基於改良式形狀上下文與最近鄰居法對人體姿態行為辨識分析

戴顯權 [a]、王峻國*[a]、沈冠宇 [a]、王國明 [a]、趙乾言 [a]
國立成功大學電腦與通信工程研究所 [a]

摘要 — 本論文中，我們利用人體縱軸及橫軸投射直方圖透過傅立葉轉換後取得特徵，並且提出一個改良式形狀上下文(Shape Context)比對方法取得另一特徵，經由最近鄰居演算法(k-th nearest neighbor, K-NN)來分類單張姿態，最後透過隱藏式馬可夫模型來判斷連續性的影像。經由上述改良的方式可降低複雜度以達成即時判斷姿態的效果，並得到一個完善的視覺監控系統架構。[1]

關鍵字： 傅立葉轉換、姿態分類、Shape Context、K-NN、隱藏式馬可夫模型

一、 前言

由於近年來人口密度不斷攀升，但是出生率卻逐年下降，家庭多半呈現少子化的狀態，因而促使社會逐漸面臨了人口老化的問題。就目前而言，由於大多數都是雙薪家庭的緣故，多數家庭只留下年邁的老人或年幼的孩童於家中，且沒有多餘的時間注意老人及小孩在家中的安全，然而對於較無應變能力的他們，一旦有狀況發生則後果不堪設想。因此，為了協助大部分家庭於這方面的困擾，我們將設計研究發展出一套視覺監控系統架構，以提升居家照顧的生活安全品質，並且大大節省人力資源的浪費以及不必要的支出。

就居家看護而言，老人和小孩最常發生的危險性動作不外乎就是跌倒，所以在我們整個系統架構中，判斷人體是否有危險性的動作是必須精確的，因此在系統分析上會特別去注意跌倒姿態的評估。

本計畫流程如下：第二章，我們討論關於姿態辨識上的主要有哪些方法，並研究這些方法中有那些缺點是需要被改善的，在第三章中，描述我們系統架構及影像前處理，第四章中描述如何擷取特徵，並於第五張中描述如何透過特徵判別姿勢以及利用隱藏式馬可夫模型來預測姿勢。第六章，透過實驗數據來分析驗證整個系統架構的準確度，最後於第七章做一個總結及未來目標。

二、 相關文獻

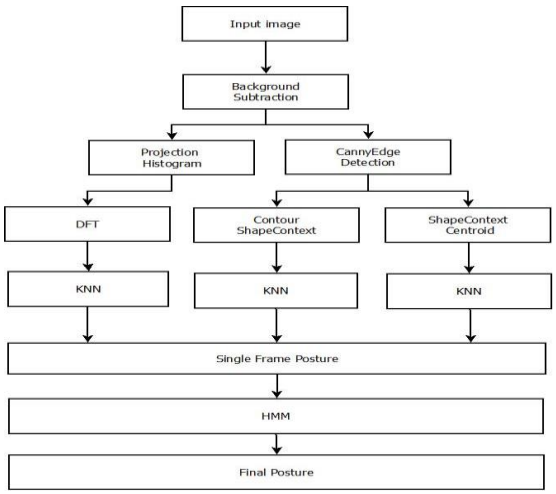
近年來有許多方法被提出用來解決於姿態辨識上的問題，如在1998年由Fujiyoshi 和 Lipton [1]利用星形式的骨幹架構去計算之間的角度來分析人類行為的變化，進而去判斷出行走與跑步的正確性。2003年由Spagnolo et al. [2]首先把影像轉換成直方圖，接著透過競爭學習系統演算法去學習分析，進而得到分析的物體。2005年由Cucchiara et al. [3]利用幾何學及顏色直方圖取出物體輪廓的特徵值，再經由隱藏式馬克夫(HMM)模組運算每一個物體的行為機率以達到辨識結果。2006年由Davis 和 Tyagi [4]提出利用隱藏式馬可夫之有限狀態機描述其每一個動作之機率值為多少，進而辨識它視為何種行為。

[1]本研究由國科會贊助，計畫編號NSC 101-2221-E-006-217。

2006年由Robertson 和 Reid [5]提出透過主成份分析得到主要特徵，經過隱藏式馬可夫(HMM)來分析人體行為姿態。2008年由Thome et al. [6]提出階層式馬可夫於多視覺影像中分析跌倒行為偵測。2010年由Gu et al. [7]以位置、方向、目前身體高度做為特徵經由馬可夫模組進行分析、分類。2010年由Zhang et al. [8]從3-D影像的樣本中選取特徵透過類神經網路進行分析來得到一個有效的人體行為分類辨識。2011年由Behrouz and Deepu [9]提出以時間與空間上的關聯距離去探討姿態分析。然而以上眾多研究者所提出的方法仍然有些問題存在，例如遮蔽物擋住主要物體、只處理單張影像的分析或是沒考慮即時問題等等，而這些判斷後的結果通常並沒有去做後續的分析及處理或應用。為了解決上述的問題，我們提出一個改良式Shape Context[10]的比對方式及利用K-NN[11]演算法的分析，並透過隱藏式馬可夫模型來校正單張影像容易因雜訊產生的失誤，藉此提高整個系統的準確度，讓整個視覺監控系統架構可以更加的完善。

三、 人體輪廓分割

本章節中，我們將開始描述整體架構流程(圖一) 及分割人體輪廓的步驟：



圖一： 系統架構

首先，我們採用單台攝影機固定設置於室內拍攝連續影像並且以二維影像做分析，首先將得到的連續影像圖片之色彩由 RGB 轉換成 YUV。其中 YUV 色彩空間主要成份為 Y(亮度)和 U、V(色度)，Y 是由彩色轉換成灰階影像的灰階值，轉換過程主要依據人類視覺對 RGB 三原色(紅、綠、藍)的不同敏感度而來，當轉換的係數越大表示顏色敏感度越高，所以三種顏色之敏感

Corresponding author:jyunguo.wang@gmail.com
DOI:10.6159/IJSE.2014.(4-1).34

度依序為綠色(0.587)、紅色(0.299)、藍色(0.114)。另外對人類視覺而言，低頻資料比高頻資料更具敏感，並且亮度變化相較顏色變化敏感，而 JPEG 影像的應用上一般採用灰階或是全彩影像，因為灰階影像只包含相對重要的亮度資訊，因此我們將採用灰階影像。透過 RGB 和 YUV 的關係式進行影像的像素值評估，如公式(1)所示：

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ U &= -0.148R - 0.289G + 0.437B \\ V &= 0.615R - 0.515G - 0.100B \end{aligned} \quad (1)$$

取得 Y 影像後把背景與前景分割出來，在這裡我們採用背景相減法計算影像之間的差異進而得到影像中的物體，但是我們知道有些物體並不是我們所想要的物件，如：雜訊、陰影，因此我們會在此步驟加入去雜訊及去陰影的方法，而去雜訊的方法有很多例如：低通濾波、平滑法、中值濾波等等。

對於不同濾波器而言，每一個頻率信號的減弱程度不同，低通濾波容許低頻信號通過，但減弱(或減少)頻率高於截止頻率的信號通過，因此較適用於處理音頻訊號，將不加以分析。而在平滑法雖然可以消除雜訊，但是會因為遮罩大小而造成影像在分析處理時影響影像模糊化，導致在後續的處理上可能有所誤差。中值濾波雖然它可以解決平滑法的問題但還是有它難處所在，例如：假設把運算值經過計算排序後，取得中間值做取代，但這個動作很容易將不是雜訊的高頻資訊濾掉，造成重要資訊的遺失。因此，我們採用較簡單的侵蝕與擴張方法來去除雜訊得到清晰的輪廓。

四、 影像特徵擷取

4.1 DFT

經由章節三的影像前處理過後，我們得到較為清晰前景的影像輪廓，接著針對此影像輪廓我們分別採取三個方法來取得我們所要的特徵，第一，我們將所取得的影像輪廓投射至垂直與水平直方圖中，我們計算影像在水平投射中每列的像素累計值，假如列的累計值小於門檻則作為是身體高度的頂端(P_{up})，接著計算水平投射影像，當某一列的累計值小於門檻則表示此列為身體高度的底端(P_{down})。另外，我們也計算垂直投射影像累計每行的像素值，若行的像素累計值小於一個門檻，則表示為身體寬度的左邊 (P_{left})及右邊(P_{right})，透過這樣的方式找到該物體的長度和寬度，進而從長度和寬度當中我們得到物體的長寬比例做為我們其中的一個特徵，如公式(2)所示：

$$P_{hjgh} = P_{up} - P_{down} \quad , \quad P_{width} = P_{right} - P_{left}$$
$$H - W_{ratio} = \frac{P_{high}}{P_{width}} \quad (2)$$

然而直接從直方圖中取得輪廓大小或是物體相關的位置作為特徵是較不具代表性，因為它會受限於攝影機的比例而導致大小不一樣，並且人體若有移動使得位置點改變將造成有誤判的情況發生。因此，為了解決上述兩種情況我們可以透過轉換空間的方法來計算，常見的轉換空間方法例如：傅立葉轉換、離散餘弦轉換等等。

在論文中我們選擇傅立葉轉換(DFT)的方法選取作為另外的特徵值，它是可以把時域空間的問題轉換成正弦分量的頻域空間進行分析，在分析上 DFT 可以把影像的空間轉換成頻率空間(時間-振幅)並將時域(time domain)函數以數學方法轉換為頻域(frequency domain)函數，如此可以有效解決上述問題。經由 DFT 轉換後的形式我們可以得知頻率的高低，用以表示資料是否屬於較為重要資訊而得到有效的特徵值，如公式(2)所示：

$$DFT_H[h] = \frac{1}{N}\sum_{y=1}^{N} H(y)\exp(-j2\pi yh/N), h = 0,.....,N-1$$
$$DFT_V[v] = \frac{1}{M}\sum_{y=1}^{M} V(x)\exp(-j2\pi xv/M), v = 0,.....,M-1 \quad (3)$$

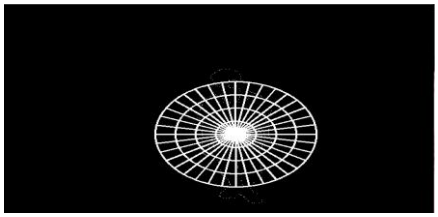
其中 M、N 為影像大小，H(y)、V(x)為水平垂直的投射直方圖，u、v 為直方圖的值，之後經運算後我們會取出較為重要的前 10 筆係數資料做正規化表示為我們的特徵，我們利用水平跟垂直的投射直方圖產生各 10 筆的係數資料加上長寬比作為我們的第一個方法的特徵。

4.2 Shape Context

在另一方面，我們把前處理後的影像做 Canny Edge Detection [12]，Canny Edge Detection 分為三個步驟來處理：降低雜訊、尋找影像當中的亮度梯度及使用 Hysteresis Thresholding 去評估是否為邊緣。經過 Canny Edge Detection 演算法處理後，我們便可取得一個完整的邊緣如圖二所示：

圖二： Canny Edge Detection

接著我們把 Canny Edge Detection 處理後產生的輪廓做 Shape Context。Shape Context 是一種物體型態描述的方式，一開始它會選擇輪廓點當中的最高點，然後把這個輪廓點當作圓心展開一個同心圓，並且根據選定的角度把同心圓切成多個片段(圖三)，然後透過公式(4)來記錄每一個片段輪廓點的分布，直到所有的片段紀錄完成，才尋找下一個輪廓點作相同的動作，直到所有的輪廓點都執行一次為止，其中公式 h_r 代表每一個片段裡面的點數，r 為展開同心圓的圓心，q 為輪廓點。



圖三： Shape Context

$$h_r(k) = \#\{q \mid q \neq r, (q-r) \in bin^k\} \quad (4)$$

然而在原先的 Shape Context 比對方式上，都是採用片段與片段之間的計算來做為匹配程度的好壞，所以在執行計算上需耗費大量的時間，導致不適合應用在即時系統上，因此我們提出改良式 Shape Context 的方法來取得特徵，我們把改良式的 Shape Context 分為兩種方式取得特徵。第一種：對邊緣偵測後所得到的影像輪廓只取 10 個點去做計算，利用這 10 個點的 Shape Context 來描述物體形狀的分布情況並做為第二個方法特徵，第二種：對影像物體輪廓的質心做 Shape Context，並做為第三個方法特徵。經由上述得到的特徵來做為第五章中分類模型辨識的輸入。

五、 姿態辨識

5.1 K-NN 演算法

K-NN 的演算法被廣泛的應用於不同領域上，例如：工業工程[13]、影像處理[14]，及訊號處理[15]方面等，因為它既簡單、複雜度低在分類上都有不錯的效果，因此我們決定採用 K-NN 演算法來進行辨識。

在我們的系統中，一開始先把 4 種姿態各 50 筆訓練樣本，配置在一個多維空間裡，其中維度取決於我們前段章節中所提到三個不同方法的特徵數量，並將這三種方法分別進行 K-NN 運算及姿態辨別，而不把這三種方法混合使用，其主要原因是在於方法物理意義不同，例如：Shape Context 中每一個片段裡面的點數並不能代表為一種特徵，僅能代表輪廓的分配情況。

方法一，我們分別取得橫軸及縱軸的投射直方圖做 DFT 後，各取前 10 筆，再加上長寬比共 21 筆的特徵，透過 K-NN 演算法進行相似度比對，最後從一開始配置的 200 筆姿態當中挑選出前 20 筆相似度最高的姿態，並記錄下來。

方法二，對經過 Canny Edge Detection 處理後的影像輪廓點中，依照由上到下、左到右的順序擷取出 300 個點，並以每隔 30 個輪廓點做一次 Shape Context，計算片段內的輪廓點數量，直到完成 10 次的 Shape Context 採集。在系統當中我們選取的角度為 10 度、bins(環的數量)為 4 個，因此一個 Shape Context 裡面會有(360/10)*4 = 144 個片段，10 個輪廓點乘上 144 格片段等於 1440 個區塊，接著透過 K-NN 來計算每個區塊之間的差異性來判斷姿態之間相似度，最後從一開始配置的 200 筆姿態當中挑選出前 20 筆相似度最高的姿態，並記錄下來。

方法三，我們對質心做 Shape Context，然而只對質心做 Shape Context，所以只有 144 區塊，經由 K-NN 比對後，從配置的 200 筆姿態中挑選出前 20 筆相似度最高的姿態，並紀錄下來。

經過上述三個方法後，我們取得3*20=60筆的姿態資訊，並從這些資訊當中去分析哪一種姿態所佔的比例較高，以做為單張辨識畫面中的姿態輸出。

5.2 隱藏式馬可夫模型

前面提到的方法是針對單張去判斷姿態，然而單張畫面的姿態判斷極容易受到光源雜訊及物體行為瞬間改變的影響，導致判斷失誤，倘若能透過一連串的影像張數來預測目前影像畫面的影像姿態，便可使得目前影像的姿態判斷更加準確。因此，為了使系統更加完善，我們引用了隱藏式馬可夫模型[16]（Hidden Markov Model，HMM）。

一個隱藏式馬可夫模型是由三種參數所組成，一個是初始狀態機率值，一個是狀態轉移機率值以及狀態觀測機率值，在隱藏式馬可夫模型中，我們所知的只有實際出現的觀測序列值，以及各觀測序列值對應各狀態的機率值，至於狀態的轉移情況則屬於未知，故稱為隱藏式馬可夫模型。

因此在系統中我們透過 HMM 的概念，根據目前畫面之前的 20 張畫面姿態做為已知序列，來預測目前畫面姿態，以提高整體系統的準確度。

六、 研究結果

在實驗方面，我們採用的平台為Window7、CPU 2.9Hz，撰寫語言為C++，使用函式庫opencv2.0，處理的影像畫面都是經由IC-7110W紅外線夜視型雲端無線網路攝影機所拍攝，解析度為640*480，程式計算上每秒鐘處理15-20張Frame，姿態辨識方面採用4個姿態來做為實驗：站立、蹲下、彎腰及跌倒。

結果的呈現，我們分別以三個部分來探討，第一部分先透過數據分析來選擇Shape Context的bin和切片區段的角度，決定Shape Context的bin和切片區段的角度後，當作為第二部分及第三部分Shape Context的參數，並透過一般情況和複雜情況的環境底下進行分析，以驗證我們整體系統架構的實用性。

6.1 Shape Context 的 bin 和角度

實驗中，我們以4個不同的bin及角度對系統進行測試，以觀察對姿態判別有何影響，在實驗環境中分為兩種情況(一般及特殊)進行探討。一般情況(圖四)定義為光源充足、無複雜衣物的情況；特殊情況(圖五)則定義為穿著複雜衣物、光線昏暗及全身未被完整拍攝。

圖四： 一般情況

圖五： 特殊情況

我們在簡單和特殊的情況下分別選取每一個姿態各250張影像進行辨識，表格I為在一般情況下，對4個bin及角度搭配的準確度，表格**II**則為在特殊情況下的準確度。根據這兩張表格分析結果得知：角度越大，效果愈差，但在bin的方面，一般情況下bin為4時準確度較高，特殊情況下bin為8時準確度較高。然而我們發現當bin的區塊切割愈細，在姿態辨識上所需計算時間也會相對增加，因此基於系統執行速度和平均準確度的考量之下，我們最後選擇以bin為4個，角度為10度來做為後續實驗的參數。

表格I 一般情況下的準確度

角度 / bin	10	30	60	90
2	86.5%	68.7%	64.7%	59.1%
4	98.6%	89.1%	91.2%	82.4%
6	87.7%	87.4%	74.7%	71.6%
8	87.4%	72.2%	70.5%	73.3%

表格II 特殊情況下的準確度

角度 / bin	10	30	60	90
2	71.8%	56.3%	58.3%	41.6%
4	86.3%	85.1%	87.8%	73.4%
6	86.9%	89.9%	81.8%	71.1%
8	92.6%	94.7%	84.2%	78.7%

6.2 一般情況下的分類情形

決定Shape Context的參數後，我們可以透過表格III來觀察系統在一般環境下分類情形，並採用4種姿態各250張的畫面進行分類。表格中，橫軸為所要分類的項目姿態，縱軸為所要辨識的項目姿態。

表格III 一般情況下姿態分類表

分類 / 辨識	站立	蹲下	彎腰	跌倒	準確度
站立	250	0	0	0	100%
蹲下	0	248	2	0	99.2%
彎腰	0	5	245	0	98%
跌倒	0	7	0	243	97.2%

根據以上結果，可以發現 4 個姿態分類準確度都在 97% 以上，而分類錯誤的姿態大多是因為在姿態上有大幅度的改變，導致輪廓判斷失誤。

6.3 複雜情況下的分類情形

表格IV採用4種姿態各250張的畫面進行分類，橫軸為所要分類的項目姿態，縱軸為所要辨識的項目姿態：

表格IV 複雜情況下姿態分類表

分類 / 辨識	站立	蹲下	彎腰	跌倒	準確度
站立	204	3	43	0	81.6%
蹲下	39	197	14	0	78.8%
彎腰	0	33	217	0	86.8%
跌倒	0	5	0	245	98%

透過表格IV我們可以發現，複雜情況下和一般情況的準確度上會有些許的落差，主要原因在於人體輪廓會因複雜的衣物、光線的強弱影響下，導致背景和人體對比不夠強烈、輪廓線條不甚明顯，造成在姿態判別容易混淆，以致準確度下降，但針對於跌倒這個姿態來看，還是可以有不錯的辨識率。

七、 結論

實驗結果證明我們的系統架構可以達到不錯的效果，並在判斷跌倒姿態的準確度方面，不論在一般或特殊的環境下都可以達到97%的辨識率，因此對於居家照顧當中的防範跌倒，可以有不錯的效果。

在未來研究方面，由於系統在背景減法上並沒有特別使用其他方式做處理，因此在前處理的影像當中還是會有一些雜訊干擾，所以倘若能利用一個自適應高斯濾波加入此系統裡，應能有效提升此系統架構的準確度。

參考文獻

[1] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," Proc. *IEEE Workshop on Applications of Computer Vision*, pp. 15 – 21, Oct. 1998.

[2] P. Spagnolo, M. Leo, G. Attolico, and A. Distante, "Posture recognition in visual surveillance of archeological sites," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1542-1547, Oct. 2003.

[3] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment**"**, *IEEE International Conference on Image Processing*, vol. 1, pp. 11-14, Sept. 2005.

[4] Davis, James W.; Tyagi, Ambrish, "Minimal-latency human action recognition using reliable-inference," *Image and Vision Computing*, Vol. 24, Issue： 5, pp. 455-472, May 2006.

[5] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, Vol. 104, Issue： 2-3, pp. 232-248, Nov. 2006.

[6] N. Thome, S. Miguet, and S. Ambellouis, "A real-time multi-view fall detection system： A LHMM-based approach," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 18, No. 11, pp. 1522-1532, Nov 2008.

[7] J. Gu, X. Ding, S. S. Wang, and Y. Wu, "Action and gait recognition from recovered 3-D human joints," *IEEE Trans. on Systems, Man, and Cybernetics—PART B： Cybernetics*, Vol. 40, No. 4, pp. 1021-1033, August 2010.

[8] B. Zhang, I. Horváth, J.F.M. Molenbroek, and C. Snijders, "Using artificial neural networks for human body posture prediction," *International Journal of Industrial Ergonomics*, pp. 414-424, February 2010.

[9] B. Saghafi and D. Rajan, "Human action recognition using Pose-based discriminant embedding," *Image Communication*, May 2011.

[10] S. Belongie and J. Malik, "Matching with Shape Contexts", *Content-based Access of Image and Video Libraries, 2000. Proceedings*, Apr 2000.

[11] T, Ivan , "A Generalization of the K-NN Rule", *Systems, Man and Cybernetics*, Feb. 1976.

[12] C.John, "A Computational Approach to Edge Detection", *Pattern Analysis and Machine Intelligence*, Nov. 1986.

[13] Penedo. F, "Hybrid Incremental Modeling Based on Least Squares and Fuzzy K -NN for Monitoring Tool Wear in Turning Processes", *Industrial Informatics*, Nov. 2012.

[14] Kumar, A , Zhang, D , "Personal recognition using hand shape and texture", *Image Processing*, Aug 2006.

[15] Eronen, A.J., "Music Tempo Estimation With K -NN Regression", *Audio, Speech, and Language Processing*, Jan 2010.

[16] N. Thome, S. Miguet, "A Real-Time, Multiview Fall Detection System：A LHMM-Based Approach", *Circuits and Systems for Video Technology*, NOV 2008.