行政院國家科學委員會專題研究計畫 成果報告

三維物件掃瞄及裸眼立體投影顯示系統開發
研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 98-2221-E-216-031-
執行期間：98年08月01日至99年10月31日
執行單位：中華大學資訊工程學系

計畫主持人：鄭芳炫

計畫參與人員：碩士級-專任助理人員：呂長鴻
碩士班研究生-兼任助理人員：郭子豪
碩士班研究生-兼任助理人員：張裕邦
碩士班研究生-兼任助理人員：陳昀蔚
碩士班研究生-兼任助理人員：徐永成

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫涉及專利或其他智慧財產權，2年後可公開查詢

中華民國 99年11月04日

行政院國家科學委員會補助專題研究計畫■成果報告
□期中進度報告

三維物件掃瞄及裸眼立體投影顯示系統開發

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC98－2221－E－216－031
執行期間：2009 年 8 月 1 日至 2010 年 10 月 31 日

計畫主持人：鄭芳炫
共同主持人：
計畫參與人員： 呂長鴻、郭子豪、張裕邦、陳昀蔚、徐永成

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各二份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
■涉及專利或其他智慧財產權，□一年■二年後可公開查詢

執行單位：中華大學資訊工程學系

中　華　民　國　99　年　11　月　2

摘要

本計畫研發以結構光(Structured Light)的方式，利用照相機與投影機取得物體表面3D座標的方法。由於照相機的鏡頭是由球面鏡所組成，因此拍攝的影像都會有變形(Lens Distortion)的問題，我們必須先用非線性的方式(non-linear)校正照相機之影像後，才能利用照相機之影像作為計算座標的來源。另外投影機所投影出來的顏色也會有色偏(Color Distortion)的現象，顏色是由不同的基本色組成，輸出顏色時會有基本色比例上的誤差，因此，在我們要投出影像之前，必須先校正所有的顏色，使投影機投出的各種顏色能夠接近理想值。結構光3D取像的方式是利用照相機與投影機間的同位(Corresponding)資訊來計算座標，我們利用 紅、綠、藍、黃、青、紫、白顏色，而我們給予每個顏色一個數字代表，並且使用這些顏色排列組合(Computation)的方式編碼來產生圖樣(Pattern)，經由投影機將圖樣投影於被掃瞄物體上，再使用照相機取像，藉由取得圖樣的解碼與照相機的同位資訊，利用這些資料來計算出物體表面的3D座標，由於編碼是由顏色所組成，所以在解碼時必須先取得影像中的顏色群組(Color Group)，將此顏色群組依照其顏色的排列，求出此一顏色群組再所有群組中的順序，進而求出此顏色群組中每一個顏色的順序，得知每一個顏色的順序後即可求得該顏色的角度。將顏色轉換成數字的過程中，我們使用色調(Hue)的數值來判別顏色，使用色調來判別的方法可以去除亮度(Luminance)對我們判別顏色的影響。另外我們也利用排列組合的特性來過濾部分物體表面所造成的錯誤，使得在缺少部分線條的狀況下，可以取得現有的正確顏色資訊，也就是提供錯誤容忍(Fault Tolerance)機制，除此之外，使用排列組合編碼，在解析度上以及執行速度上都有不錯的表現，在我們的實驗中也掃描各類物件，實驗證明此方法可以正確的取得物件3D座標，並且與其他方式比較解析度，我們使用的編碼方式確實能夠提升物件3D取像的解析度，與目前最高的 De Brujin 編碼方式比較，可以提升46%的解析度。

關鍵字：結構光，三維掃描，面部掃描，圖像校正，顏色校正，三維坐標，三維重建

Abstract

This project provides an approach to scanning a real object by *Structured Light* and reach 3D coordinate by digital camera and DLP projector. Camera shoots composes by lens and it brings the *image distortion* of capturing picture. Therefore, we must use *non-linear* correction to solve this issue. The image must be corrected before calculating the 3D coordinate of objects. Moreover, *Color distortion* is also happened on common projectors. Colors of projector are composed with base-colors which filtered by *Color Wheel*. Some colors are not perfect with the RGB ratio from projector. That is why we need to correct the colors used by projector. Thus, we need to adjust the RGB ratio of each color to ideal values. The coordinates of object surface are computed with *corresponding* of camera and projector. We use the colors *Red*、*Green*、*Blue*、*Yellow*、*Cyan*、*Magent*a and *white*. Assign the number *1~7* to each color and use permutation coding to generate the pattern. Beam the pattern on object surface by projector and capture the picture by digital camera. Then obtain the 3D coordinate information of object from the corresponding of camera and projector. The pattern is composed with *Color Groups* which contents some color stripes. We need to decode the color groups by color stripes of scan line. Obtain the sequence of each color stripe of current color group and reach the Index of color group. The color stripes index number can be obtained by offset value with the color group index. Then transfer the index numbers of stripes to angles. Index numbers of stripes are implied in each color group of pattern. This information can be obtained by pattern decoding. In order to get the information, we use the *Hue* value to determine the number of color. Hue detection is better than RGB approach because Hue value doesn’t include the Luminance information. That means we can reduce the effect of *intensity* that reflects from non-uniform object surface. Pattern coding by permutation also can reduce the *Error Rate* when color stripes lost in a single color group. We can say it provides the *Fault Tolerance* ability and reduces the overhead of computing. The result of resolution and performance of Permutation Coding is better than other approaches in the same strategies. We increase 46% resolution which is more than De Brujin 3D scanning of Spatial Neighborhood.

Keywords: Structured Light, 3D Scanning, Facial Scanning, Image Calibration, Color Calibration,3D coordinate, 3D Reconstruction.

目錄

1. 前言

許多現實生活的各種物體皆可以以數位的3D資訊來表現，而將實際物體掃描成為數位資訊的方法為 3D 取像技術，目前 3D 取像的應用越來越廣泛，而針對不同的掃描物件都有不同的掃描方式以及技術背景,掃描的方式與人類辨識物體遠近原理是相同的，主要是利用兩眼視線的影像，根據物體的特徵以及焦距來判定物體的距離。

目前3D影像的運用相當廣泛，所以研究簡易方便以及準確度高的投影方式是主要的目標,3D 取像的應用如: 辨識[1]、醫療[2]、工商業模型建立[3]、娛樂、珍貴文物的數位化、3D 顯示的資料來源建立、動作的擷取等，因此未來在立體取像的需求也會日益提高，其中結構光的立體取像方式為較簡易且建構成本較低，取像過程的速度較快，且可重建原物體之原始表面材質的特性。

2. 研究目的

由於立體取像與顯像的應用越來越多,再不同產業也漸漸導入立體取像的技術，為了能夠滿足較平價的硬體價格，也能夠提供不錯的掃描解析度，這樣的方式能夠讓立體取像的技術更普及，也更能方便取得與使用，綜合各方面的評估，使用結構光的取像方式是較好的方式，在硬體的成本上是最能夠普及,解析度的表現上也能夠符合目前的使用需求,使用上也較為便利，另外結構光的運用範圍也比較廣泛,除了靜態的物體掃描，若搭配適合的演算法也能夠做動態的物件掃瞄，或是動作的立體取像等等，這些是其他掃描方式較不易達到的目標。

所以這個研究的計畫是建立在結構光掃描技術,加強結構光掃描過程中所使用的圖樣設計，圖樣的設計對結構光的立體取像有相當關鍵的影響，目前結構光主常用的類型可分為兩類，一類為使用單一張照片作為座標計算的來源，圖樣則需加入空間編碼的設計，讓計算時能夠辨識相對位置來計算座標，另一類的圖樣則是以時間編碼[4]為依據，在不同時間投影不同的圖樣，藉由不同時間圖樣的變化來辨識座標，然而單一照片使用空間編碼，在運算的計算技術上有較多的改善空間，也較使用時間編碼的方式複雜，但空間編碼的演算法也能夠套用在時間編碼的取像上，因此空間編碼的結構光取樣是本計畫的研究重點。

3. 文獻探討

所謂結構光(Structured Light) 3D取像, 即是由一台投影機以及一台照相機所組成的系統，主要用意是利用投影機來製造特徵圖像，而照相機再取得圖像辨識特徵，經由辨識特徵以及投出已知的特徵資訊，依照已知的各項資訊來計算出精

確的物體表面座標。

3.1 結構光(structured light)取像的原理

利用投影機投出已知的圖像，使影像在欲取得的物體表面呈像，並且使用照相機取得物體表面之影像，藉由投影與取像的行經光線角度計算出表面每一個點像的位置，如圖 1 所示

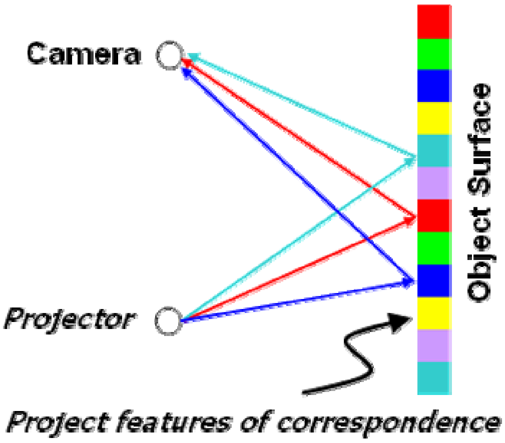


圖 1 ： 結構光取像原理

結構光的取像類型可分成三個主要類型，分時投影(Time-multiplexing)、空間區域(Spatial Neighborhood)、與指向編碼(Direct Coding)，不同的類型提供不同的掃描特性，也應用在不同的使用環境或是物件類型，這三個類型的掃描方式主要都是使用照相機與投影機做為硬體架構，並且使用三角函數的方式來進行座標的計算，差別在於使用圖樣(Pattern)的差異。

3.2 照相機影像校正(Image Calibration)

由於 3D 座標之計算需要照相機與投影機的同位(Corresponding)資訊，因此影像的正確性會影響取像的準確度，目前的照相機的鏡頭皆由球面鏡所組成，所以照相機的拍攝影像都有變相的問題存在，在拍攝影像時必須先校正照相機之影像，才能使用校正後的影像來進行座標計算，照相機的影像校正的方式是先量測影像的變形分布，在依照與實際物體的比對，將變形的區域修正回原始的影像，而影像校正的方式主要分成兩個類型，模型量測分析(Modeling Based)[11]與影像變形量測[12]:

4. 研究方法

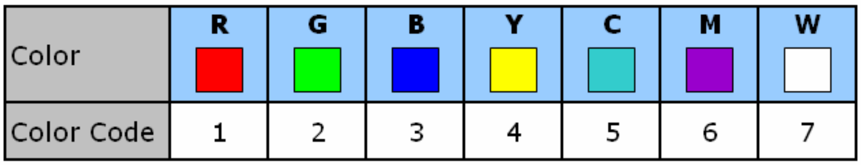
4.1 圖樣編碼設計

利用排列組合將 R、G、B、Y、C、M、W 七個顏色做編碼，其中W做為

區隔碼，也就是兩個W中間一定會有另外的六個顏色所組成的編碼，所以六個碼的排列組合共有720種組合，每個組合由七個顏色所構成。

定義:

將七個顏色編上一個數字作為代碼(如表格1 所示)，白色為群組分隔顏色，其餘的六個顏色在色調(Hue)[13]上剛好都是差距60度，也就是任一顏色距離鄰近的兩個顏色都是最大的間距，這方便我們在取像後分離顏色。

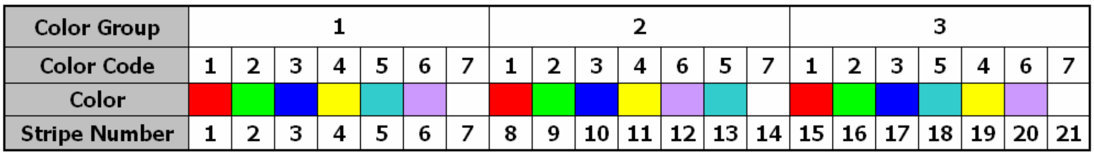


Color	R 🟥	G 🟩	B 🟦	Y 🟨	C ■	M 🟪	W ⬜
Color Code	1	2	3	4	5	6	7

表格 1 ： 顏色代碼表

產生編碼:

依照排列組合的方式，將顏色以代碼的方式排列，第一組為 123456，第二組為 123465，第三組為 123546 以此類推，以此方式編碼總共可編出 720 組數字，依照順序將第一組編號取出並且顯示對應的顏色，在最後加上白色，接著取出第二組編號，同樣顯示對應的顏色，在最後加上白色，如圖 2 所示，重複同樣的步驟直到顯示的點數達到預設的長度，例如解析度為 1024x768 的投影機，當繪製的長度到 1024 就可以停止，圖 3 為實際的圖樣影像的一部份。



Color Group	1							2							3						
Color Code	1	2	3	4	5	6	7	1	2	3	4	6	5	7	1	2	3	5	4	6	7
Color	🟥	🟩	🟦	🟨	■	🟪	⬜	🟥	🟩	🟦	🟨	🟪	■	⬜	🟥	🟩	🟦	■	🟨	🟪	⬜
Stripe Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

圖 2 ： 圖樣產生方式與代碼對應

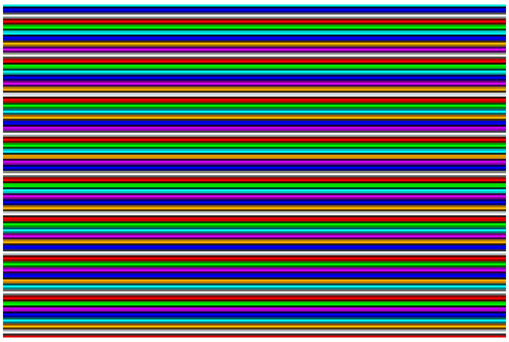
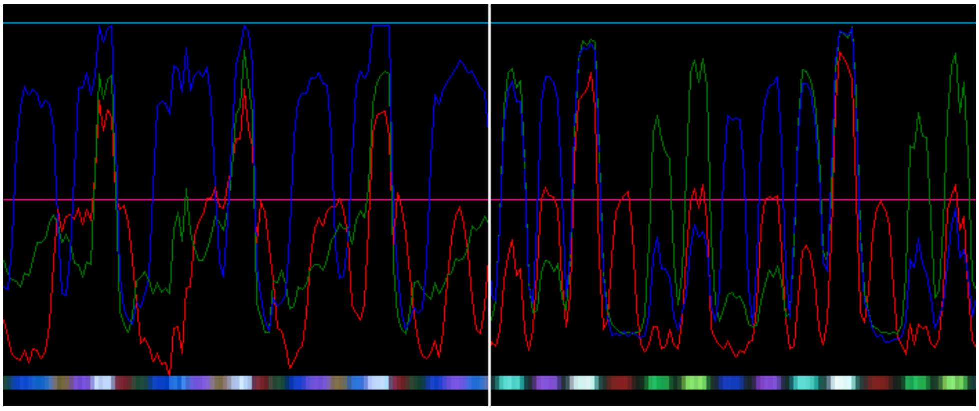


圖 3 ： 圖樣編碼的一部份

由於光線散射的問題，所以當光線投射到物體表面時會與週遭的光線混合，造成

溶色的結果，將會造成解碼時的困難，甚至無法解碼，在每個顏色中間必須插入一個黑色的空格，圖樣原本的色調並不會因為投影的散射而造成顏色的變化，另外加入黑色間隔較容易突顯每個顏色的分隔，也就是可以從亮度(Luminance)來判別顏色的峰值(Peak)，圖4(a) 中無法分辨每一個顏色的位置，(b)圖中可以明顯的判別每個顏色的位置與峰值



(a) (b)

圖 4 ： 圖樣有無黑色間隔線條的顏色分佈比較

顏色碼產生公式:

$$ColorCode(j) = \sum_{i=N-1}^{0} s(q(c(i,j),j))\, x10^{(N-i+1)}$$

（方程式 1）

N:使用的顏色數量(不包含白色)
j:第 j 組

$$q(x,j) = \begin{cases} N-1, \ j=N-1 \\ q=mod(q,x), \ otherwise \end{cases}$$

$$C(i,j) = \begin{cases} j/(i!) \ , \ i>0 \\ 0, \ i=0 \end{cases}$$

S (k): Get the k-th number from array.
If k = 2，取出”2” ，若下次取出時 k=2，取出的數字為”3” 。(圖 5)

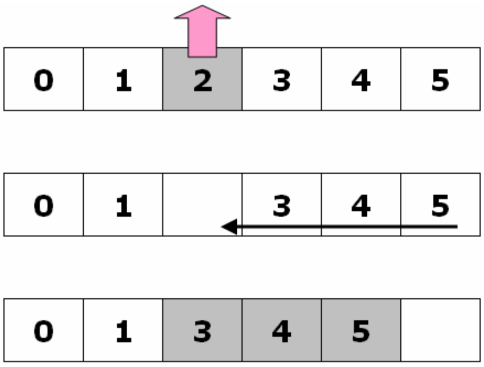
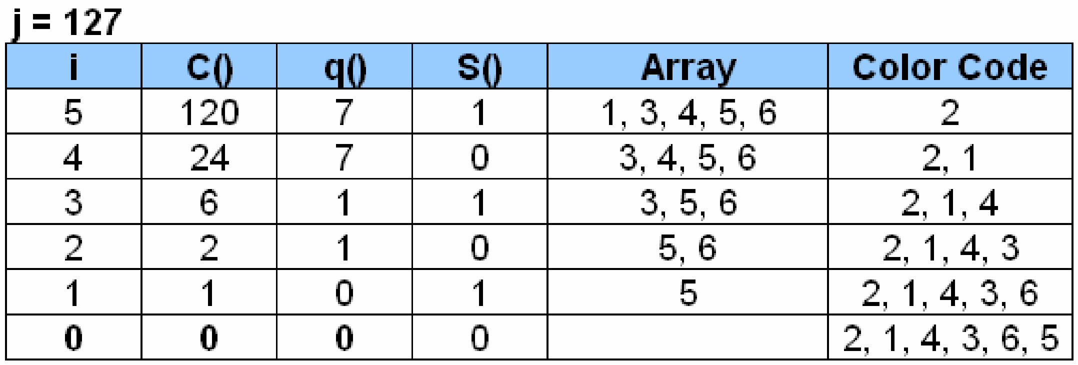


圖 5 ： 編號抓取示意圖

以 j = 127 為例，取得顏色碼的方式:



j = 127

i	C()	q()	S()	Array	Color Code
5	120	7	1	1, 3, 4, 5, 6	2
4	24	7	0	3, 4, 5, 6	2, 1
3	6	1	1	3, 5, 6	2, 1, 4
2	2	1	0	5, 6	2, 1, 4, 3
1	1	0	1	5	2, 1, 4, 3, 6
0	0	0	0		2, 1, 4, 3, 6, 5

表格 2 ： 取得顏色碼

由 Color Code 欄位的結果得知，j = 127 (顏色群組=127) 所計算出的顏色碼為 2, 1, 4, 3, 6, 5，因此以顏色編碼後的結果為”G R Y B M C” 。

4.2 距離與角度的轉換

座標計算式使用三角函數的方式計算，因此我們必須將投影機的投影解析度與相機的解析度換算成角度，轉換方式是將投影的面以及相機的圖片，將每一條水平線與垂直線的位置轉換成角度，以投影機的水平角度轉換為例: 若投影機的解析度為1024x768，所以水平方向由1024個點所組成，假設投影畫面的寬度為W，點與點之間的距離interval=W/1024，如圖31所示，投影機距離被投影平面的距離為distance，則角度計算方式如下:

$$A(i) = \text{atan}\left(\frac{\text{interval x i}}{\text{distance}}\right) \qquad \text{(方程式 2)}$$

Ai: 第 i 點的角度

中心點右側與左側為鏡射關係，所以將右側角度鏡射至左側並乘上負號。

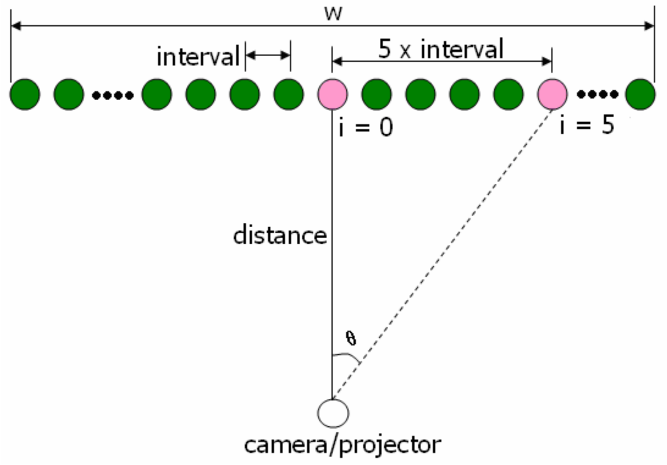
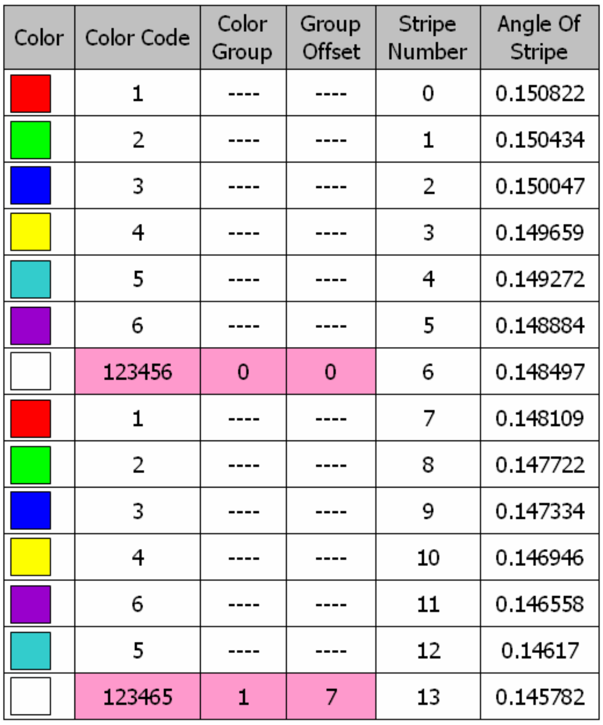


圖 1 ： 角度轉換方式

在計算座標之前必須先計算的角度有(1) 照相機影像之垂直線的水平角度(2) 照相機影像之水平線的垂直角度 (3) 投影機投出每條顏色線的垂直角度，這些資訊在座標計算中使用。

4.3 取像與解碼

在取得相機擷取影像後，從上而下(或由左而右; 依照相機與投影機的擺放形式而定)讀取點像的資訊，並且依照編碼原則中的顏色代碼定義給予相對的代碼，如圖 6 所示，當讀取到白色資訊時，將之前的六個碼依照讀取順序排列，以圖 6 為例，讀取到顏色資訊為 123456，白色本身的資訊不納入計算，編碼是 123456 解碼後可得知這是第一組碼，而第一組碼的表現範圍為 1~7，依照顏色可以得知每一條線的順序，取得線條的順序後，可利用 3.1 (方程式 1)的概念來計算出每一條顏色線的投影角度。



Color	Color Code	Color Group	Group Offset	Stripe Number	Angle Of Stripe
🟥	1	----	----	0	0.150822
🟩	2	----	----	1	0.150434
🟦	3	----	----	2	0.150047
🟨	4	----	----	3	0.149659
▨	5	----	----	4	0.149272
🟪	6	----	----	5	0.148884
⬜	123456	0	0	6	0.148497
🟥	1	----	----	7	0.148109
🟩	2	----	----	8	0.147722
🟦	3	----	----	9	0.147334
🟨	4	----	----	10	0.146946
🟪	6	----	----	11	0.146558
▨	5	----	----	12	0.14617
⬜	123465	1	7	13	0.145782

圖 6 : 顏色碼與順序的對應

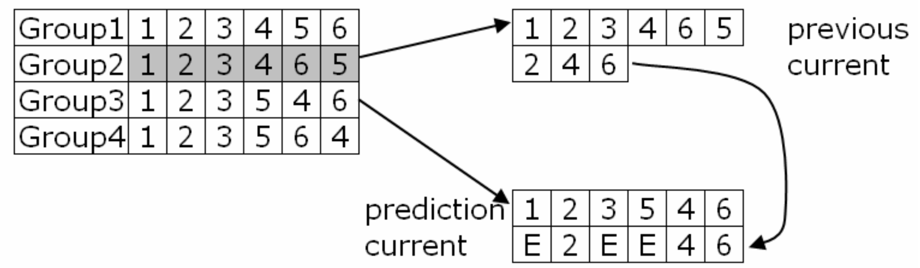
若取得的顏色順序為 ”GRYBMC” ，色碼為 2, 1, 4, 3, 6, 5，因可得知顏色群組為127，每個顏色群組由七個顏色構成，每個顏色中間加入黑色區隔，所以一個顏色群組總共包含14個點，127顏色群組的起始位置為 14 X 127 = 1778，顏色 G 所在位置為1778，顏色 R 的所在位置在 1780，以此類推可以得知此顏色群組每個顏色的順序，在使用(方程式 1)來取得角度資訊。

容錯能力:

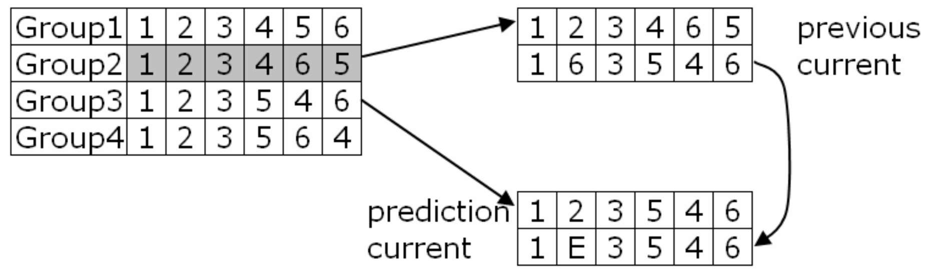
由於結構光是將顏色光頭影至物體表面，在使用照相機取得影像，但由於物件表面可能為非白色表面，或是反射率不均的情況，或是物體表面有不連續的狀況發生，以上這些變數都會造成顏色光投影至物體表面，卻無法正確取得影像的狀況，因此如何在資訊有遺失的情況下來計算，是十分重要的方法。

在3.1章節中本計畫使用排列組合的方式來產生圖樣，由此可知所有的線條的順序都是已知的，因此我們可以利用已知的線條資訊，再來預測有異常線條區域的資訊，如線條短少、多出現條、顏色錯誤等等問題。

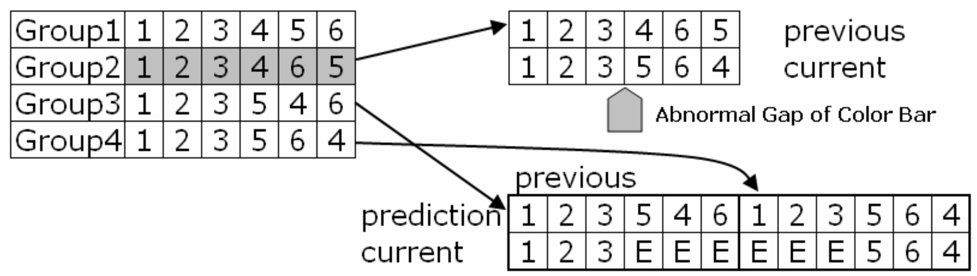
線條錯誤 1:



線條錯誤 2:



線條錯誤 3:



若有線條多出的問題，先依照重複編號線條的數值來比較，將相同編號數值較低的移除，移除後再依照上述的三種狀況進行資訊偵測。

4.4 座標計算

座標計算的概念如圖 33 所示，主要是利用照相機與投影機的已知條件，若加上物件表面的任一點，即構成一個三角形的區域，利用三角函數的方式來求取物件表面該點的座標。

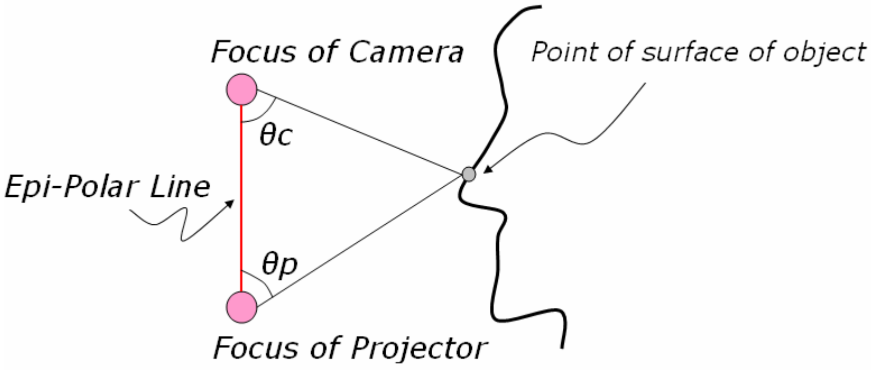


圖 7 : 座標計算的概念

θp 投影機夾角:

θp 為點 Pi 與投影機中心線的夾角，計算方式由圖樣解碼所取得，如圖 8 所示，拍攝影像後取得一垂直線的所有顏色資訊，以 3.3 的解碼方式進行解碼，並且取得顏色群組與其所包含所有顏色的資訊，在換算成角度資訊，每一垂直線中的所有顏色都轉換成角度資訊，而此角度資訊及為θp。

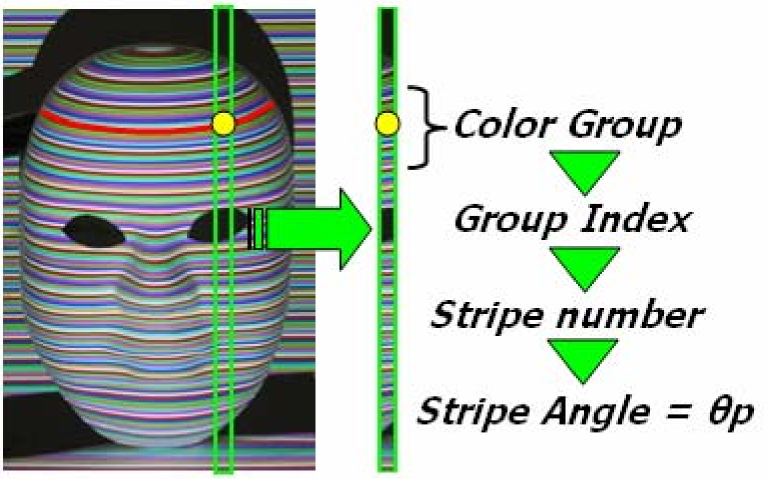


圖 8 : θp 投影機夾角

θc 照相機夾角:

θc 為點 Pi 與照相機的夾角(圖 9)，但θc 夾角不需要使用任何顏色資訊來計算，僅需要 Pi 在垂直線的位置(Index)即可以換算成角度資訊。所以在座標計算之前，會先計算出每個有效的θp 與其所在位置的θc 兩個角度。

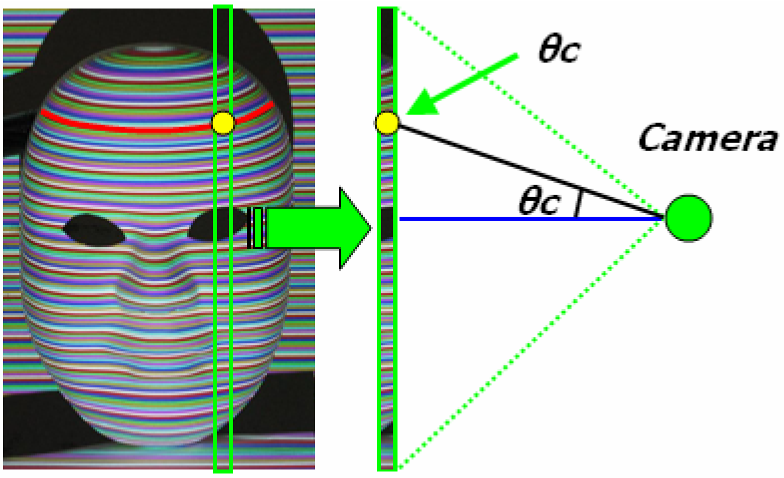


圖 9 : θc 照相機夾角

座標計算公式:

座標計算僅需要使用三角函式即可計算出點 Pi 的 3D 座標，以圖 10 左圖為例，而以幾何方式顯示為右圖，為計算方便重新定義角度位置如圖 11 所示，

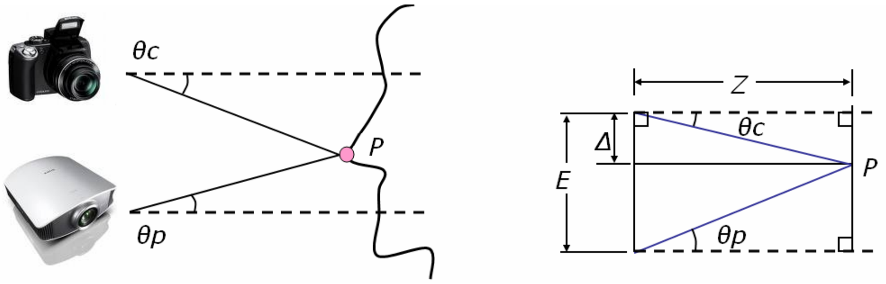


圖 10 ： 座標計算之示意圖

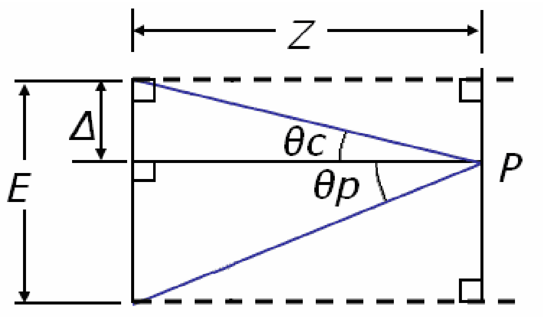


圖 11 ： 座標計算之示意圖

$$\tan\theta c = \frac{\Delta}{Z}$$

$$\tan\theta p = \frac{E\text{-}\Delta}{Z}$$

$$\frac{\Delta}{\tan\theta c} = \frac{E\text{-}\Delta}{\tan\theta p}$$ ， 結合上面兩個式子得到此式。

$$\Delta = \frac{E}{\frac{\tan\theta p}{\tan\theta c}+1}$$ （化簡後的結果）

其中θp與θc為已知資料，E(Epi-Polar)也是已知條件，因此可以求的Δ的數值，已知前定義來說，Camera 的中心線為原點，Δ則是等於 Y 軸的數值，所以此處的Δ數值及為-Y。則 Z 可由下列式子求出:

$$Z = \frac{\Delta}{\tan\theta c}$$

座標計算時依照待測點的位置，可分成三個區域，計算時需依照這三個區域的分類使用不同的計算公式，區域的分界限為 (1)照相機中央水平線以上區域

(2) 照相機中央水平線以下區域與投影機水平線以上的重疊區域 (3) 投影機水平線以下的區域，如圖 38 所示。三個區域使用相同的概念分別化簡公式，依照掃描點所在的位置，來判定此掃描點再哪一個區域，在套用對應的公式來計算座標。

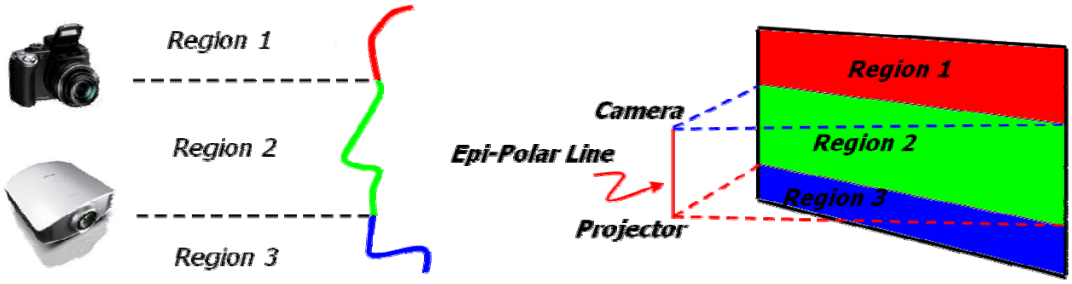


圖 12 ： 座標計算區域分界

第一區 (Region1):

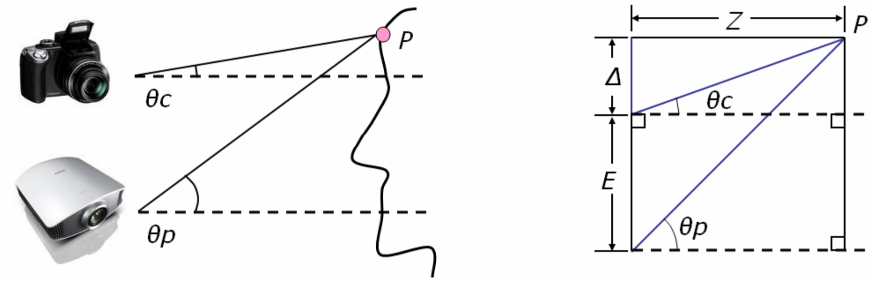


圖 13 ： 第一區座標計算範圍

條件: $\theta c >= 0$ and $\theta p > 0$

$\theta H = |\theta c|$，$\theta L = |\theta p|$，E=Epi-Polar（照相機與投影機的鏡頭內焦點的距離）

$$\Delta = \frac{E}{\frac{\tan(\theta_c)}{\tan(\theta_p)} - 1} \quad \text{(方程式 3)}$$

$$Z = \frac{\Delta}{\tan(\theta_c)} \quad \text{(方程式 4)}$$

$$Y = \Delta \quad \text{(方程式 5)}$$

$X = Z * \tan(\theta x)$; θx 為 P 點的水平角度

第二區 (Region2):

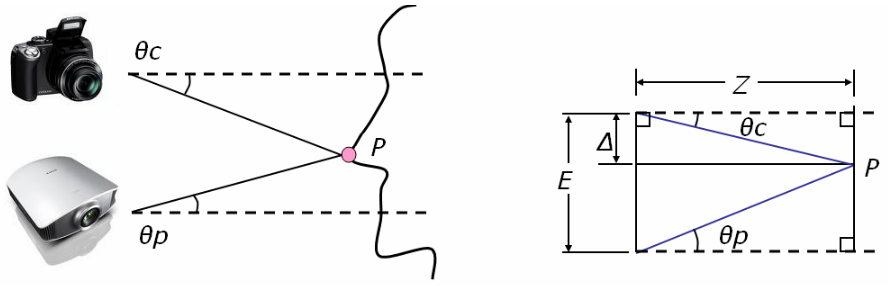


圖 14 ： 第二區座標計算範圍

條件: θc < 0 and θp > 0
θH = |θc|，θL=|θp|，E=Epi-Polar（照相機與投影機的鏡頭內焦點的距離）

$$\Delta = \frac{E}{\frac{\tan(\theta_p)}{\tan(\theta_c)} + 1}$$ （方程式 6）

$$Z = \frac{\Delta}{\tan(\theta_c)}$$ （方程式 7）

$$Y = -\Delta$$ （方程式 8）

$X = Z * \tan(\theta x)$; θx 為 P 點的水平角度

第三區 (Region3):

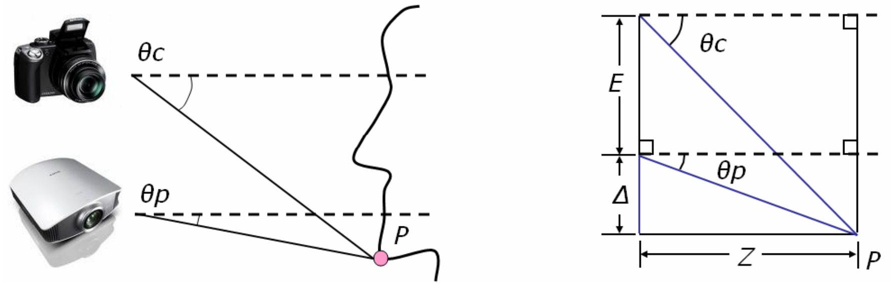


圖 15 ： 第三區座標計算範圍

條件: θc < 0 and θp <= 0
θH = |θc|，θL=|θp|，E=Epi-Polar（照相機與投影機的鏡頭內焦點的距離）

$$\Delta = \frac{E}{\frac{\tan(\theta_c)}{\tan(\theta_p)} - 1}$$ （方程式 9）

$$Z = \frac{\Delta}{\tan(\theta_p)}$$ （方程式 10）

$$Y = -(E+\Delta)$$ （方程式 11）

$$X = Z * \tan(\theta x)$$ （方程式 12）

θx 為P點的水平角度(如圖16所示)，Z 的數值於上述的算式中獲得，θx 可由照相機的影像中獲得(如圖17)，因此可由此公式獲得X的數值。

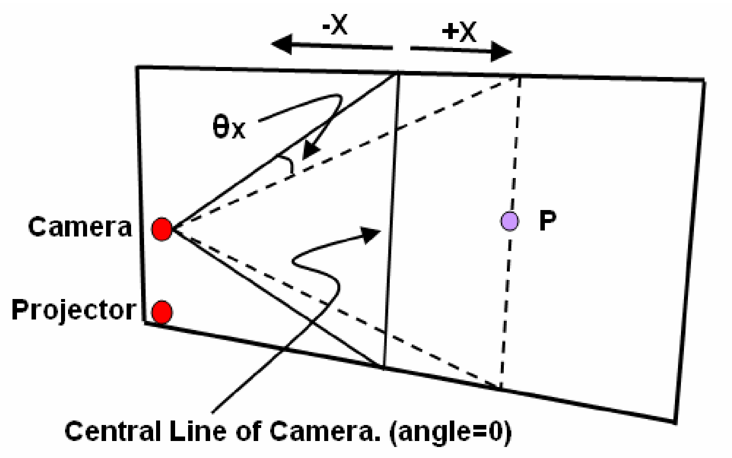


圖 16 ： 水平夾角示意圖

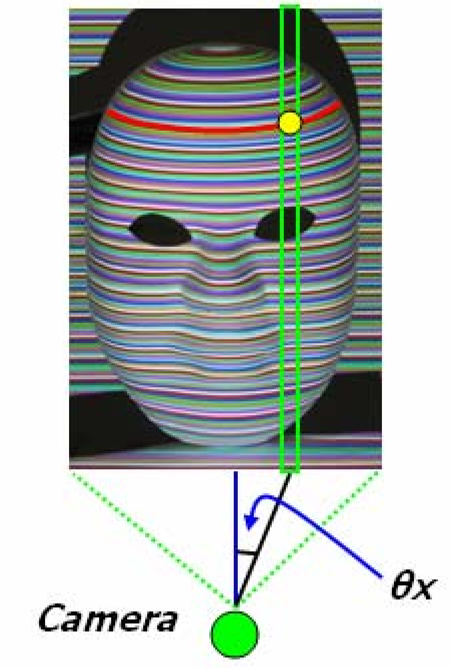
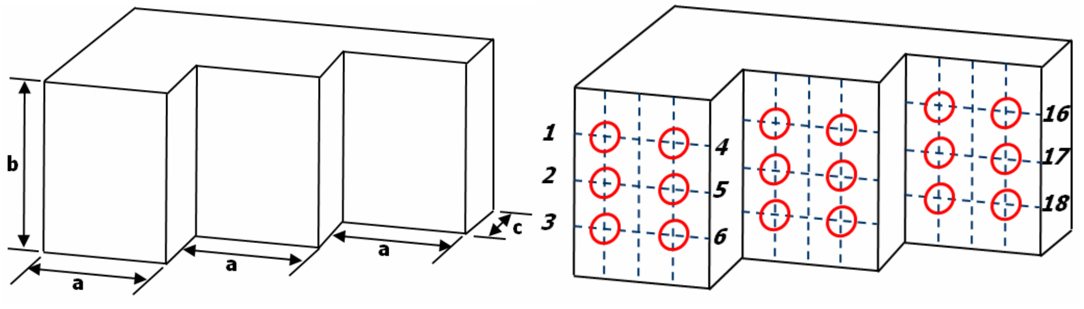


圖 17 ： 水平夾角取得方式

5. 結果與討論

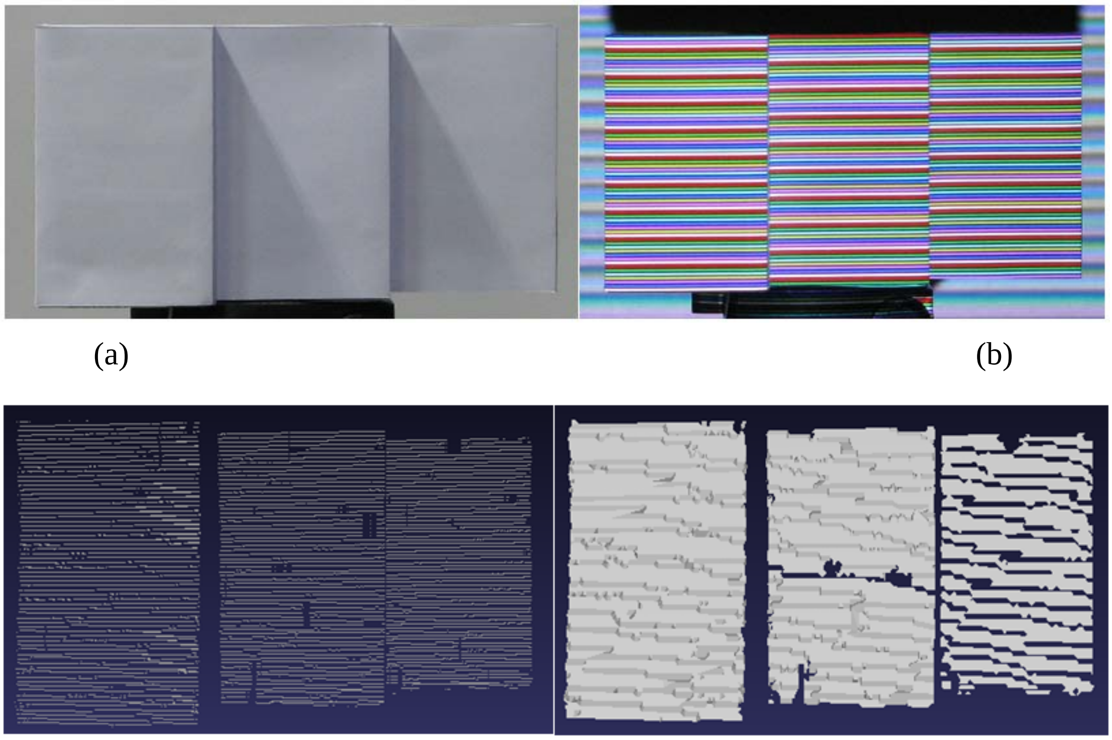
5.1 準確度掃描分析

在座標準確性的測試中，我們針對一階梯物件進行測試，此物件分成三個階層，物件的尺寸為圖 18(a)所示，其中 a = 56 (mm), b = 88 (mm), c = 23 (mm)，每個階層量測六個測試點，整個物件總共量測 18 個測試點(圖 18(b))。圖 19 為階梯物件之掃描結果。圖片中(a)為白光下物件之原始影像，(b)為圖樣以投影機投影於物體之影像，(c)為掃描後之 3D 座標影像，(d)以 3D 座標所建立網格後之影像。



(a) (b)

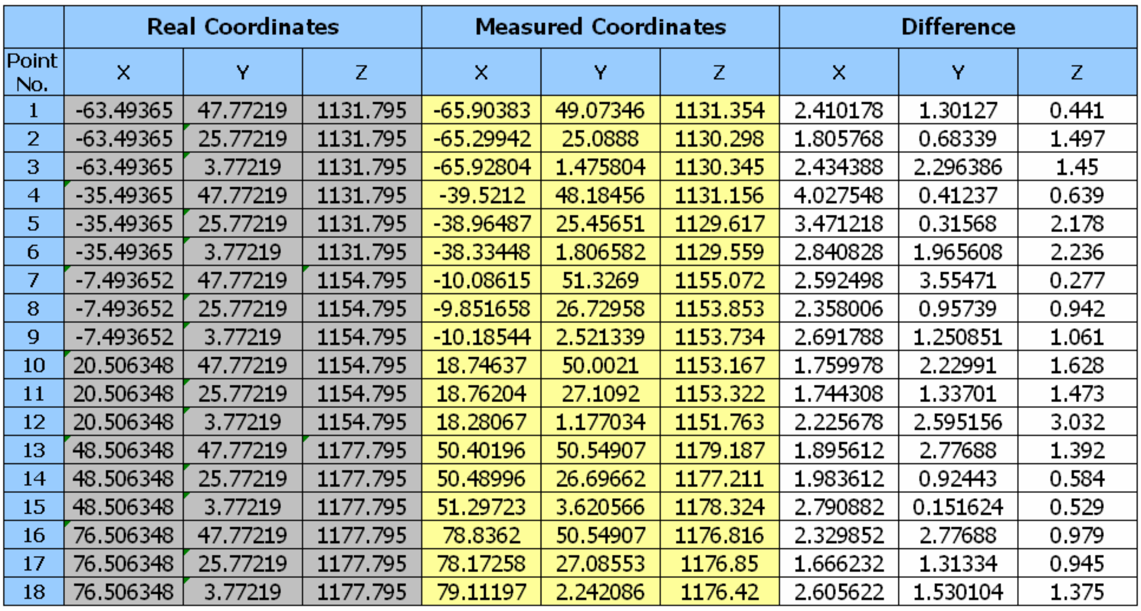
圖 18 : 階梯物件之尺寸與量測點



(a) (b)

(c) (d)

圖 19 : 階梯物件之掃描結果



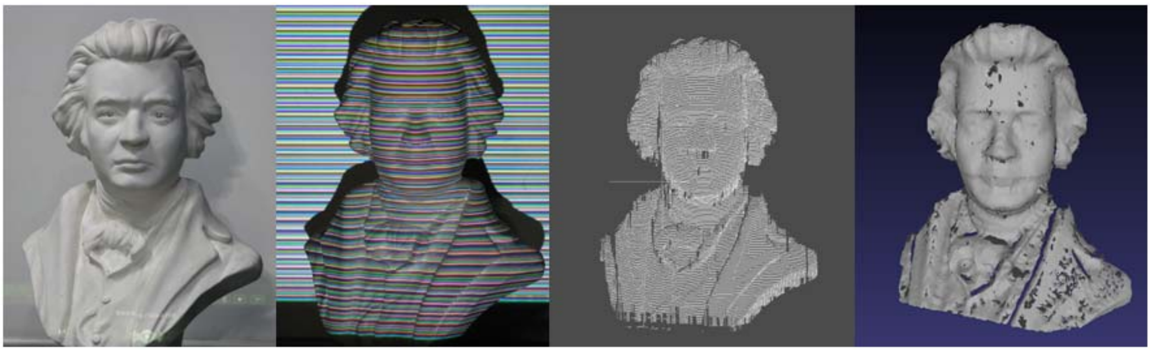
	Real Coordinates			Measured Coordinates			Difference		
Point No.	X	Y	Z	X	Y	Z	X	Y	Z
1	-63.49365	47.77219	1131.795	-65.90383	49.07346	1131.354	2.410178	1.30127	0.441
2	-63.49365	25.77219	1131.795	-65.29942	25.0888	1130.298	1.805768	0.68339	1.497
3	-63.49365	3.77219	1131.795	-65.92804	1.475804	1130.345	2.434388	2.296386	1.45
4	-35.49365	47.77219	1131.795	-39.5212	48.18456	1131.156	4.027548	0.41237	0.639
5	-35.49365	25.77219	1131.795	-38.96487	25.45651	1129.617	3.471218	0.31568	2.178
6	-35.49365	3.77219	1131.795	-38.33448	1.806582	1129.559	2.840828	1.965608	2.236
7	-7.493652	47.77219	1154.795	-10.08615	51.3269	1155.072	2.592498	3.55471	0.277
8	-7.493652	25.77219	1154.795	-9.851658	26.72958	1153.853	2.358006	0.95739	0.942
9	-7.493652	3.77219	1154.795	-10.18544	2.521339	1153.734	2.691788	1.250851	1.061
10	20.506348	47.77219	1154.795	18.74637	50.0021	1153.167	1.759978	2.22991	1.628
11	20.506348	25.77219	1154.795	18.76204	27.1092	1153.322	1.744308	1.33701	1.473
12	20.506348	3.77219	1154.795	18.28067	1.177034	1151.763	2.225678	2.595156	3.032
13	48.506348	47.77219	1177.795	50.40196	50.54907	1179.187	1.895612	2.77688	1.392
14	48.506348	25.77219	1177.795	50.48996	26.69662	1177.211	1.983612	0.92443	0.584
15	48.506348	3.77219	1177.795	51.29723	3.620566	1178.324	2.790882	0.151624	0.529
16	76.506348	47.77219	1177.795	78.8362	50.54907	1176.816	2.329852	2.77688	0.979
17	76.506348	25.77219	1177.795	78.17258	27.08553	1176.85	1.666232	1.31334	0.945
18	76.506348	3.77219	1177.795	79.11197	2.242086	1176.42	2.605622	1.530104	1.375

表格 3 : 掃描與實際物體座標之分析 (單位:mm)

量測實際物體與 3D 掃描之數據比較(表格 3),平均的誤差值分別為: X = 2.4241109, Y = 1.5762772 and Z = 1.2587778 (單位:mm),由於在量測 3D 物件座標時,是使用目視的方法來進行取樣,因此在量測值包含人為量測的誤差,所以實際的誤差值會比量測誤差更小。

5.2 形體之掃描分析

我們在實驗中針對石膏像以不同方向進行掃描,並且依照掃描的步驟進行比較與分析,石膏像的尺寸為: 385(mm) X 272(mm) X 618(mm),表面顏色為白色。掃描結果如圖: 20、21、22、23 所示,圖片中(a)為白光下物件之原始影像,(b)為圖樣以投影機投影於物體之影像,(c)為掃描後之 3D 座標影像,(d)以 3D 座標所建立網格後之影像。

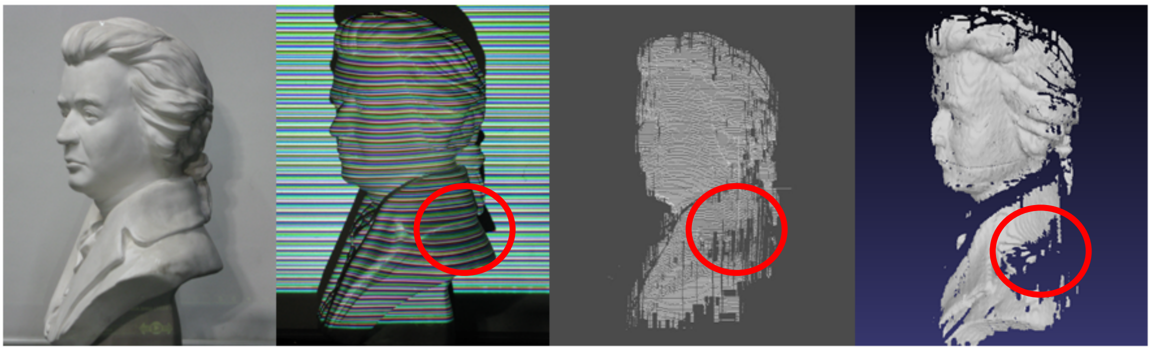


(a) (b) (c) (d)

圖 20 : 石膏像之正面掃描

圖 20 中有較佳的掃描結果,影像之完整度較高,唯有邊緣的部份因為資訊不足,

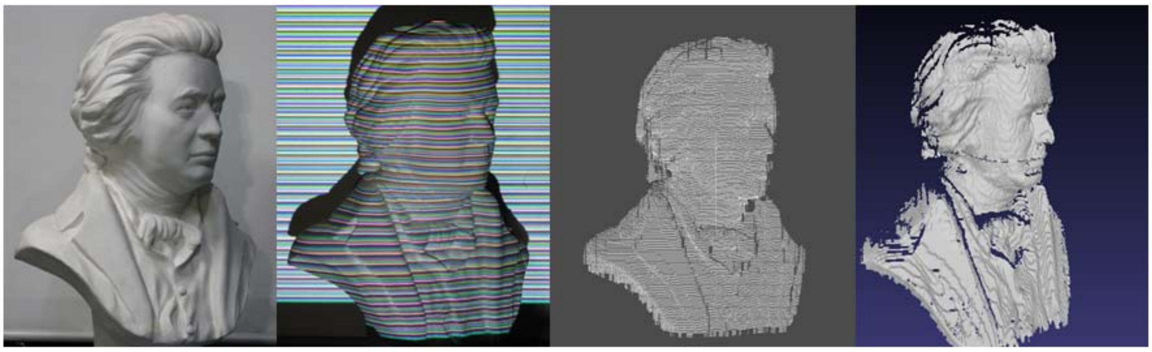
而造成影像的失真現象。



（a）　（b）　（c）　（d）

圖 21: 石膏像之左側掃描

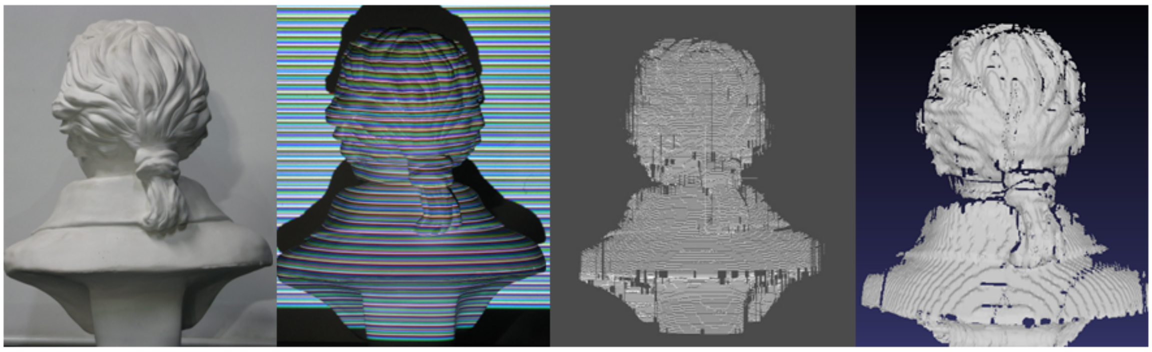
圖 21 圓圈部份，因為角度與陰影所產生色偏現象，深度變化較大時，投影在失焦情況下容易產生色偏，使得運用顏色解碼的效過降低。



（a）　（b）　（c）　（d）

圖 22 : 石膏像之右側掃描

圖 22 整個構圖完整，掃描結果與圖 20 相同，座標資訊也較完整。



（a）　（b）　（c）　（d）

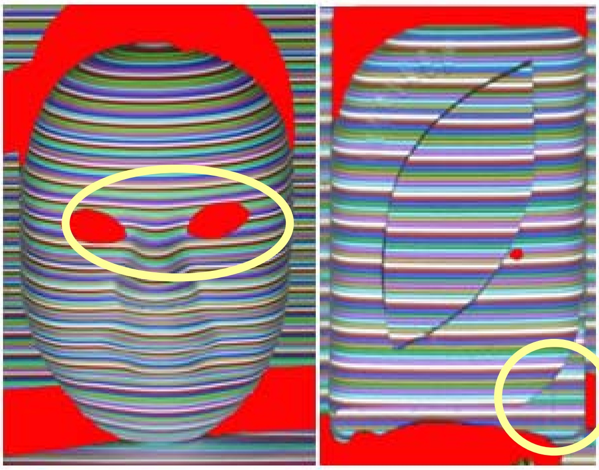
圖 23 : 石膏像之背部掃描

圖 23 有零散的小區域失真，整體掃描結果都有不錯的表現，頭髮部份的形狀變化性高，整個髮型的紋路變化都能夠正確掃描。

5.3 討論分析

5.3.1 非連續表面

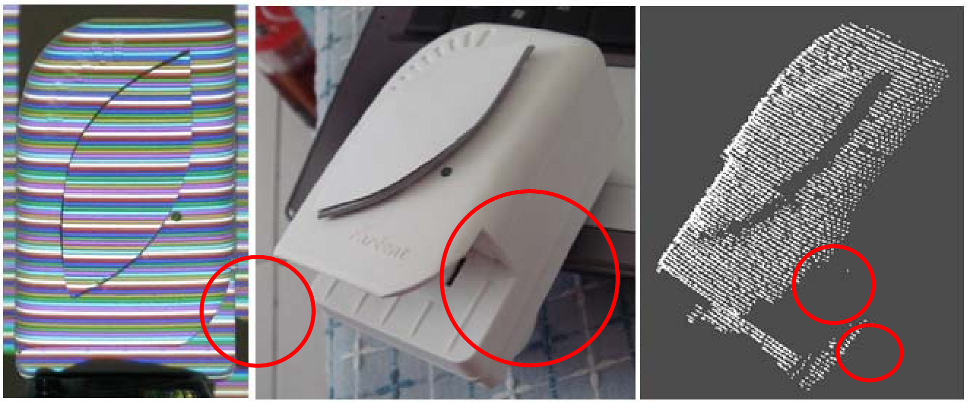
圖樣編碼的容錯能力設計，是針對解決不連續物體表面所造成的陰影，也就是掃描線中有空缺的部份(如圖57中紅色區域所示)，掃描線被中斷後，在下一段掃描線開始後，必須能夠正確的判別顏色群組，以及對應的角度資訊，而顏色為顏色群組的構成要素，若顏色缺少或是錯誤的比例過高，將造成顏色群組解碼失敗。



(a) (b)

圖 24：陰影與不連續表面

一般來說不連續的表面會造成陰影，如圖24(a)眼部，產生不連續掃瞄線，但是某些情況下，不連續物體表面仍然會產生連續物體表面，如圖24(b)，雖然有深度部分的差異，但是圖樣仍然可以投射在物體表面，這樣的情況將會是顏色群組解碼錯誤，而造成3D做標的計算錯誤，上方與下方邊緣處產生鋸齒現象，一部分的下方的資料被歸類至上方表面，或導致顏色群組異常而無法解碼，下方物體表面存在一部份的空白，如圖25(c)所示。



(a) (b) (c)

圖 25：非連續物體表面

5.3.2 物體表面顏色的影響:

由於我們是使用彩色圖樣的取像方式，也就是投出彩色圖樣的影像，若物體表面已經存在某些顏色，將會導致使用顏色解碼的問題，狀況就是我們可能讀取到非圖樣所產生的顏色，或是改變圖樣本身的顏色，這兩種情況都會造成顏色群組中顏色的遺失，而造成顏色群組解碼的失敗(圖 26)。

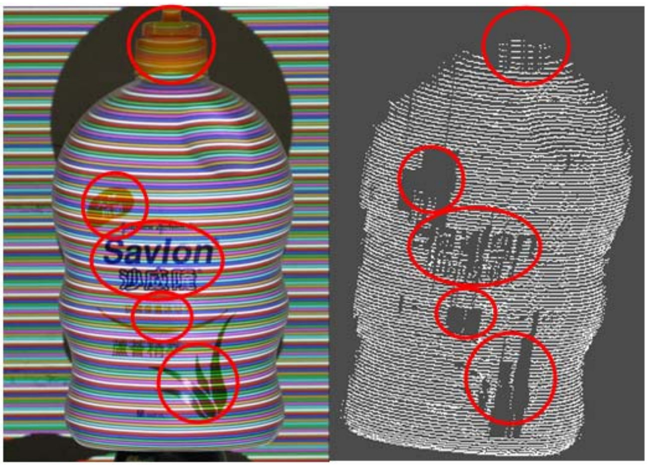


圖 26 : 物體表面顏色的影響

5.4 結論與未來展望

本計畫提供了圖樣的編碼方式以及投影機顏色校正的概念,在圖樣的編碼除了可以降低運算的時間，也提供了一定能力的錯誤容忍能力，使得錯誤率降低，相對的以提高了掃描的解析度，以往的研究中，有些使用 Dynamic Programming [15][16]的方式來解決解碼的問題，這樣的方式在解碼的速度上有較差的成效，也消耗較多的計算時間，當然也有其他研究提出類似的群組概念，例如 De Brujin[17]編碼方式，但又無法提供較好的錯誤容忍力，本計畫使用的編碼方式，除了能夠提供較高解析度的圖樣，在解碼的計算上也將低了運算時間，使得解碼的複雜度為 O = CN, 在顏色校正方面,以往許多研究中都一樣使用色調[18]來來區分顏色以及編碼,但是都未提出類似的概念,使用顏色校正後能夠提高顏色的準確率，也使得解碼的過程能夠有較好的判別結果。

目前物件 3D 掃描系統已經能夠正確的掃描出物體的表面座標，但是在影像變形校正[19]以及顏色校正還是以手動的方式來進行校正,希望未來能夠完成這兩項功能的自動化，目前影像變形校正，是經由實際測量變形狀況，將數據填入計算公式中，在校正後觀察變化情況，雖然可以實踐影像變形校正的工作，但是花費較多的時間與不便，投影機顏色校正也是取像後，再依照顏色的分布逐一調整 RGB 的數值,也是花費較多的時間與較差的效益,因此影像變形的自動化與顏色校正的自動化會是未來可以研究的方向。

另外由於編碼的方式，解碼時也必須依照編碼的特性分群，也就是說解碼時的最小可辨識的群組必須是七個顏色的掃描線組成，若物體表面無法讓足夠的掃描線投影，會造成無法辨識的區域產生，因此如何改善編碼的方式也是另一個主要的方向，例如每一單色掃描線加入其他資訊等，使得包含的資訊更多，能夠在較小的面積上取得足夠的資訊，在利用這些資訊來計算出物體表面左座標。

參考文獻

[1] C. Beumier, 3D Face Recognition, *IEEE International Conference on Industrial Technology 2006*, ICIT 2006, Mumbai, India, Dec 2006.

[2] Chia-Hsiang Wu, Yung-Nien Sun, Chien-Chen Chang, “Three-Dimension Modeling from Endscopic Video Using Geometric Constraints Vis Feature Position”, *IEEE Transactions on Biomedical Engineering. July 2007*.

[3] Jianfeng Li, Youngkang Guo, Jianhua Zhu, Xiangdi Lin, Yao Xin, Kailiang Duan, Qing Tang, “Large depth-of-view protable tree-dimensional laser scanner and its segmental calibration for robot vision”. *Optics and Lasers in Engineering 2007.*

[4] C. Rocchini, P.Cignoni, C.Montani,P. Pingi and R.Scopigno, “A low cost 3D scanner based on structured light, *EUROGAPHICS 2001*.

[5] J. L. Posdamer, M. D. Altschuler, Surface measurement by space-encoded projected beam systems, Computer Graphics and Image Processing 18 (1) (1982) 1–17.

[6] D. Caspi, N. Kiryati, J. Shamir, Range imaging with adaptive color structured light, Pattern analysis and machine intelligence 20 (5) (1998) 470–480.

[7] E. Horn, N. Kiryati, Toward optimal structured light patterns, Image and Vision Computing 17 (2) (1999) 87–97.

[8] D. Bergmann, New approach for automatic surface reconstruction with coded light, in: Proceedings of Remote Sensing and Reconstruction for Three-Dimensional Objects and Scenes, Vol. 2572, SPIE, 1995, pp. 2–9.

[9] M. Maruyama, S. Abe, Range sensing by projecting multiple slits with random cuts, Pattern Analysis and Machine Intelligence 15 (6) (1993) 647–651.

[10] P. Vuylsteke, A. Oosterlinck, Range image acquisition with a single binary-encoded light pattern, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (2) (1990) 148–163.

[11] Peter Eisert, "Model-based camera calibration using analysis by synthesis techniques," in Proc. 7th International Workshop MODELING, AND VISUALIZATION 2002, November 2002, pp. 307-314, Erlangen, Germany.

[12] R. Tsai, A versatile camera calibration technique for high accuracy 3D machine

vision metrology using off-the-shelf TV cameras and lenses, IEEE Journal of Robotics and Automation **RA-3** (1987), no. 4.

[13] Tehrani, M.A.; Saghaeian, A.; Mohajerani, O.R. “A New Approach to 3D Modeling Using Structured Light Pattern”, *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008.*

[14] J. Salvi, J. Pag`es, and J. Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition, 37(4):827–849, 2004.*

[15] David B. Wagner, "Dynamic programming," *The Mathematica Journal, 1995.*

[16] L. Zhang, B. Curless, and S. M. Seitz, “Rapid shape acquisition using color structured light and multi-pass dynamic programming,” *in The 1st IEEE International Symposium on 3D Data Processing*, Visualization, and Transmission, pp. 24–36, June 2002.

[17] H. Fredricksen, A survey of full length nonlinear shift register cycle algorithms, Society of Industrial and Applied Mathematics Review 24 (2) (1982) 195–221.

[18] Fechteler,P. “Fast and High Resolution 3D Face Scanning”, *IEEE International Conference. Image Processing, 2007.*

[19] Cui Haihua, Dai Ning, Yuan Tianran, Cheng Xiaosheng, Lio Wenhe, “Calibration Alogrithm for Structured Light 3D Vision Mesuring System”, *2008 Congress on Image and Signal Processing.*

[20] Molinier, Thierry; Fofi, David; Salvi, Joaquim; Gorria, Patrick,"2D virtual texture on 3D real object with coded structured light", *Image Processing: Machine Vision Applications. Proceedings of the SPIE*, Volume 6813, pp. 68130Q-68130Q-9 (2008).

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

附件三

98 年 9 月 18 日

報告人姓名	鄭芳炫	服務機構及職稱	中華大學資工系教授
時間 會議地點	2009/9/11~2009/9/15 日本京都	本會核定補助文號	NSC98-2221-E-216-031
會議名稱	(中文) 第五屆智慧型資訊隱藏及多媒體訊號處理國際研討會 (英文) The Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP-2009)		
發表論文題目	（中文）以動作識別為基礎的人類行為描述模型 （英文）Human Behavior Description Model Based on Action Recognition		

一、參加會議經過

本次會議為避免會議當天人數太多，特別安排在前一天可先行註冊。本人 9 月 10 日搭華航早上飛機至日本關西機場已是當地中午時間，隨即坐利木津巴士到京都，進住飯店時已是下午 4 點。稍做休息後隨即前往京都車站附近的會議地點註冊(見圖一)，註冊處負責服務的同學大都是大陸去的留學生，其中只有一位是台灣去的留學生，簡單閒聊了一下後，領取了會議資料完成註冊手續後即離開會場回飯店休息。

メルパルク KYOTO
The Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing
IEEE
ストラン円山 個室にての宴会プラン
ご予約・お問い合わせ
075-361-1233
メルパルク KYOTO

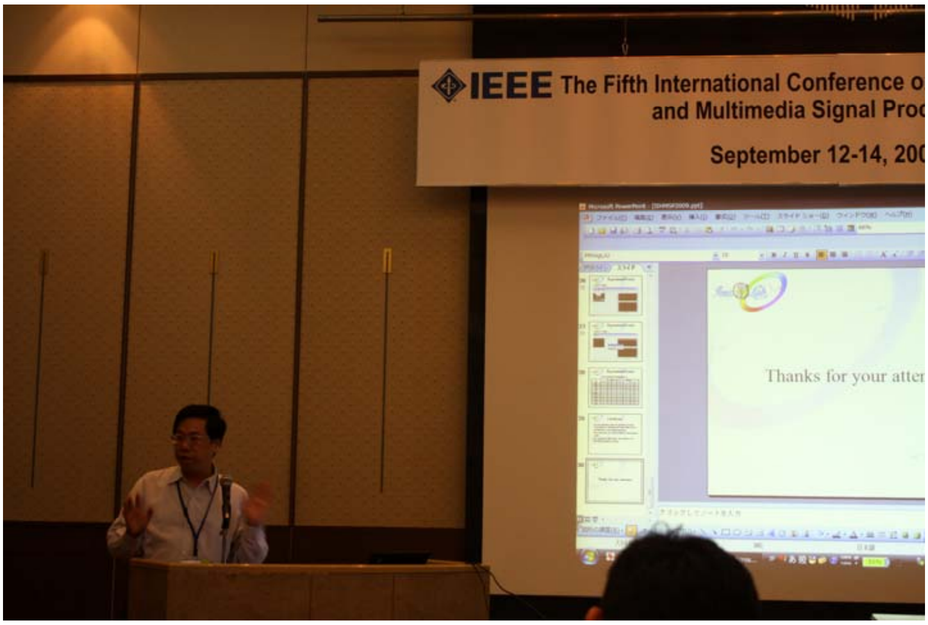
圖一

會議的開幕典禮由主辦單位與會議的委員會主席簡單的致歡迎詞後，隨即展開(見圖二)。由於本會議為一個大型的研討會,因此共安排了五個場地同時進行。本屆會議並未受到新流感的影響參加人數很多，共有超過 25 個國家 410 篇論文投稿，每篇論文都經過二位評審嚴格的審查後，最後通過 326 篇論文。大會安排了三天的會議議程,安排方式是每天上午先有一場專題演講再加一場論文口頭報告分五個場地同時進行，下午則是有二場論文口頭報告亦分五個場地同時進行。

IEEE The Fifth International Conference on Intelligent Information Hiding
and Multimedia Signal Processing (IIH-MSP2009)
September 12-14, 2009, Kyoto, Japan
IIHMSP-2009
Welcome to Kyoto, Japan

圖二

三天會議安排了三個場次之專題演講分別是第一天由日本京都大學的副校長Takashi Matsuyama 教授演講，主題是 The State of the Art of 3D Video Technologies，第二天由韓國大學的 Hyoung Joong Kim 教授演講，主題是 Data Compression by Data Hiding，第三天由中國大陸浙江大學 Jianrong Tan 教授演講，主題是 Multimodal Information Fusion in the Virtual Environment and Its Applications in Product Design。這三位教授在各自領域均學有所成，演講內容亦十分精彩，尤其是京都大學的 Takashi Matsuyama 教授的演講內容與本人目前國科會計畫的研究內容有相關，因此可說是收獲良多。詳細論文報告場次規劃可參考議程表。本人之論文『Human Behavior Description Model Based on Action Recognition』被安排在第一天的上午 section A01 之場次發表，如圖三所示。本次會議尚有許多台灣之其他論文發表,經過三天完整會議研討,與會者均有豐富的收穫。此外會議主辦單位特別在第一天晚上安排了歡迎酒會及第二天晚上安排晚宴，使與會者之間有互相交流的機會。


IEEE The Fifth International Conference o
and Multimedia Signal Proc
September 12-14, 200
Thanks for your atter

圖三

二、與會心得

此次會議的主辦單位為 Ritsumeikan University, 會議地點安排在京都車站附近的會議中心,而不是在校園內。這樣的安排應該是考慮到來自世界各國與會者的方便,不過若是能安排在校園內, 不僅在經費上可較為節省,應該更能感受學術交流的氣息。本會議雖是大型的研討會, 但定位上仍是以專業精緻之研討會自許, 與一般大雜燴式之大型研討會不同。主要目的是讓與會之學者能真正達到充份的學術交流, 而不是走馬看花。三天的會議安排得十分緊湊, 每天都是從上午 9:00 至下午 6:20 止。本會議在第二天晚上安排的晚宴中除了報告此次會議的相關數據外也頒發了最佳論文獎, 同時亦宣佈下一屆在德國舉辦。

國際研討會是學術研究交流很好的場合, 可結合全世界相同研究領域的學者互相切磋。主辦單位除了安排專題演講及論文發表的議程外, 若能安排半天的行程到會議主辦大學做參訪, 應該更能達到交流的目的。

三、考察參觀活動(無是項活動者省略)

本會議的定位是專業精緻之研討會, 三天的會議安排得十分緊湊, 每天都是從上午 9:00 至下午 6:20 止, 因此並無時間做考察參觀活動。

四、建議

每次參加研討會常常會在會場碰到許多台灣去的學者教授, 若在出國前就可以互相聯繫一起出席, 不僅在費用上可以比較節省, 在會議上也可以整合力量為台灣之學術界出聲讓國際能充份了解台灣在學術領域之實力。本次會議共有超過 25 個國家的研究學者參加, 台灣大概有幾十位教授及學生參加, 除本校中華大學外, 尚有政治大學、中正大學、元智大學、文化大學、長榮大學、台科大、雲科大、高應大、崑山科大、龍華科大、元培科大等。也許國科會可以在現有之網站上另闢一個出席國際會議之交流園地, 讓國內之研究學者可以互通訊息, 不僅可以整合大家的力量, 也可知道國內在國際學術界之活動能量。

五、攜回資料名稱及內容

本次會議攜回一本紙本的會議導引手冊, 資料名稱為 IIH-MSP-2009 Conference Guide Book。另有一片資料光碟為本次會議論文集之光碟版。

六、其他

2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

HUMAN BEHAVIOR DESCRIPTION MODEL BASED ON ACTION RECOGNITION

Fang-Hsuan Cheng and Cheng-Yuan Chang

Dept. of Computer Science and Information Engineering
Chung Hua University
E-mail: fhcheng@chu.edu.tw

ABSTRACT

Based on human action recognition, this paper proposes a new description model to record human behavior. The paper makes use of action recognition result and regards time information accumulated in action recognition as features, records human actions and time spent in these actions, then identifies events through action combination and gives effective processing toward these identified events. In order to prove feasibility of human behavior description model, we take events produced when pedestrians pass through cross-road as example. Under cross-road context in the experiment, total 60 films are shot when five pedestrians are passing through cross-road, producing 191 events. 187 events are correctly detected in the experiment with correct rate of 98%.

1. INTRODUCTION

With science and technology gaining progress, more and more intelligent surveillance systems use digital media to store image data, whereas these image materials are not effectively treated on the first time spot of sudden accidence occurring, and the happening events of video characters are only known after people are required to review films. Regarding this, we propose a description model of human behavior to identify human action, and detect events through action combination. When event is identified, it will be given with most instant and correct processing. However, even though it solves the general problem of behavior recognition system in action recognition, we still have no way to understand action and behavior modes happening on human, thus it is expected to achieve preventative protection and emergence event processing in future.

System framework of this paper is described in Fig. 1. We summarize a complete behavior analysis framework, which is divided into object detecting, object tracking, action recognition, human behavior description model and event detecting, event recording, event control and processing. It is wished to introduce event prediction in new methods in future. This paper follows methods proposed by our lab. in the first three parts and introduces new characteristics value and new action in action recognition. After action recognition is made, the processing procedure will adopt a model framework proposed by this paper and describe event occurring in human, thus make decision identification toward processing mechanism.

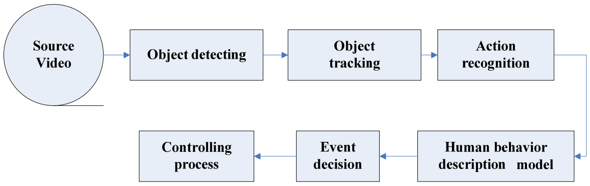


Fig.1 System framework diagram

2. RELATED WORK

Current techniques are mainly based on action recognition. In the recent years, research and relevant articles on events gradually emerge. Articles relating to events are currently under establishment. Therefore, this paper proposes a human behavior description model, taking pedestrian passing through cross-road as example, make whole and experimental analysis toward event detection, and thus validate the probability of this description model. This paper mainly integrates multiple actions to provide reference for event detection. Time is very essential in action recognition, thus the method is divided into time-unrelated and time-related in the research areas.

Time-unrelated method:

Thomas Kleinberger [1] proposes a certain conditions required in ambient intelligence based home care systems (AHCS), makes discussion toward

978-0-7695-3762-7/09 $26.00 © 2009 IEEE
DOI 10.1109/IIH-MSP.2009.104

possibly occurring recognition problems and gives possible solution and answers. It is advantageous to make full discussion toward problems occurring in most surveillance system. However, it is disadvantageous that data in practical experience is lacking and details of various aspects could not be integrated in systematic way. PIRUELA, J.A. Mr. [2] proposes a verification system, which uses shape comparison to surveillance sensitive area when being intruded. The advantages lie in that design of recognition result could be relatively easier within limited surveillance range .The disadvantages lie in that it is more easily affected by foregoing detected objects through shape comparison, influencing the whole correct rate.

Time-related method:

The following researches introduce conceptions of action and time thus strengthens results of behavior recognition. Yun Yuan and Shaohai Hu [3] propose that human behavior analysis method is adopted to detect moving objects and uses anthropomorphous characteristics to recognize human action, and finally display the recognized action and time spot of action occurring. Its advantages lie in that introduction of action and time spot of action occurring makes people understand what action occurs in character. However, disadvantages lie in that relevant connection between actions is unavailable, thus human behavior is unknown. Juang and Chang [4] propose a calculus method of neural fuzzy network to distinguish human action, which is applied in home security to detect actions of people falling down or lying down for rest. It is advantageous to put time consideration into action, thus falling down and lying down for rest could be distinguished. But it also has disadvantaged that time is put into consideration in single action. If times of multiple actions are introduced, recognition will be more correct. Caroline Rougier and Jacqueline Rousseau [5] propose a new way to detect event of human falling down. When old people stay at home alone, automatic security surveillance is made by combing motion history image (MHI) and change in the human shape. Its advantage lies in that many falling down actions are more identified accurately by using relevance between actions. Its disadvantage is an overall module framework is lacking for behavior recognition toward a single action. Naresh P. Cuntoor and Rama Chellappa [6] propose to adopt hidden Markov model (HMM) to decide occurrence probability of each event. Occurrence probability of each event is detected from action and provides information for training sample. Occurrence probability of events could be any time spot. In case two events occur at the same time, similarity between these two events could be calculated. If it is a high similarity, they could be judged to be the same event. It is advantageous to adopt HMM in training event. If occurring event has a high similarity with event in training sample, the occurring event is categorized to be the same as training event. Its disadvantages lie in that some events show different event meanings in different time spot, even though the same action is made, thus it could increase training time and cause training failure due to complexity of events. Gaitanis Kosta and Macq Benoît [7] propose a new method of recognizing human behavior, which mainly transforms hidden markov model (HMM) into multi-agent hidden markov model (M-AHMEM) and uses actions in different parts of human to decide human behavior in ranking way. Its advantages lie in that multiple actions are promoted to be combined and increase recognition rate of single action. Its disadvantages lie in that we know which action is made by human rather than effectively understand which event occurs in human.

Thus this paper proposes a description model of human behavior to simulate human behavior, and makes further knowledge about which event occurs, and then establishes twelve different and meaningful events under context of pedestrian passing through cross-road to validate probability of this model.

3. HUMAN BEHAVIOR DESCRIPTION MODEL AND EVENT DECISION

We propose a new human behavior description model which model framework is illustrated in Fig. 2. This description model is applicable to any human behavior. From Fig. 2, it is known that human behavior model is mainly consist of input action, action queue and event conditions. In this paper, human actions is input in human behavior description model, and each action and occurrence time of action is input and stored in action queue. When action recognition finishes, information in action queue is transferred to event conditions for event decision. Different event is decided by putting characteristics values into event conditions. In case occurring behavior of human conforms to these conditions, satisfied event conditions are triggered; Event conditions include decision formula in each situation. When event occurs, the decision condition of event is satisfied. This paper shows event decided in event conditions.

In this paper, when human makes action, time of accumulating actions will be recorded. When all event conditions are satisfied, an event will emerge. One occurring event is based on multiple combinations of actions, whereas similar action possibly shows different meaning. Therefore, this paper aims at finding what event we need to recognize, take occurring action of human and accumulating action time as parameter, and thus understand what actions are made by human in an event. We also record occurring situation through action combination, based on action recognition result, record spending time of individual i-action as (T_i). We record time of producing all actions by human in a period and record it as T , seen in Eq. (1):

$$T = \sum_{1}^{n} T_n \tag{1}$$

Method of storing action queue in Fig. 2 is seen in Eq.(2), among which action queue collects all recognized actions of human in frame with total number of i. After colon, how much time is experienced from start to end of i-action moving is recorded. Null means no action occur or no action yet to be recognized in this paper. Therefore, when action doesn't occur or no action is to be recognized, sustaining time of Null is still recorded, seen in Eq. (3).

$$AQ = \{Action\_1 : T_1, Action\_2 : T_2, ..., Action\_i : T_i\} \tag{2}$$

$$AQ = \{Null\_1 : T_1, Action\_2 : T_2, ..., Action\_i : T_i\} \tag{3}$$

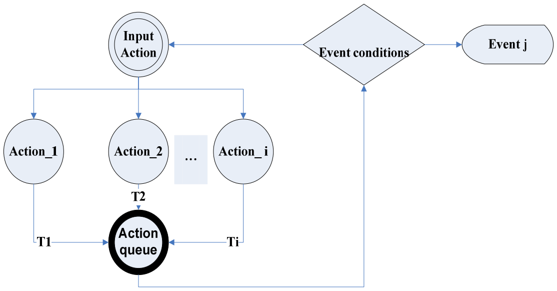


Fig. 2 Human Behavior Description model

In order to help understanding functions of this model, we provide a simple example for explanation. This paper takes walking actions made by pedestrian under yellow traffic light when passing through cross-road as example. Pedestrian will walk for about 2 seconds, and fall down in the middle of crossing road for 9 seconds. We actually put this example in our model. Fig. 3 put the following characteristics in event conditions, including action queue, emergency time and pedestrian location Cx_i。

Action queue (AQ) record occurring action of human and occurrence time of action, and deliver them in event conditions. Emergency time (ET) is recorded from time of falling down to recovering walking or sounds alarm due to time of falling down lasts too long. We use C to represent gravity center location of pedestrian, Cx represents coordinates in X axis of gravity center, Cx_1, Cx_2 represents coordinates in X axis of gravity center when pedestrian stays in different time.

We use a step-by-step procedure to make people understand the whole procedure, seen as follows:

Step1:
When pedestrian passes through road and enters in frame, we could know walking action of human through action recognition.

Step2:
We transfer the walking action and time of accumulating actions to action queue.

Step3:
We deliver action and time information of actions accumulated in action queue to event condition for decision. At this time, pedestrian only walks 2 seconds conforming to condition c rather than a and b condition, thus step 4 is returned.

Step4:
If not conforming to event condition, we still return state of inputting action recognition.

Step5: Action recognition system detects action of falling down of pedestrian; AQ=FD.

Step6:
Lasting time of falling down is recorded. Falling down for 9 seconds of pedestrian is transferred to action queue.

Step7:
At this time, AQ=FD, Cx_1, Cx_2>0, ET>8sec, decision formula a, b and c is valid, and skip to Step8 .

Step8:
If conforming to condition of decision formula, it is known that pedestrian falls down and lies down for 9 seconds, event 12 will be output.

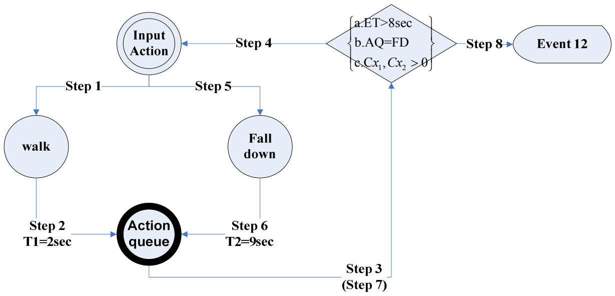


Fig. 3 Human Behavior Description model (example)

4. EXPERIMENTAL RESULT

In order to validate feasibility of human behavior description model proposed in this paper, we make an experiment toward pedestrian who are passing through cross-road which is shot by DV and regarded as the experimental scene. There are totally 60 films for five different pedestrians shot in the experiment, which include 191 events based on predefined 12 kinds of event conditions. Undetected amount of events and wrongly decided amount of events are calculated for each kind of event condition. In Table 1, first column shows twelve events defined by system. Second column shows amount of each event occurring. Third column divides detection amount into two kinds-miss amount means undetected events in event films; error amount means events wrongly regarded for others in event films. Last column is divided into miss rate, error rate and correct rate. Correct event recognition rate in the experiment is 98% and error rate is 2%. Through the human behavior description model, most event decision is conducted and human behavior could be effectively and correctly recognized.

Table1 Event recognition result and recognition rate

Event case	Event amount	Detection Rate		Recognition Rate		
		Miss	Error	Miss%	Error%	Correct%
Event1	21	0	0	0%	0%	100%
Event2	15	0	0	0%	0%	100%
Event3	15	0	2	0%	14%	86%
Event4	12	0	2	0%	17%	83%
Event5	20	0	0	0%	0%	100%
Event6	5	0	0	0%	0%	100%
Event7	9	0	0	0%	0%	100%
Event8	11	0	0	0%	0%	100%
Event9	11	0	0	0%	0%	100%
Event10	14	0	0	0%	0%	100%
Event11	27	0	0	0%	0%	100%
Event12	31	0	0	0%	0%	100%
Summary	191	0	4	0%	2%	98%

5. CONCLUSION

A whole intelligent surveillance system is consist of following sections, such as object detecting , object tracking, action recognition, human behavior description model and event detecting, event record, event control processing, event prediction. In the processing following action recognition, this paper proposes a human behavior description model to describe events occurring to human and exerts as decision and processing mechanism, finally shows excellent experimental achievement. As long as each action is combined into meaningful event at premise, behavior description model in this paper could be flexibly applied in various event decisions, finally realize goal of behavior analysis.

Based on a human behavior description model, this paper could correctly and instantly decide human behavior. When human behavior produces a meaningful event, this system could make effective and correct recognition. The paper takes pedestrian passing through cross-road as example. Correct rate in the experiment is 98% and error rate is 2%. Through the human behavior description model, most event decision is conducted and human behavior could be effectively and correctly recognized. Additionally, in order to validate model feasibility, the experiment includes events occurring when five pedestrian cross road, which is divided into emergency behavior and general behavior. With regard to emergency event, an instant control of traffic light is conducted, so as to protect security of pedestrian； General behavior is mainly used to protect security of walking pedestrian and remind pedestrian for safety.

ACKNOWLEDGEMENT

The authors would like to express their sincere thanks to National Science Council of Taiwan for her partial financial support to this research with grand no. NSC96-2221-E-216-039-MY2.

REFERENCES

[1] Becker, Martin; Werkman, Ewoud, “Approaching Ambient Intelligent Home Care Systems”, Digital Object Identifier, Page(s):1 – 10, 2006.

[2] Pinuela, J.A., Amador-Reyes, A.; Andina, D., “DOMOIIC SECUHIIY SYSTEM RASED ON HUMAN BEHAVIOR ANAI.YSIS”, World Automation Congress, 2004. Proceedings Volume 17, 28 June - 1 July Page(s):149 – 154, 2004.

[3] Yun Yuan, “Real-Time Human Behavior Recognition in Intelligent Environment”, Digital Object Identifier 10.1109/ICOSP.345793, 2006.

[4] Chia-Feng Juang; Chia-Ming Chang;, “Human Body Posture Classification by a Neural Fuzzy Network and Home Care System Application”, Digital Object Identifier 10.1109/TSMCA 897609, 2007.

[5] Rougier, Caroline; Meunier, Jean; “Fall Detection from Human Shape and Action History using Video Surveillance”, Digital Object Identifier 10.1109/AINAW.181.2007.

[6] Naresh P. Cuntoor, “Activity Modeling Using Event Probability Sequences”, Digital Object Identifier, Volume 17, Issue 4, Page(s):594 - 607, April 2008.

[7] Kosta, G.; Benoit, M., “Group Behavior Recognition for Gesture Analysis”, Circuits and Systems for Video Technology, IEEE Transactions on Volume 18,Issue 2, Page(s):211 - 222 Digital Object Identifier, Feb 2008.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

99 年 9 月 30 日

附件三

報告人姓名	鄭芳炫	服務機構及職稱	中華大學資工系教授
時間 會議地點	2010/9/21~2010/9/24 中國上海	本會核定補助文號	NSC98-2221-E-216-031
會議名稱	(中文) 第 11 屆太平洋區多媒體國際研討會 (英文) The 11-th Pacific Rim Conference on Multimedia (PCM-2010)		
發表論文題目	（中文）一個嶄新以 ASM 為基礎的二階段臉部地標偵測方法 （英文）A novel ASM-based two-stage facial landmark detection method		

一、參加會議經過

本次會議地點在上海，原本桃園上海就是熱門航線不容易訂機位，再加上適逢世博期間，更是雪上加霜，因此便訂了較冷門的航線桃園寧波，然後再搭陸路走杭州灣跨海大橋進上海。杭州灣跨海大橋全長 36 公里，目前是全世界最長的跨海大橋。雖然工程十分浩大，但完全是由中國獨立設計及建造完成，沒有借助國外的技術，顯見中國大陸在開革開放後的進步。我和系上另一位老師一同參加此會議，9 月 20 日出發當天因凡那比颱風侵襲大陸，造成回程飛機延飛，以致於我們的飛機也連帶被耽誤了時間，延後近二小時才飛。因此之故，行程中一直耽心會趕不上當天進住上海的旅館，所幸在一切順利的情況下安全的抵達。本次會議由上海復旦大學主辦，會議地點選在學校附近自營的復宣酒店內，因我們並不住在復宣酒店內，因此一早便搭地鐵出發再加上短程的計程車很順利的到達了會場報到，見圖一。

圖一

會議前後安排共四天，第一天為 tutorial，第二天到第四天則為論文發表。會議的開幕典禮由主辦單位與會議的委員會主席簡單的致歡迎詞後，隨即展開,見圖二。由於本會議為一個中型的研討會,因此只安排了二個場地同時進行。本居會議共有將近 20 個國家 261 篇論文投稿，每篇論文都經過二位評審嚴格的審查後，最後只通過 75 篇口頭發表論文(oral)及 56 篇海報論文(poster)，論文通過率只有 50%，算是十分嚴謹的。大會安排了三天的論文會議發表議程,安排方式是每天上午先有一場專題演講再加一場論文口頭報告分二個場地同時進行及一場海報論文發表，下午則是有二場論文口頭報告亦分二個場地同時進行及一場海報論文發表。

圖二

會議第一天Tutorial分成上下午各一個主題，分別由新加坡南洋理工大學Jianxin Wu教授主講Histogram Intersection Kernel Learning for Multimedia Applications及韓國光州理工學院Yo-Sung Ho教授主講MPEG Activities for 3D Video Coding。接下來三天會議亦安排了三個場次之專題演講分別是第一天由中國南京大學的Zhi-Hua Zhou教授演講，主題是A New Machine Learing Framework with Application to Image Annotation，第二天由亞洲微軟研發部門主管Yong Rui博士演講，主題是The Evolution of Image Search，第三天由亞洲微軟研發部門資深研究員 Tie-Yan Liu 博士演講，主題是 Learning to Rank : Pushing the Frontier of Web Search。這些主講教授在各自領域均學有所成，演講內容亦十分精彩，因此可說是收獲良多。詳細論文報告場次規劃可參考議程表。本人之論文『A Novel ASM-based Two-stage Facial Landmark Detection Method』被安排在第二天的下午13:30~15:10之場次發表，如圖三所示。本次會議尚有許多台灣之其他論文發表,經過四天完整會議研討,與會者均有豐富的收穫。此外會議主辦單位特別在第二天晚上安排了黃埔江夜遊及第三天晚上安排晚宴，使與會者之間有互相交流的機會。

A novel two-stage facial landma
detection by object detection a
active shape model

圖三

二、與會心得

此次會議的主辦單位為復旦大學，會議地點安排在學校附近的復宣酒店。本會議雖是中型的研討會,但定位上仍是以專業精緻之研討會自許，與一般大雜燴式之大型研討會不同。主要目的是讓與會之學者能真正達到充份的學術交流，而不是走馬看花。四天的會議安排得十分緊湊，每天都是從上午 9:00 至下午 5:10 止。國際研討會是學術研究交流很好的場合，可結合全世界相同研究領域的學者互相切磋。主辦單位除了安排專題

演講及論文發表的議程外，若能安排半天的行程到會議主辦大學做參訪，應該更能達到交流的目的。

三、考察參觀活動(無是項活動者省略)

本會議的定位是專業精緻之研討會，四天的會議安排得十分緊湊，每天都是從上午9:00 至下午 5:10 止，因此並無安排做考察參觀活動。只有在 9 月 22 日晚上有安排黃埔江夜遊，可看到漂亮的黃埔江夜景及租界區各國的建築，感受上海的風華。

四、建議

PCM 會議從 2000 年第一次在澳洲雪梨舉辦至今已十一屆，期間曾在北京,新竹,新加坡,東京,濟州,浙江,香港,台南,曼谷等處舉辦過,台灣在過去十屆中就舉辦過二次。每次參加研討會常常會在會場碰到許多台灣去的學者教授,不過這一次台灣與會者並不多,只遇到中研院,工研院與清華大學的學者。中國大陸近年來積極的主辦大型的國際研討會,但其中良莠不齊,常常有些研討會只要繳註冊費就好了,出不出席研討會不是很重要,而PCM 是一個專業的研討會,今年已是第十一屆,建議台灣的學者若選擇大陸舉辦的研討會,應選擇審查嚴謹且較有歷史的研討會如 PCM2010 即是。希望明年 PCM2011 在澳洲雪梨舉辦會有更多台灣的學者參予。

五、攜回資料名稱及內容

本次會議攜回二本紙本的會議論文集,資料名稱為 Advances in Multimedia Information Processing – PCM2010, LNCS 6297 及 6298。此次會議並沒有提供會議的資料光碟。此外，有提供第一天的二場 Tutorial 的投影片講義。另外在會議現場有其他研討會的宣傳資料，如 PCM2011 在澳洲雪梨, ICASSP 2012 在日本京都舉辦等。

六、其他

A Novel ASM-Based Two-Stage Facial Landmark Detection Method

Ting-Chia Hsu, Yea-Shuan Huang, and Fang-Hsuan Cheng

Computer Science & Information Engineering Department,
Chung-Hua University, Hsinchu, Taiwan

Abstract. The active shape model (ASM) has been successfully applied to locate facial landmarks. However, in some exaggerated facial expressions, such as surprise, laugh and provoked eyebrows, it is prone to make mistaken detection. To overcome this difficulty, we propose a two-stage facial landmark detection algorithm. In the first stage, we focus on detecting the individual salient corner-type facial landmarks by applying a commonly-used Adaboosting-based algorithm, and then further apply a global ASM to refine the positions of these landmarks iteratively. In the second stage, the individual detection results of the corner-type facial landmarks serve as the initial positions of active shape model which can be further iteratively refined by an ASM algorithm. Experimental results demonstrate that the proposed method can achieve very good performance in locating facial landmarks and it consistently and considerably outperforms the traditional ASM method.

Keywords: Active Shape Model, Facial Landmark Location.

Introduction

Facial feature extraction is a very popular research field in the recent years which is essential to various facial image analyses such as face recognition, facial expression recognition and facial animation. In general, based on different kinds of information extraction, the technology of facial feature extraction can be divided into two categories. First, local method, which is to detect local face components such as eye pupils, eye corners, mouth corners, etc. Secondly, global method, which makes use of the whole geometric structure of face components to locate the interested facial landmarks. In local method, because the feature models of facial landmarks are mutually independent, the detection result is easy to be affected by the variation of lighting and poses. In global method, because it uses a set of feature landmarks to form a global facial structure model, it usually has more ability to endure the detection error of individual landmark. Therefore, the global method generally obtains better performance in locating facial landmarks. At present, three kinds of the most commonly-used methods are deformable templates (DT) [1], active shape models (ASM) [2][3][4] and active appearance models (AAM) [5]. Both ASM and AAM are provided by Cootes, they iteratively decrease an energy function to obtain the optimized facial landmark locations.

G. Qiu et al. (Eds.): PCM 2010, Part II, LNCS 6298, pp. 526–537, 2010.

In recent years, ASM has been successfully applied to medical image analysis, such as computed tomography (CT), and it also can be applied to locating facial feature landmarks. However, the accuracy of the facial feature localization is still a problem because face images are much complex than medical images. Therefore, researchers keep on proposing new methods to improve its performance, such as Haar-wavelet ASM [6], SVMBASM [7] and ASM based on GA [8]. In general, these new methods have better accuracy than the original ASM, but they all are still prone to make mistaken detection in exaggerated facial expressions.

In this paper, we present a novel two-stage algorithm to improve the performance of facial landmark detection. The traditional method of ASM uses the average facial shape template to initialize the positions of facial landmarks, and it iteratively finds the best landmark positions only along the normal direction of edge contours. This process may contain two kinds of drawbacks. First, the average facial shape template may deviate considerably from the genuine landmark positions, therefore the landmarks are not able to be found correctly. Secondly, the genuine landmark position may not be located on the normal direction of edge contour, which will accordingly produce unsatisfactory landmark positions. Furthermore, when people have made exaggerated facial expressions, the traditional ASM often performs poorly because the shapes of exaggerated facial expressions usually are very different from the average facial shape. However, through analyzing the structure of human face compositions, we can understand the shape variation of human face mainly depends on the positions of the left/right eye inner and outer corners, the left/right inner and outer eyebrow corners and the left/right mouth corners. If these corner positions can be found correctly at the first stage, it will be able to set more approximate initial positions for the facial landmarks. Accordingly, better landmark locations can be found through ASM iteration and the accuracy of landmark detection can be much improved. From the above discussion, an improved landmark detection method is proposed which detects the corner-type landmark first, uses the detected corner-type landmarks to initialize the facial feature positions in the second stage, and then applies ASM to obtain the final landmark positions.

This paper is organized as follows. Section 2 introduces the classical ASM method and Section 3 describes the proposed two-stage ASM. Experimental results are given in Section 4, and finally, conclusions are drawn in Section 5.

2 Review of the Active Shape Model (ASM)

ASM is one of statistical models, which contains a global shape model and a lot of local feature models. Section 2.1 decides the shape model; Section 2.2 describes the local feature models and Section 2.3 describes the ASM algorithm.

2.1 The Shape Model

Suppose there are n facial feature points and each one is located at obvious face contour. The positions of these n points are arranged into a shape vector X, that is

$$X = [\, x_1, y_1, x_2, y_2, \dots, x_k, y_k, \dots, x_n, y_n \,]^T \tag{1}$$

where x_k and y_k are the horizontal coordinate and the vertical coordinate of the kth feature point respectively.

Using the PCA operation, the eigenvectors of the covariance matrix corresponding to main shape variations can be generated. Then, a shape model can be represented as:

$$x = \bar{X} + Pb \tag{2}$$

where $\bar{X}$ is the mean shape model, $P = [\Phi_1\ \Phi_2\ ...\ \Phi_t]$ is the eigenvectors corresponding to the t largest eigenvalues, and b is the shape parameter which is the projection coefficient that X projects onto P. Usually, b_i is constrained within the range of $\pm 3\sqrt{\lambda_i}$, so that a constructed face shape will not degenerate too much.

2.2 The Feature Model

In general, we suppose a landmark is located on a strong edge. According to the normal direction of a landmark, we can get m pixels on both sides of this landmark. So, for each landmark, there are in total 2m+1 gray-level values which form a gray-level profile $\mathrm{g_i} = [\mathrm{g_{i0}}, \mathrm{g_{i1}}, ..., \mathrm{g_{i(2m)}}]$, where i is the landmark index. In order to capture the frequency information, the first derivative of profile $\mathrm{dg_i}$ is calculated as

$$dg_i = [g_{i1} - g_{i0}\ , g_{i2} - g_{i1}, ..., g_{i(2m)} - g_{i(2m-1)}]. \tag{3}$$

In order to lessen the influence of image illumination and contrast, dg_i is normalized as

$$y_i = \frac{dg_i}{\sum_{k=0}^{2m-1} |dg_{ik}|}, \text{ where } dg_{ik} = g_{i(k+1)} - g_{ik}. \tag{4}$$

The feature vector y_i is called “grayscale profile”.

2.3 ASM Algorithm

The ASM searching algorithm uses an iteration process to find the best landmarks which can be summarized as follows:

1. Initialize the shape parameters b to zero (the mean shape).
2. Generate the shape model point using the $x = \bar{X} + Pb$.
3. Find the best landmark z by using the feature model.
4. Calculate the parameters b$'$ as $b' = P^T(z - \bar{X})$.
5. Restrict parameter b$'$ to be within $\pm 3\sqrt{\lambda_i}$.

If $|b' - b|$ is less than a threshold value, then the matching process is completed; else $b = b'$, and return to step 2.

3 The Proposed Method

The traditional ASM only uses the grayscale profile as its feature model. However, the grayscale profile is inherently a one-dimensional feature which in general is too simple to represent the distinct information of a landmark point. Basically, there are two other drawbacks of the traditional ASM. The first is that it selects the target points only from the candidates along the normal direction of edge contour. If the target point is not located in the candidate points, it will cause the found target point incorrect. The second is it uses a fixed search range for different landmark points. But, different landmarks in fact may require different search ranges because they have different variation extents.

In general, the facial feature landmarks can be attributed into two types: corner-type landmarks and edge-type landmarks. The corner-type landmarks (such as the left/right eye inner/outer corners) have very unique 2-D shapes looked like corners, but the edge-type landmarks (such as the landmarks of eyelid or mouth lip) have non-unique 1-D shapes shown as a line. Fig. 1 shows some examples of corner-type landmarks and edge-type landmarks. Obviously, the corner-type landmarks are much easier to detect than the edge-type landmarks. Therefore, in this paper we propose a novel two-stage facial landmark detection algorithm. The first stage is to locate the corner-type landmarks, and the second stage is to locate the whole facial landmarks by using the locations of the detected corner-type landmarks in the first stage as the initial positions of ASM. In this study, we define a total of 10 corner-type landmarks which are the left/right eye inner and outer corners, the left/right eyebrow inner and outer corners, and the left/right mouth corners. Another difference of our method to the traditional ASM is to define variable search ranges for different edge-type landmarks according to their variation degrees. That is if from the training data the positions of an edge-type landmark differ a lot, the search range of this landmark will be accordingly large. On the contrary, if the positions of an edge-type landmark are very stable, the corresponding search range will be relatively small. The proposed method will be introduced in the following.

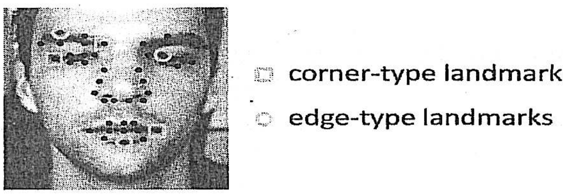


Fig. 1. Examples of the corner-type landmarks and the edge-type landmarks, where '○' denotes corner-type landmarks and '□' denotes edge-type landmarks

3.1 The First Stage

Adaboosting algorithms have been extensively used for object detection and they often obtain outstanding detection performance. Therefore, for each corner-type landmark we used the Adaboosting algorithm [9] to construct a detector in the first stage. Samples of the 10 corner-type landmarks are shown in Fig. 2 in which the black spots indicate the corner representative positions and they are not necessary to be located at the center of the image blocks. In order to improve the issue on search range, we defined different search ranges for different corner-type landmarks in Fig. 3.

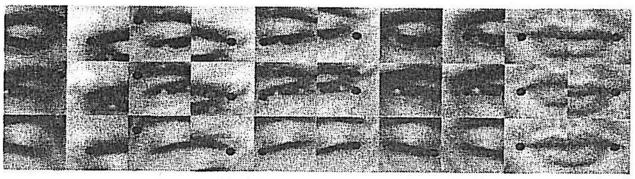


Fig. 2. Image samples of corner-type landmarks

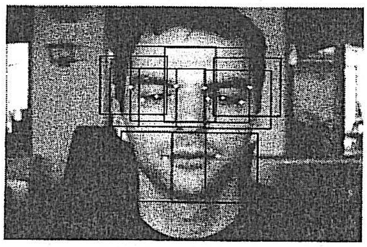


Fig. 3. The two-dimensional search ranges of 10 corner-type landmarks

With a constructed Adaboost-based detector, it may obtain several candidates of one landmark in the defined search range, and accordingly it is necessary to select the correct one among them. Because different corner-type landmarks are located at different facial geometric compositions (i.e. eyebrow, eye and mouth), their candidate selections should be designed according to their own affecting external factors (such as hair and glasses). With the above understanding, we categorized the 10 facial landmarks into three groups (eyebrow, eye and mouth) based on their functions and their geometric positions. Let e(x,y) be the edge strength of pixel (x,y) and s(x,y) be the detection score of a specific Adaboost-based corner-type landmark detector. Then, each group has its own candidate selection design as described in below.

3.1.1 Candidate Selection of Eyebrow Corners

Conceptually, an eyebrow corner should have a strong horizontal edge strength and a weak vertical edge strength. But, because the eyebrow may be covered by hair, just using the edge strength cannot get good candidate selection result. Instead, a HOG (Histogram of Oriented Gradients) [12] feature is also used to select the eyebrow corners. Therefore, in order to select the correct candidate, three factors are taken into consideration as

$$F_{eyebrow}(x,y) = \alpha\, log\, s(x,y) + \beta\, log\, e(x,y) + \gamma\, log(\frac{1}{h(x,y)}) \tag{5}$$

where α, β and γ are three weight parameters, s is the detection score of the Adaboost-based eyebrow detector, e is the edge strength and h is the Mahalanobis distance of the HOG features between the corresponding eyebrow model and the eyebrow candidate at pixel (x,y). Among the eyebrow corner candidates, the one having the largest $F_{eyebrow}$ is the selected candidate.

3.1.2 Candidate Selection of Eye Corners

Because an eye corner and its near pupil present a rather stable distance, this property can be used to select the eye corners. Since our face detection algorithm can detect not only face positions but also both pupil positions, both eye corners accordingly can be roughly estimated from the detected pupil positions. Among the eye corner candidates, the one closest to its estimated eye corner is selected. Fig. 4 displays an example of eye corner selection.

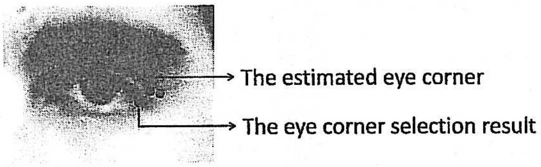


Fig. 4. An example of eye corner selection, where '●' denotes an eye corner candidate, and '▲' denotes the estimated eye corner

3.1.3 Candidate Selection of Mouth Corner

Because the mouth corner candidates usually are located either at the correct mouth corners or at the facial wrinkle corners with medium-large edge strengths, it will be ineffective if the edge strength is used to select the correct mouth corner candidate. However, the two kinds of candidates have very different variances. In general, a true mouth corner has a larger variance than a wrinkle corner does. With this understanding, $F_{mouth}(x,y)$ is designed to reflect the possibility that a candidate is truly a mouth corner as $F_{mouth}(x,y) = \mu \, log \, s(x,y) + \omega \, log \, v(x,y)$, where μ and ω are weight parameters, v is a variance function. Among the mouth corner candidates, the one having the largest F_{mouth} is selected to be the correct one.

Sometimes, especially when one opens his/her mouth widely or compresses his lip mightily; the largest $\mathrm{F}_{\mathrm{mouth}}$ may correspond to a wrong candidate. In fact, when a mouth is widely opened, it is difficult to detect by an adaboost-based detector because the current mouth image deviates significantly from the normal mouth appearance, and sometimes even all the detected candidates do not contain the correct mouth corner. Similarly, when one compresses his/her lip mightily, the variance of a facial wrinkle corner may be larger than that of the correct mouth corner. Therefore, we further proposed a method to improve the correctness of mouth corner selection.

3.1.4 Further Improvement of Mouth Corner Selection

If the angle between the line passing two eye pupils and the line passing two mouth corner candidates is larger than a threshold, it indicates the current mouth direction is inconsistent to the current eye direction and this constitutes an abnormal face composition. Therefore, it will be very useful for us to make a certain modification so that a wrongly selected candidate can be updated to a correct one. Our experiments showed when encountering an abnormal face composition, most probably one mouth corner candidate (called 'candidate A') is correctly selected and the other one (called 'candidate B') is incorrectly selected. Therefore, we try to predict the correct position of candidate B by using the correct candidate A. Experiments showed the two eye pupils are easier to detect than the two mouth corners and they have higher detection accuracy. So we can remedy the wrongly detection mouth corners from the detected eye pupils. First, from the two selected mouth corner candidates, we need to decide which one is correct and which one is incorrect. To serve this end, we simply select the candidate with the larger detection score as the correct one and the other one as the incorrect one. The selected correct candidate is called the "base point". From the two detected pupils, a middle separating line can be constructed which has the same distance to the two pupils. Then, from the "base point" and the middle separating line, an "anchor point" located at the other side of the middle separating line can be found. The base point and the anchor point have the same distance to the middle separating line. Then, a segment can be defined by taking the anchor point as its center, having 1/3 length of the distance between two pupils, and being along the line direction parallel to the two pupils. Within the segment, the most appropriately predicted candidate is obtained by the following design. For each candidate C, two sub-blocks can be defined, one is in the left side of C and the other is in the right side of C. For a true mouth corner candidate, one of its sub-blocks contains a large portion of lip pixels which is called "lip-attributed sub-block" (LASB), and one of its sub-blocks contains a large portion of skin pixels which is called "skin-attributed sub-block" (SASB). Basically, the most appropriately predicted candidate C satisfies two conditions. First, the intensity of the corresponding LASB is smaller than that of the corresponding SASB. Second, the intensity variance of the corresponding LASB is larger than that of the corresponding SASB. However, for each pixel candidate, it is not necessary to explicitly decide which sub-block is the LASB and which is the SASB. Instead, this can easily be decided by simply considering the physical composition of the current processing candidate. If the candidate under consideration corresponds to a left mouth corner, the left sub-block is the SASB and the right sub-block is the LASB. On the contrary, if the candidate under consideration corresponds to a right mouth corner, the left sub-block is the LASB and the right sub-block is the SASB. Therefore, for a true mouth corner candidate, it must follow

$$\begin{cases} Var(\text{LASB}) > Var(\text{SASB}) \\ Avg(\text{LASB}) < Avg(\text{SASB}) \end{cases} \tag{6}$$

Here, Var and Avg denote the intensity variance and the average intensity of a sub-block, respectively. If there are more than one candidate meet the above conditions, the one having the largest sum of Var(LASB) and Var(SASB) is selected to be the most appropriately predicted candidate.

In Fig. 5, the square point and the circle point denote the selected candidates of the left mouth corner and the right mouth corner, respectively. The triangle point denotes the found anchor point and the line segment passing the triangle point is the search range of which the most appropriately predicted candidate of the right mouth corner is decided.

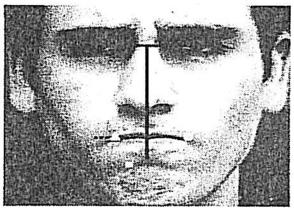


Fig. 5. An illustrative graph for further improving mouth corner selection

3.2 The Second Stage

This second stage is to detect the whole facial landmarks by using the detected locations of the corner-type landmarks from the first stage as the initial positions of the second-stage ASM model. Beside the initialized landmark positions of eyes, eyebrows and mouth, the nose landmark positions are also initialized according to a new composition of eye corners and mouth corners. The average position (S_x, S_y) of the 4 corner-type landmarks (including the left-eye inner corner, the right-eye inner corner, the left mouth corner and the right mouth corner) taken from the average face shape is computed as a reference point. The new nose landmark positions can be estimated by the following three steps:

Step 1. Compute the new average position (C_x, C_y) of the 4 corner-type landmarks obtained from the first stage;

Step 2. Compute the displacement (d_x, d_y) between the reference point and the new reference point, i.e.

$$(d_x, d_y) = (C_x\text{-}S_x\,,\ C_y\text{-}S_y)$$

Step 3. Shift each nose landmark position $\bar{X}_i$ by (d_x, d_y), i.e

$$\bar{X}_i = \bar{X}_i + (dx, dy), i \in \text{landmarks of nose}$$

Fig. 6 shows the initialization method of nose landmarks, and the blue solid circle is the reference point, the blue hollow circle is the new reference point of the currently being processed face, the black triangle denotes the original nose shape, the red triangle denotes the re-initialized nose shape after the displacement of d_x and d_y.

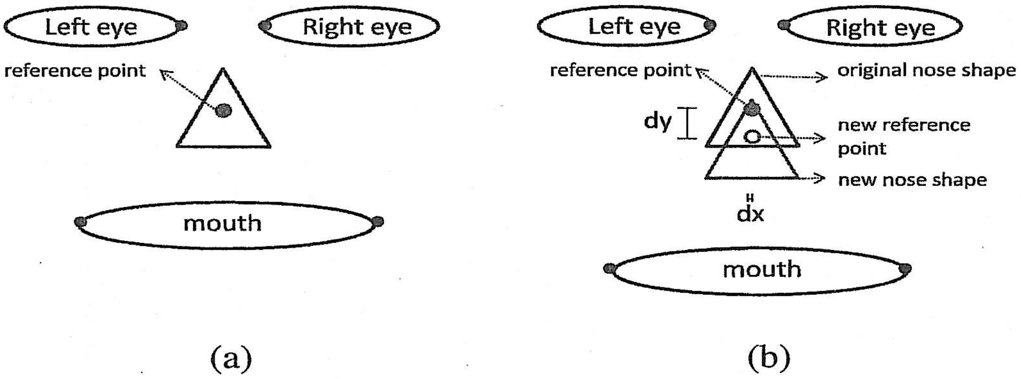


Fig. 6. Initialization of nose landmarks (a) average shape;(b)result of first stage

As for defining the appropriate search range of each edge-type landmark, the standard deviation of its position is adopted. First, the feature landmarks are divided into six groups, as shown in Fig.7. Group 1 denotes the eyebrow-related landmarks; Group 2 denotes the eye-related landmarks; Group 3 denotes both side nose-both-side-related landmarks; Group 4 denotes the bottom nose-bottom-related landmarks; Group 5 denotes the upper-lip-related landmarks; and Group 6 denotes the lower-lip-related landmarks. On purpose, all the landmarks in the same group use a same search range SR_k which is defined as:

$$sd(j) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_{ij} - \bar{x}_j)^2} \tag{7}$$

$$SR_k = \max_{j \in group_k} sd(j) \tag{8}$$

where k is the group index, j is the landmark index, $\bar{x}_j$ is the average position of the jth landmark, and $sd(j)$ is the standard deviation of the jth landmark position.

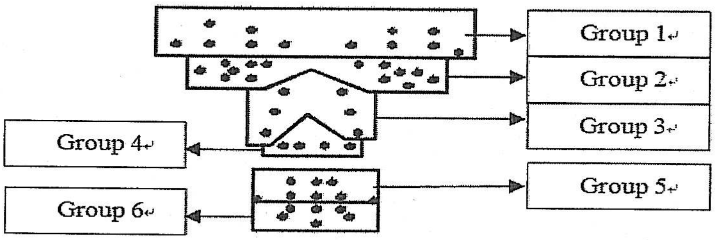
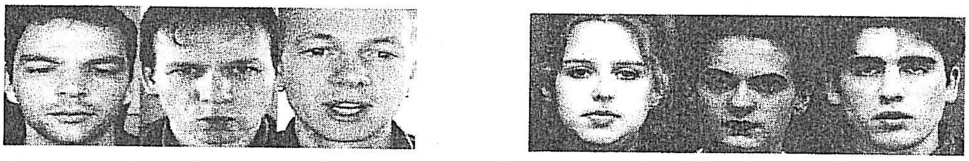


Fig. 7. A illustrative diagram of the six-groups facial landmarks

4 Experimental Results

We use the well known BioID face database and a part of the Cohn Kanade database to train the ASM shape model and the 10 corner-type Adaboost-based detectors. In total, there are 3016 BioID face images (include mirrored images) and 6588 Cohn Kanade face images used in the training stage. Because the Cohn Kanade database in total contains 9005 face images, the remaining 2417 face image is used to test the performance of landmark localization. Fig. 8 shows some samples of both databases. The 50 landmark points are manually labeled for all the images.



(a) (b)

Fig. 8. (a) Examples of the BIOID database, and (b) Examples of the Cohn Kanade database

The hit rate of each corner-types landmark is calculated and is listed in Table 1. In this paper, the hit rate of each landmark is defined as

$$hit\ rate(\%) = \frac{\sum_{i=0}^{N} f(i)}{N} * 100\% \tag{9}$$

$$f(i) = \begin{cases} 1 & ,\text{if } |M_i - D_i| < 0.3 * M^w \text{ and } \frac{M^w}{1.5} < D_i^w < 1.5 * M^w; \\ 0 & ,\text{otherwise} \end{cases} \tag{10}$$

where N is the total number of test images, M_i is the manually marked position of this landmark of the i-th image, D_i is the representative position of the detected landmark block of the i-th image, D_i^w is the width of the detected landmark block of the i-th image, and M^w is the width of the manually marked landmark block.

Table 1. The hit rates of different corner landmarks. The index 1/2 is the left outer/inter eyebrow corner, the index 3/4 is the right outer/inter eyebrow corner, the index 5/6 is the right outer/inter eye corner, the index 7/8 is the left outer/inter eye corner, the index 9/10 is the right/left mouth corner.

Index	1	2	3	4	5	6	7	8	9	10
Hit rate(%)	97.4	94.9	93.9	99.3	96.3	96.6	98.8	98.4	94.7	98.2

For evaluating the accuracy of landmark localization, the error rate E is defined as

$$E_j = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{pt_{ij} - ans\_pt_{ij}}{dist_i} * 100\%\right) \tag{11}$$

where pt_{ij} is the detected position of the j-th landmark of the i-th image, ans_pt_{ij} is the manually marked position of the j-th landmark of the i-th image, and $dist_i$ is the distance between the two pupils of the i-th image.

The overall performance of the proposed two-stage ASM method and the traditional ASM are compared by showing their individual errors of each landmark in Fig. 9.

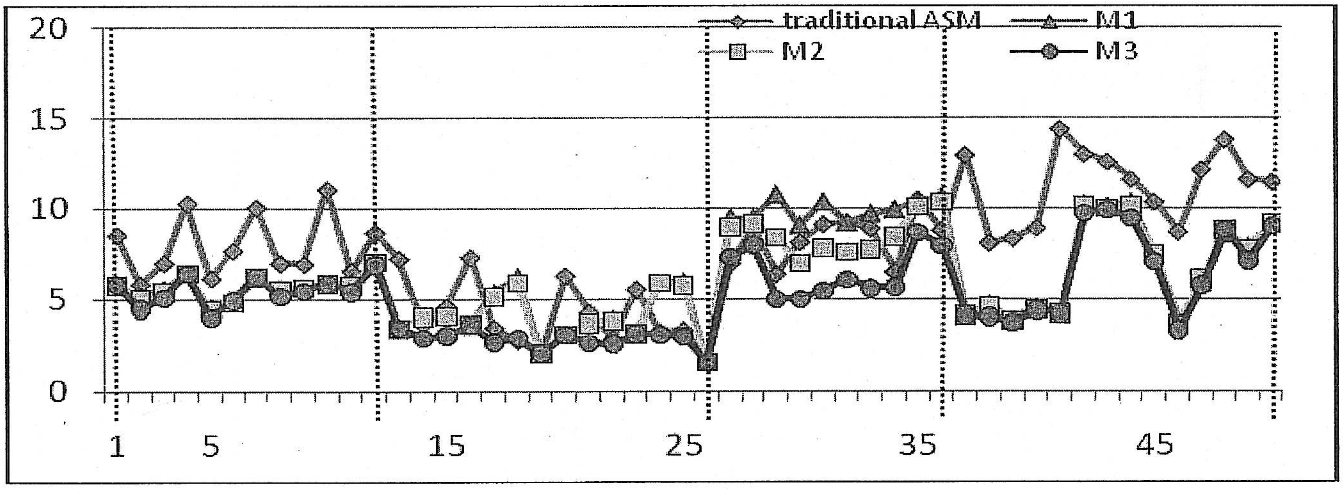


Fig. 9. The error rate of each corner landmark. The horizontal axis is the corner index in which No.1~No.12 correspond to the eyebrow-related landmarks, No.13~No.27 correspond to the eye-related landmarks, No.28~No.36 correspond to nose-related landmarks, and No.37~No.50 correspond to mouth-related landmarks. The vertical axis indicates the error rate.

There are several improvements proposed in this paper. In order to verify their effectiveness, several experiments are conducted by using different combination of the proposed methods. M1 denotes using the first stage to locate the corner-type landmarks with the Adaboost-based detectors and the traditional ASM to locate the edge-type landmarks without initializing the nose shape, M2 denotes using the first stage 1 to locate the corner-type landmarks and the improved ASM to locate the edge-type landmarks with initializing the nose shape, and M3 denotes using the first stage to locate the corner-type landmarks and the improved ASM to locate the edge-type landmarks by using both nose shape re-initialization and different landmark-related search ranges.

Form Fig. 9 we can see the error of M1 in the nose part is larger than traditional method. Because when we initialized the eye, eyebrow and mouth without initialized the nose, sometimes it would undermine the overall facial structure. Therefore, it will cause large errors. But in M2, because we using the eye corner and mouth to initialize the nose position, the error rate in nose can be reduced.

Obviously, M3 performs much better than the traditional ASM. This reveals that both the Adaboost-based corner-type landmark detectors and the variable rectangular search ranges are very useful in detecting the corner-type landmarks of eyebrow, eye and mouth, such as the 1th, 4th, 7th, 10th, 13th, 16th, 20th, 23th, 37th and 41th landmarks in Fig. 9. When a human face has made an exaggerated facial expression, due to the two-stage ASM design, most landmarks can still be detected correctly. By using different search ranges for different landmarks can also improve the landmark localization accuracy. Although none of nose-related landmarks belongs to the corner-type landmark, they can still using the eye corner and mouth corner to improvement. Fig. 10 shows the detected positions of 50 landmarks by using the traditional and the proposed two-stage ASM methods, individually, the first row is the processed results of the traditional ASM, and the second row is the processed results of the proposed method M3. Obviously, the proposed method M3 gets much better results than the traditional ASM.

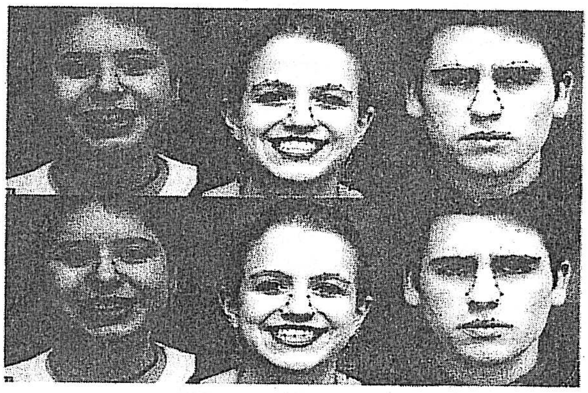


Fig. 10. Some results on the Cohn Kanade database. The top row shows the detected landmarks by the traditional ASM method and the bottom row shows the detected results by the proposed ASM method.

5 Conclusion

In this paper, we have proposed a two-stage ASM method to improve the facial landmark detection. The first stage uses an Adaboosting algorithm to locate 10 corner-type landmarks, which are attributed into three classes (i.e., eyebow, eye and mouth) and each class has its own candidate selection method from the detected candidates. The second stage is to detect the whole facial landmarks by using the locations of the detected corner-type landmarks in the first stage as the initial positions of ASM, and different facial landmarks correspond to different search ranges based on their variation extents. From the experimental results, it demonstrates clearly that the proposed method outperforms the traditional ASM algorithm, especially in corner-type landmarks. In the future work, we will try to design a 2D feature model instead of the tradition 1D feature model for the edge-type landmarks. Expectedly, it can further improve the accuracy of localizing facial landmarks.

Acknowledgments. This research is supported by Taiwan NSC under contract NSC 98-2221-E-216-029.

References

1. Zhang, B., Ruan, Q.: Facial feature extraction using improved deformable templates. In: The 8th International Conference on Signal Process., vol. 4 (2006)
2. Coots, T.F., Taylor, C., Cooper, D., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
3. Zhou, D., Petrovska-Delacr'etaz, D., Dorizzi, B.: Automatic Landmark Location with a Combined Active Shape Model. In: IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (2009)
4. Pu, B., Liang, S., Xie, Y., Yi, Z., Heng, P.-A.: Video Facial Feature Tracking with Enhanced ASM and Predicted Meanshift. In: Second International Conference on Computer Modeling and Simulation (2010)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Proc. European Conference on Computer Vision (1998)
6. Zuo, F., de With, P.H.N.: Fast facial feature extraction using a deformable shape model with Haar-wavelet based local texture attributes. In: Proceedings of IEEE Conference on ICIP (2004)
7. Du, C., Wu, Q., Yang, J., Wu, Z.: SVM based ASM for facial landmarks location. In: 8th IEEE International Conference on Computer and Information Technology, CIT 2008 (2008)
8. Wan, K.-W., Lam, K.-M., Ng, K.-C.: An accurate active shape model for facial feature extraction. Pattern Recognition Letters 26(15) (November 2005)
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2001)
10. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) (2005)

國科會補助計畫衍生研發成果推廣資料表

日期:2010/11/02

國科會補助計畫	計畫名稱：三維物件掃瞄及裸眼立體投影顯示系統開發
	計畫主持人：鄭芳炫
	計畫編號：98-2221-E-216-031-　　　　　學門領域：影像處理

無研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：鄭芳炫　　計畫編號：98-2221-E-216-031-

計畫名稱：三維物件掃瞄及裸眼立體投影顯示系統開發

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	1	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	4	4	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	1	1	100%		
國外	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	2	2	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	此研究成果將與工研院電光所合作開發以紅外線結構光為基礎的 3D 深度攝影機.

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 ■洽談中 □無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本研究發展以結構光的 3D 立體取像方式, 為一個較簡易且建構成本較低的方式，不僅取像過程速度較快，且可重建原物體之原始表面材質的特性, 由於 3D 立體取像的需求日益提高，其應用也日益廣泛如 3D 互動辨識、醫療、工商業模型建立、娛樂、珍貴文物的數位化、3D 顯示的資料來源建立、動作的擷取等。