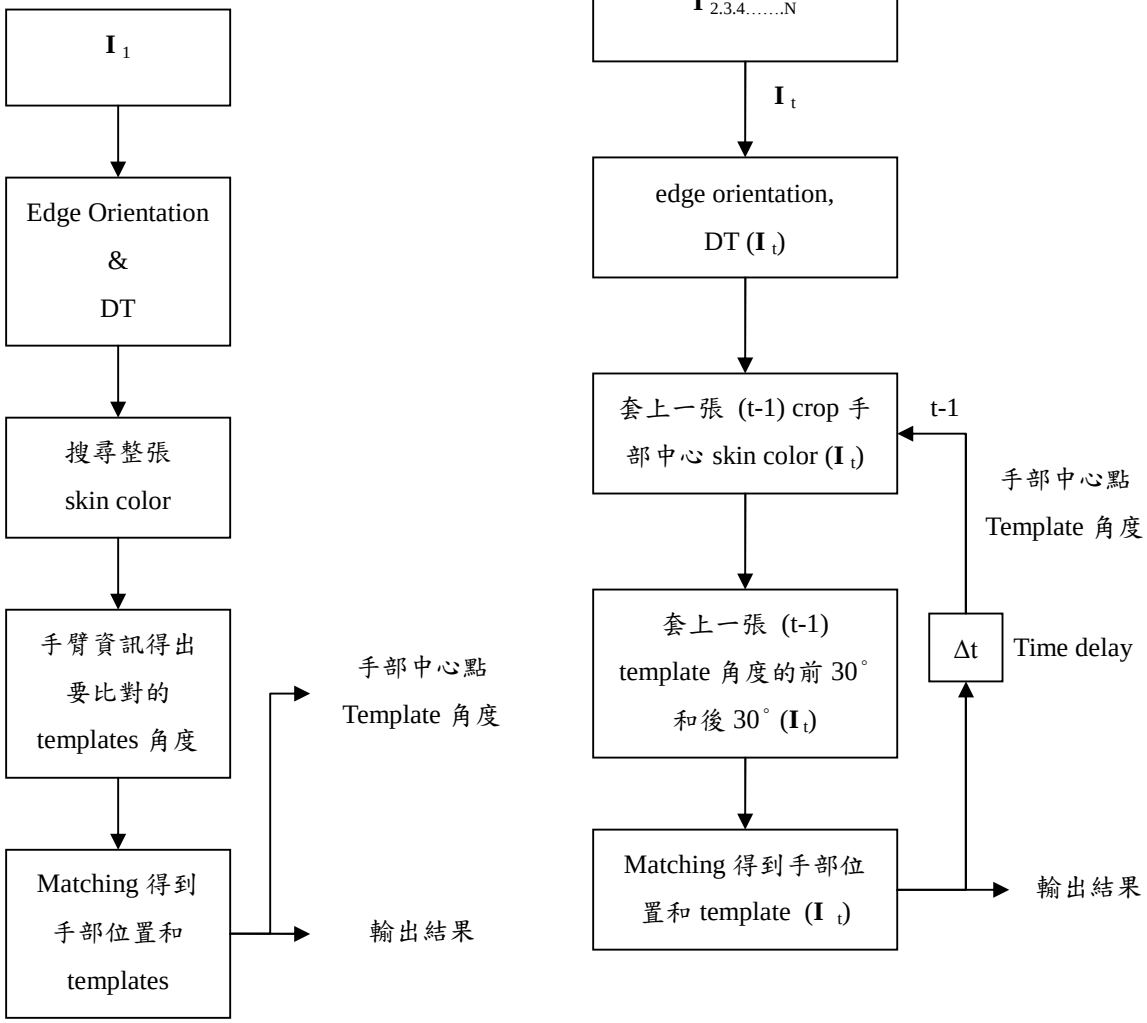
中不會突然的放大或縮小，因此我們認為手部在移動的前後距離以 10 公分為基準，其變化大可都在一個 scale 的範圍內。

如果再建立 template 的過程中，加入 3 個 scale 的 template，我們以兩手勢為主，template 數目為 2*3*36=216。

在追蹤的過程中還是以手勢為主，分層的往下比對，template 越多，我們比對的速度相對增加，這也是未來將要克服的問題。

圖 3.9 將手部 scale 分成 3 種。

(三) Tracking



(a) 偵測模組

(b) 追蹤模組

圖 3.10 手部追蹤流程圖(a)(b)。

追蹤分成兩個模組：偵測模組與追蹤模組。本手部追蹤主要為連續的影像追蹤，因此將追蹤分為兩部分，一部分為第一張初始值，第二部份為第一張之後，利用初始值所得的資訊追蹤手部。

第一張影像初始值極為重要，因此我們加入手臂的資訊，將第一張手的位置及角度更為顯著。

在偵測手臂部份，其方法是認為在膚色的範圍裡，手臂部分的特性為 edge 上方向的一致性是最多的，其方向可以代表我們套用手掌角度的中間值(如圖 3.6)。這部分我們只針對第一張(初始值)做手軸的處理，第二張以後得出正確角度模板就 crop 中心點位置搜尋範圍和套用前三後三張模板，以增加速度。

利用手臂資訊得出角度之後，必須挑選適當的模板，我們利用影像 DT 值和 6 個 templates 上 edge 的角度等相關資訊求出最小的分數。首先，對原始影像的 edge 分成六種 orientation，然後再對六種角度分別做 DT。

接下來第二張之後，會利用前一張所擁有的資訊來加快速度和更精準的追蹤，得知前個時間點的手部位置之後就可框選更小的搜尋範圍，並且減少 template 的套用。因為在前個時間點手部位置的移動和手部變化並不大，所以我們就可以利用此特性來加快執行時間和精準的追蹤。

四、結論與未來方向

在本實驗結果中，兩手勢在連續的影像 match 裡是相對準確的。因此之後利用即時方式處理，以及 3D 資訊，便可達到理想的空間滑鼠。

本研究試著採用影像處理的觀念，來進行手部的辨識與追蹤，我們的方法包含建立模版與手部追蹤兩個主要部份。本論文主要是開發一個可用於複雜背景、不需貼 marker 的手部追蹤系統。利用手的輪廓模型，目前的系統已可以追蹤不同遠近、不同角度的手，且也可以作簡易的手勢辨識。

未來工作包含(a)讓系統更穩定偵測與追蹤到人手；(b)增加更多種的手勢變化追蹤；(c)希望改善追蹤的速度。

五、參考文獻

[1] B. D. R. Stenger. *Model-Based Hand Tracking Using a Hierarchical Bayesian Filter*. PhD thesis, University of Cambridge, March 2004.

[2] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland, "Visually Controlled Graphics," *IEEE Trans. Patt. Anal. Mach. Intell.* 15, 6 (1993) 602-605.

[3] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic, San Diego, CA, U.S.A., 1988.

[4] S. Basu, I. Essa, and A. Pentland, "Motion Regularization for Model-Based Head Tracking," *Proc. Int. Conf. on Pattern Recognition*, Vienna, Austria, pp. 611-616, 1996.

[5] I. Haritaoglu. *W(4): A Real-Time System for Detection and Tracking Of People and Monitoring Their Activities (Visual Surveillance, Silhouette, Remote Sensing)*. Ph.D. Dissertation. University of Maryland, College Park, MD, USA, 1999.

[6] I. Haritaoglu, D. Harwood and L. S. Davis, "W4: Who, When, Where, What: A Real Time System for Detecting and Tracking People," *Proceedings of 3th International Conference on Face and Gesture Recognition*, Nara, Japan, pp. 222-227, 1998.

[7] I. Haritaoglu, R. Cutler, D. Harwood and L. S. Davis, "Backpack: Detection of People Carrying Objects Using Silhouettes," *Proceedings of 7th IEEE International Conference on Computer Vision*, Greece, pp. 102-107, 1999