

# 行政院國家科學委員會專題研究計畫 成果報告

## 物件式前景物影像抽取技術 研究成果報告(精簡版)

計畫類別：個別型  
計畫編號：NSC 95-2218-E-216-004-  
執行期間：95年11月01日至96年07月31日  
執行單位：中華大學資訊工程研究所

計畫主持人：黃雅軒

計畫參與人員：碩士班研究生-兼任助理：顏華慶、劉偉成、張偉禕

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：本計畫可公開查詢

中華民國 96 年 10 月 22 日

# 物件式前景物影像抽取技術結案報告

本計畫的研究成果已發表在 2007 International Conference on Artificial Intelligence and Pattern Recognition (2007/7/9~2007/7/12)，論文名稱為” Local Structure Based Foreground Object Extraction”，請參考附錄。

## 一. 前言：

在過去數十年裡，大部分視訊監控系統都使用於閉路電視系統，其主要功能只是消極的錄影存證，並不能主動提供偵測資訊，例如機場、車站、銀行、飯店的監控系統，主要是利用人為方式來辨識是否有恐怖分子和犯罪嫌疑人的存在，這需要人員不停地在監控系統的屏幕上搜尋，然而隨著數位化影音技術的進步，及大量資料儲存之價格降低，加上光學攝影器材成本下降，使視訊(Video)訊號數位化處理更為普及，更重要的是人工智慧技術日趨成熟，使得智慧型視訊監控系統為大眾所矚目，尤其是在美國發生911事件後，世界各國無不把視訊監控技術列為重要的研究課題，加上台灣發生SARS後，對於病患的隔離與醫護人員的進出，更須利用到此監控系統來作有效的管制。

過去十年中，國內外已有許多研究機構作這方面的研究。例如，1997年美國國防高級研究院DARPA訂定了視訊監控(VSAM)為重大的研究計畫項目，主要研究用於戰場及普通民用環境進行監控的自動視訊理解技術；著名的研發成果有(A) MIT之Intelligent Room[1]，其主要功能為可追蹤多人位置，並可辨識坐、立、行之姿勢以及由手指（或雷射）在螢幕指出之位置；(B) Microsoft之Easy Living[2]，其主要功能為可追蹤多人位置及分辨身份，如使用者在螢幕A欣賞影片，若使用者移至另一處時，該影片會在螢幕A停止播放，轉而在靠近使用者之螢幕B處繼續放映；(C) Tokyo University之Intelligent Space，其主要功能為可追蹤人體3D位置，已可找出人體中之頭、手、足、眼等之位置；(D) 馬里蘭大學發展一套追蹤系統稱之為W<sup>4</sup>系統[3]，不但能夠定位和分割出人的身體部分，而且利用外觀模型的建立可實現多人的跟踪，並可以檢測人是否攜帶物體等簡單行為；(F) 北卡大學也發展一套人類行為辨識 (Human Activity Recognition) 系統[4]，可以辨識出街道上所發生的竊盜行為並發出警告；(G) 卡內基美濃大學[5]已經發展一套視訊監視與自動監控系統(video surveillance and monitoring, VSAM)系統，用來做一般的物體偵測與追蹤，在此系統中，包括了SPU、OCU與GUI 等模組；(H) 哥倫比亞大學研究學者[6]整合鏡面反射系統至傳統的攝影系統中，設計成一個全方位的攝影機系統，如同於現行的商店中，於各個角落加裝一面反射鏡，如此一來，可以清楚地看到各個角度的畫面，只要從鏡面中偵測移動目標即可，此外包括一個傳統鏡頭，可以得到一般的正常透視投影的影像，此兩支鏡頭將裝在一起，形成一個全方位的攝影裝置；(I) 布朗大學[7]則針對人類行為分析進行了一系列的研究，並且有了許多基礎的技術，例如人類肢體辨識，人類動作辨識、臉部運動辨識等，他們將分析人類行為的步驟分成以下三個部分，Human Motion Estimation, Recognizing Human Motion, Recognizing Human Behavior。至於國內研究單位則是處於起步階段，如台灣大學、清華大學、交通大學、成功大學、中正大學、中央大學、元智大學、中研院以及工研院等。近年來，亞洲大學 蔡文祥校長率領國內二十多位教授一起執行學界科專「以視覺為基礎之智慧型

環境的建構」計畫，經過二年多的努力，已累積許多技術和技轉的成果。

所謂智慧型視覺監控系統就是利用電腦視覺的方法，在不需要人力干預的情況下，對攝影機拍攝的影像進行自動分析，進而對目標物進行追蹤、識別和行為分析。在分析過程中，不論最後的目標是進行追蹤或辨識，首先必須將前景物件影像與背景影像先作出區別，之後才能針對前景物件影像作進一步的處理(如物件辨別和行為分析等)。因此，前景/背景物體影像分離即成為一項非常關鍵的處理技術。其正確性將密切且絕對的影響後續的分析結果。若被抽取出的前景物件影像嚴重破碎成好幾部分，則無論接下來是進行物件追蹤或物件辨識，都將因物件資訊不完整而極易得到錯誤的處理結果。有鑑於此技術的重要性，中華民國影像處理與圖形識別學會(IPPR)於2006年首度舉辦技術競賽，題目為「運動物體偵測技術」，也就是本研究探討的「前景/背景分離技術」。希望藉由此競賽，能增進大家對此議題的認識瞭解與相互交流，更而鼓勵更多研究力量的投入，以激發出新的解決方法，使得在監控產業蓬勃發展的今天，能早日開發出實用的智慧監控系統。

## 二. 研究目的：

如何在一連串的視訊資料流中，判斷哪些影像點是屬於前景物件而哪些影像點是屬於背景畫面是一項非常重要且關鍵的技術。其處理結果的好壞將密切的影響視覺監控系統之視訊分析的正確和實用性。有鑑於現今技術所面臨的技術瓶頸，本計畫提出一種新穎性的前景物件影像抽取方法，以更豐富的影像資訊來得到更正確與完整的前景物件影像抽取結果。本計畫技術有望能帶動視覺監控產業相關應用系統的發展。

本發明將傳統的前景/背景微觀處理與巨觀的物件偵測處理進行巧妙的結合，大大提升了所抽取的前景影像之正確性和完整性。本計畫技術方法容易實施，具有高度實用性。

## 三. 文獻探討：

一般而言，分離前景物體與背景環境的方法主要可區分成三種類型：背景相減法(Background Subtraction)[1,5,19,28,29]、時間序列畫面差異法(Temporal Differencing)[30]和光流場法(Optical Flow)[31,32,33,34]。背景相減法分辨出目前影像與所設定的背景模型之差異，為目前最普遍被應用於分離前景的方法。背景模型中每個影像點經常採用單一高斯模型或者是混合式高斯模型(Gaussian Mixture Model)來表達。此方法簡單容易應用，但是卻容易由於一些背景的變化而無法分離出正確的前景，像是光源的變化或是出現後靜止的物體。時間序列畫面差異法則是利用前後不同時段所拍攝的影像畫面來進行直接相減，對任何影像點而言，只要此差異量的絕對值大於所設定的臨截值，則它將被判斷為一前景物件影像點；否則，它將被判斷為一背景影像點。至於光流場法會對移動的物體之每個影像點計算出在二個影像畫面中的位移量，當有前景物件進入攝影中的畫面時則會引起光流場的變化。因此，利用光流場的大小即可判斷移動物體的影像位置。一般而言，背景相減法和時間序列畫面差異法的執行速度相當快速，單用軟體方式即可以達到即時處理的目的；光流場法則因計算複雜(每個影像點需做區域性的比對搜尋)，以致若無硬體的配合，則無法滿足即時處理的要求。背景相減法常會因為前景物件影像和背景影像之顏色相近，造成所抽取的前景物件影像相當破碎。另外，陰影亦常會被判斷成為

前景物件影像。時間序列畫面差異法雖對陰影問題有較大的容忍度，但若前景物件移動緩慢或靜止時，則大部分前景物件影像均會被判斷為背景影像。通常，具有邊界或線條的前景物件影像之局部性區域較容易被抽取出來，而那些沒有太大顏色變化的前景物件區塊影像(如近距離攝影時的衣服、褲子或臉孔)，亦常會被誤判為背景影像。光流場法雖然在許多情況中可以得到較完整的前景物件影像，但因其執行速度無法滿足即時處理的要求，一般不會被採用在實際系統中。

目前這三類方法(背景相減、時間序列畫面差異和光流場)在處理時，都是針對整張影像以同一條件來分離所有的前景/背景影像點。但是單獨影像點之資訊特徵(如色彩資訊)的區分能力有限，以致時常無法做出正確的前景/背景分離判斷。例如，當一個人穿著黃色衣服站在淺黃色的牆前面，則可能因衣服和牆面的顏色差異沒有大過於所設定的差異臨界值，以致於無法正確的將衣服部分判定為前景影像點。又例如一個人站立而他的雙腳和身軀所造成的陰影投射到附近的地面或牆上時，倘若將臨界值設定為較大的值時，雖然陰影可以不被判斷為前景影像點，但此時與背景相近的前景影像點則很可能就被誤判為背景影像點；倘若將臨界值設定為較小的值時，則陰影大部份會被誤判為前景影像點。為了改善差異臨界值無法滿足不同情況的現象，我們將擴展每一影像點的資訊特徵成為除了自己所擁有的色彩特徵外，也將包含此點與其鄰近影像點所形成的紋理結構特徵。另外，我們也將原本是低階影像點處理的運作，改變成深具處理目的的高階物件運算處理。這些改變不但強化了影像特徵的區分能力，也運用了在人類認知過程中影像分離與物件偵測間相互影響的訊號增強模式，先找到影像畫面中含有重要資訊的區域(即有興趣的區域，Regions of Interests，也就是存在著特定物件影像的區域)，再針對這些區域以更具體的方式去執行前景/背景分離處理。相信本研究所提出的做法，能大幅改善前景物體影像的抽取結果，得到更正確、更完整、且更具處理目的的前景物件影像。

#### 參考文獻：

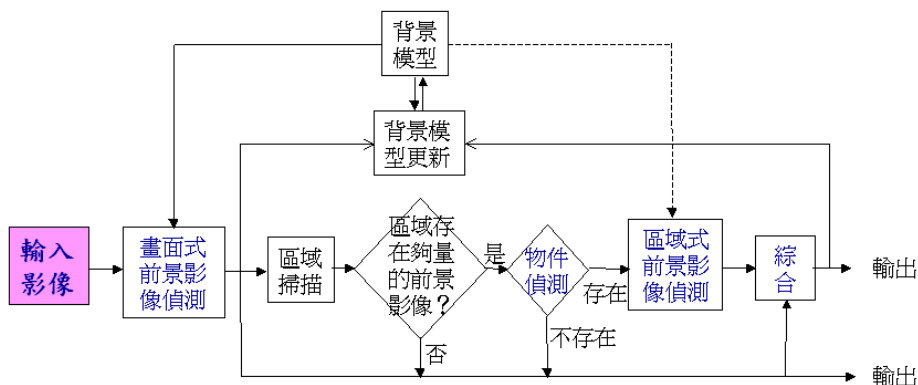
- [1] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body", IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, 1997.
- [2] R. Campbell and J. Krumm, "Object Recognition for an Intelligent Room", Computer Vision and Pattern Recognition, vol. 1, pp. 691-697, 2000.
- [3] I. Haritaoglu, D. Harwood and L. S. Davis, "W4 : Who? When? Where? What? A Real-Time System for Detecting and Tracking People", Proc. International Conference on Face and Gesture Recognition, April, 14-16, 1998.
- [4] J. Batista, P. Peixoto, and P. Araujo, "Real-Time Vergence and Binocular Gaze Control", IRSO907-IEEE/RS Int. Conf. On Intelligent Robots and Systems. Grenoble, France, September, (1997)
- [5] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt and L. Wixson, "A System for Video Surveillance and Monitoring", Tech. Rep., The Robotics Institute, Carnegie Mellon University, 2000. CMU-RI-TR-00-12
- [6] I. Haritaoglu, D. Harwood, and L. Davis, "Active Tracker: Surveillance with Active Camera", [http://www.umiacs.umd.edu/users/hismail/ActiveTracker\\_Outline.htm](http://www.umiacs.umd.edu/users/hismail/ActiveTracker_Outline.htm)
- [7] H. Sidenbladh and M. J. Black, "Learning the statistics of people in images and video", International Journal of Computer Vision. Vol. 54, Issue 1-3, pp. 183-209, Aug.-Oct. 2003.
- [8] C. Anderson, P. Burt, and G. V. D. Wal, "Change detection and tracking using pyramid transformation techniques", In Proc. Of SPIE-Intelligent Robotics and Computer Vision, Vol. 579, pp. 72-78, 1985.
- [9] P. H. Kelly, et al., "An architecture for multiple perspective interactive video", Proc. ACM Conf. Multimedia, pp.201-212, 1995.
- [10] Q. Cai, J. K. Aggarwal, "Tracking Human Motion in Structured Environments Using a Distributed Camera System", IEEE PAMI, Vol. 2, No. 11, pp.121-1247, Nov. 1999.

- [11] L. Lee, R. Romano, G. Stein, "Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frames," *IEEE Trans. on PAMI*, pp.758-768, Aug. 2000.
- [12] V. Kettner, R. Zabih, "Bayesian Multiple Camera Surveillance", *Proceeding of Computer Vision and Pattern Recognition*, For Collins, CO, pp.253-259, June 1999.
- [13] H. Pasula, et al., "Tracking Many Objects with Many Sensors," In *Proc. IJCAI-99*, Stockholm, 1999.
- [14] O. Javed, et al., "Tracking Across Multiple Cameras with Disjoint Views", *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 1-6, 2003.
- [15] S. Khan, et al., "Human Tracking in Multiple Camera", *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 1-6, 2002.
- [16] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques", *International Journal of Computer Vision*, vol. 12, no. 1, pp. 42-77, 1994.
- [17] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Condition Density", In *Proc. Of the 1996 European Conference on Computer Vision*, pp. 343-356, 1996.
- [18] Y. Ricquebourg and P. Bouthemy, "Real-Time Tracking Persons by Exploiting Spatio-Temporal Image Slices", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 792-808, 2000
- [19] C. Stauffer and W. E. L. Grimson, "Learning Patterns of Activity Using Real-Time Tracking", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747-757, 2000
- [20] F. Oberti, A. Teschioni, and C. S. Regazzoni, "ROC curves for performance evaluation of video sequences processing systems for surveillance applications", *Proc. 1999*, pp. 949-953.
- [21] L. Marcenaro, F. Oberti, and C. S. Regazzoni, "Change detection methods for automatic scene analysis by using mobile surveillance cameras", *Proc. 2000*, pp. 244-247.
- [22] L. Marcenaro, C. S. Regazzoni and G. Vernazza, "Automatic generation of the statistical model of a non-rigid object in a multiple-camera environment", *Proc. 2000*, pp. 530-533.
- [23] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin." A System for Video Surveillance and Monitoring." *Tech. Rep. The Robotics Institute, Carnegie Mellon University*, 2000. CMU-RI-TR-00-12.
- [24] I. Haritaoglu, D. Harwood and L. S. Davis."W4: Who? When? Where? What? A Real-Time System for Detecting and Tracking People." *Proc. International Conference on Face and Gesture Recognition*, April, pp. 14-16, 1998.
- [25] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body." *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 19, no. 7, pp. 780-785, 1997.
- [26] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246-252, 1999.
- [27] P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection", *2<sup>nd</sup> European Workshop on Advanced Video-Based Surveillance Systems*, 2001.
- [28] Z. Zivkovic and F. Heijden, "Efficient Adaptive Density Estimation per Image Pixel for the Task of Background Subtraction," *Pattern Recognition Letters*, vol. 27, pp.773-780, 2006.
- [29] S. C. Cheung and C. Kamath, "Robust Background Subtraction with Foreground Validation for Urban Traffic Video," *EURASIP Journal on Applied Signal Processing*, vol.14, pp. 2330-2340, 2005.
- [30] C. Anderson, P. Burt, and G. V. D. Wal, "Change Detection and Tracking Using Pyramid Transformation Techniques," In *Proc. of SPIE Intelligent Robotics and Computer Vision*, Vol. 579, pp. 72-78, 1985.
- [31] Y. Altunbasak, P. E. Eren and A. M. Tekalp, "Region-Based Parametric Motion Segmentation Using Color Information", *Graphical Models and Image Processing: GMIP*, 60(1), pp. 13-23, 1998.
- [32] J. G. Choi, S. W. Lee and S.D. Kim, "Spatio-Temporal Video Segmentation Using a Joint Similarity Measure", *IEEE Transactions on Circuits and Systems for Video Technology*, 7(2), pp. 279-286, 1997.
- [33] F. Dufax, F. Moscheni and A. Lippman, "Spatio-Temporal Segmentation Based on Motion and Static Segmentation," *IEEE Conference on Image Processing*, pp. 306-309, 1995.
- [34] Y. Tsaig and A. Averbuch, "Automatic Segmentation of Moving Objects in Video Sequences: A Region Labeling Approach," *IEEE Transactions on Circuits and Systems for Video Technology*, 3(5), pp. 597-612, 2002.
- [35] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian Detection of Independent Motion in Crowds," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 594-601, 2006.

- [36] B. Chen and Y. Lei, "Indoor and Outdoor People Detection and Shadow Suppression by Exploiting HSV Color Information," Proceedings of the Fourth International Conference on Computer and Information Technology, pp. 137-142, 2004.
- [37] C. T. Hsieh, E. Lai, Y. K. Wu and C. K. Liang, "Robust Real Time Tracking with Shadow Removal in Open Environment," 5<sup>th</sup> Asian Control Conference, pp. 901-905, 2004.
- [38] Y. Matsushita, K. Nishino, K. Ikeuchi and M. Sakauchi, "Shadow Elimination for Robust Video Surveillance," Proceedings of the Workshop on Motion and Video Computing, pp. 15-21, 2002.
- [39] A. Leone, C. Distanto and F. Buccolieri, "A Shadow Elimination Approach in Video-Surveillance Context," Pattern Recognition Letters, vol. 27, pp. 345-355, 2006.
- [40] A. Monnet, A. Mittal, N. Paragios, and V. Ramesh, "Background Modeling and Subtraction of Dynamic Scenes," Proc. IEEE International Conference on Computer Vision, pp. 1305-1312, 2003.
- [41] M. Heikkila and M. Pietikainen, "A Texture-Based Method for Modeling the Background and Detecting Moving Objects," PAMI, vol. 28, pp. 657-662 2006.
- [42] T. Ojala et al., "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns," PAMI, 24(7), pp. 971-987, 2002.
- [43] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascaded of Simple Features", in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 511-518, 2001.
- [44] B. Frba and A. Ernst, "Face Detection with the Modified Census Transform", Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (AFGR), Seoul, 2004, pp. 91-96.
- [45] P. Wilson and J. Fernandez, "Facial feature detection using Haar classifiers," Journal of Computing Sciences in Colleges, 21(4), pp. 127-133, 2006.
- [46] D. Cristinacce and T. Cootes, "Facial feature detection using AdaBoost with shape constraints," British Machine Vision Conference, 2003.
- [47] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55(1), pp. 119-139, 1997.

#### 四. 研究方法：

本計畫首先採用一般的前景影像抽取方法(可以選用任何一種已知的前景影像抽取技術)來對整張輸入影像產生與背景模型差異較大的前景影像點。事實上，此前景影像抽取的過程中亦可包含一些常見的影響處理步驟(例如陰影判斷和型態學運算等)，以便將一些可能是雜訊所造的前景影像點刪除。接下來，再將佔有足夠前景影像比例的特定形狀區域(可能為矩形，橢圓形或其它形狀)進行物件偵測處理。假如某一區域具有足夠多的前景影像而且它也被判斷為擁有某種物件的訊號，則此區域會以較寬鬆的條件來進行再一次的區域式前景影像抽取。最後，則將前後兩次的前景影像之抽取結果綜合在一起來得出最終的前景影像輸出結果。依據本專利的精神，可以設計出多種不同的實施模組，圖一為一種我們所實施方法的處理流程架構，可用來說明本專利的主要概念和處理步驟。以下為此架構之重要處理模組的描述：



一種結合物件偵測之多階段前景影像抽取方法 5/ 12

圖一、一種用來實施本計畫的處理流程架構

A. 畫面式前景影像偵測模組：

本模組可採用任何一種常用的前景影像偵測演算法(例如：背景相減或是畫面差異法等)。本模組主要目的是用來指出在整張影像中有哪些部分是屬於較明顯的前景影像，因此可選用較嚴謹的參數，並且亦可與其它雜訊處理(如陰影判斷和形態學運算等處理)搭配，使得大部分的雜訊和陰影都不會被誤判為前景影像點。在此，我們採用最常見的背景相減法。假設  $I(P)$  代表影像中第  $P$  點的特徵訊號值(如顏色或灰度值)， $B(P)$  為  $P$  點的背景模型特徵訊號值，則前景/背景的判斷方程式  $D(P)$  為

$$D(P) = \begin{cases} 1 & , \text{when } |I(P)-B(P)| > \text{thresh;} \\ 0 & , \text{others.} \end{cases} \dots\dots\dots(1)$$

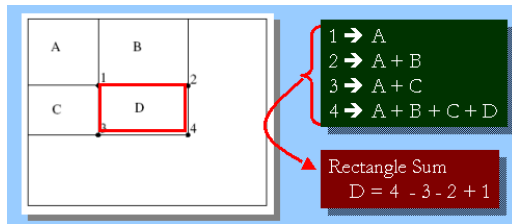
其中， $\text{thresh}$  為一個預設的差異臨界值。此方程式主要說明當  $I(P)$  和  $B(P)$  的特徵訊號差異夠大時，影像點  $P$  會被判斷為前景點(即  $D(P)=1$ )；反之，當  $I(P)$  和  $B(P)$  的特徵訊號差異不夠大時，影像點  $P$  則會被判斷為背景點(即  $D(P)=0$ )。當然，當  $D(P)$  為 1 時，可再用陰影判斷方式來判別點  $P$  是否屬於陰影狀態，當點  $P$  被認為是在陰影狀態時，則點  $P$  則會被更改成為背景點(即  $D(P)$  改設為 0)。

B. 區域掃描模組：

本模組採用某種特殊形狀的影像區域以由左至右由上到下的方式來逐一掃描整張影像，當某一區域影像中擁有足夠的前景影像訊號時，則代表此區域影像有可能存在某種前景物件，因而需要將此區域影像進一步的物件偵測模組來判斷是否有真正物件的存在。當一區域影像中並沒有夠量的前景影像時，則此區域影像會被認為並不存在前景物件，因此被忽略而再去掃描下一個區域影像。圖二顯示一種可以用來快速計算矩形區域內所有點之訊號總和的方法，它首先計算此影像的集成影像(Integral Image)。假設  $f(x,y)$  是一  $M \times N$  影像平面中座標為  $(x,y)$  之影像點之訊號值，其中  $1 \leq x \leq M$  而且  $1 \leq y \leq N$ ； $II(x,y)$  為此影像中以  $(1,1)$  點為左上角而點  $(x,y)$  為右下角之矩形內所有影像點之訊號總和，也就是  $II(x,y) = \sum_{1 \leq a \leq x} \sum_{1 \leq b \leq y} f(a,b)$ 。利用迭帶式的計算， $II(x,y)$  可以被快速地計算如下：

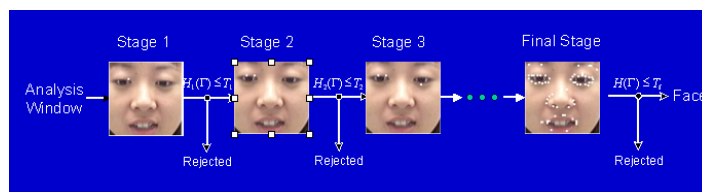
$$\begin{aligned} s(x,y) &= s(x,y-1) + f(x,y); \\ II(x,y) &= II(x-1,y) + s(x,y); \\ s(x,0) &= 0; \\ II(0,y) &= 0. \end{aligned}$$

在得到集成影像之後，若要計算矩形  $D$  之訊號總和，則只要先將點 4 之集成影像的對應值(即  $\Pi(4)$ )加上點 1 之集成影像的對應值(即  $\Pi(1)$ )，然後再減去點 2 和點 3 之集成影像的對應值(即  $\Pi(2)$  和  $\Pi(3)$ )即可，也就是矩形  $D$  的訊號總和  $= \Pi(4) + \Pi(1) - \Pi(2) - \Pi(3)$ 。由圖二的圖示可清楚的了解上述計算的基本原理。



圖二、矩形訊號集成的快速演算示意圖

- C. 物件偵測模組：本模組判斷某一區域影像是否存在著某種特定物件。目前，我們採用速度和效能均優異的 Adaboost 演算法來實現本模組。這種促進式(Boosting)演算法已經非常成功地被應用到人臉偵測、車輛偵測以及人形偵測等方面。經由學習過程，這種演算法能夠自動地建構出多階層式(Cascaded)物件過濾器(Filter)，其中每一個階層都會濾除掉一些不具有物件的區域。一個影像區域只有成功地通過所有的階層判斷才會被視為是存在物件的區域。圖三為一個可偵測人臉物件的多階層式人臉過濾器。由於每一階層都只是用簡單的特徵來進行判斷，而且大部分不存在前景物件的區域在前面少數階層中即會被過濾出來，因此 Adaboost 演算法具有運算快速的特性。



圖三、一個多階層式人臉過濾器

D. 區域式前景影像偵測模組：

當某一區域影像擁有夠量的前影像訊號以及此區域影像又被判別存在著某種特殊物件時，本模組才會被執行。由於本模組是針對已被判別存在有物件的區域來處理，所以它的對象和範圍都是有限的。因此可用一些自然界中存在的物理現象為條件來強化本模組的處理效能，例如：

- a、在此區域範圍內，中心部分的影像較邊界部分的影像應該具有更高的物件歸屬機率；
- b、在此區域範圍內，下面的影像較上面的影像應該具有更大陰影發生的機率。

因此，在此區域範圍內每一點的差異臨界值可依對物件的瞭解而有不同的設定，圖四為一種簡化的示意說明，主要顯示對於存在人形的區域而言，其中間白色矩形之內部影像較此矩形之外部影像可使用較小的差異臨界值。





圖四、人形區域不同位置的影像可對應不同的差異臨界值之示意圖，例如中間白色矩形之內部影像較此矩形之外部影像可使用較小的差異臨界值。

另外，當已知某一區域內有存在物件時，為了讓所抽取的物件較完整，可選用的差異臨界值較畫面式前景影像偵測模組所選的值為小。而且，為了執行速度的考量，在具有物件訊號之區域的影像點中，若是某一影像點已於畫面式前景影像偵測時被判斷是前景影像，則此點不需要於本模組中再進行前景/背景的判斷。

註解 [A1]:

E. 綜合：

本模組是將畫面式前景影像偵測模組和區域式前景影像偵測模組所各自找到的前景影像點綜合起來產生最終的輸出結果。在此，我們用最簡單的”OR”運算來完成，那就是在畫面式和區域式前景影像偵測處理時，只要在任何一種偵測模組中被判斷為前景影像點，則此點即被認定是前景影像點。

F. 背景模型更新模組：

當選用背景相減法來判斷前景影像點時，如何保持即時更新和有效的背景模型是一個非常重要的議題。因為不良的背景模型是不可能產生滿意的前景影像偵測結果。由於偵測過程中可以得到許多方面的判斷資訊(如存在物件的影像區域、前景影像點、背景影像點等)，因此我們的背景模型更新能更多樣且有效，例如

$$B(P) = \begin{cases} (1 - \alpha) B(P) + \alpha F(P) & , \text{當} P \text{ 被判斷為背景點且} P \text{ 不存在於具有物件的影像區域內} \\ (1 - \beta) B(P) + \beta F(P) & , \text{當} P \text{ 被判斷為背景點且} P \text{ 存在於具有物件的影像區域內} \\ (1 - \gamma) B(P) + \gamma F(P) & , \text{當} P \text{ 被判斷為前景影像點} \end{cases}$$

(2)

其中， $B(P)$ 為影像點  $P$  的背景模型， $F(P)$ 為輸入畫面之影像點  $P$  的特徵值，而且  $\alpha > \beta > \gamma$ 。有些時候，雖然真正是屬於物件的前景影像點，可能會因其訊號特徵與背景模型相似而被錯誤地判斷為背景點，我們的背景更新模組就會因為它是存在於具有物件的影像區域內而能不急著去更新它，這使得我們所建構出的背景模型能與前景影像具有更大分辨的能力。

## 五. 結果與討論：

為了驗證所提方法的有效性，我們選用一段總共包含 150 個畫面的影片來做實驗，其中前面 30 個畫面用來訓練初始背景模型，後面 120 個畫面才是用來測試。所有測試的畫面都有人工標註的標準答案(稱為”剪影”)，例如圖五顯示一些測試畫面以及其相對的標準答案。

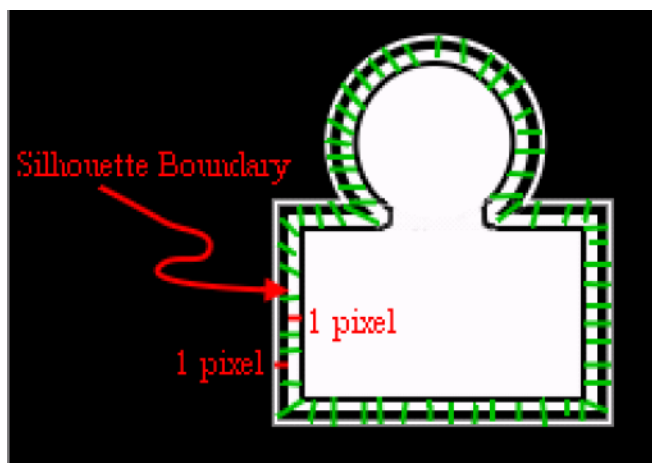


圖五、三張視訊影像和他們的剪影圖

為了降低剪影邊界的不確定性，在剪影輪廓邊緣正負一個像素點(pixel)範圍內的區域將不列入計分，此區域 R 定義如下：

$$R = \left\{ (x, y) \mid \sqrt{(x-u)^2 + (y-v)^2} \leq 1, (u, v) \in \text{Silhouette Boundary} \right\}$$

以下圖為例，綠色區域將不計分。



圖六、位於剪影輪廓邊緣正負一個像素點範圍內的區域示意圖

若程式所產生的剪影圖片以  $O(x,y)$  表示，人工所產生的剪影圖片以  $S(x,y)$  表示。其中，剪影部份以“1”表示，背景部份以“0”表示，剪影輪廓邊緣區域以  $R$  表示，而  $I$  是整張影像，則評分方式定義為

$$E = \frac{1}{N(I)} * \sum_{(x,y) \in I \text{ and } (x,y) \notin R} O(x,y) \oplus S(x,y) \dots\dots\dots(3)$$

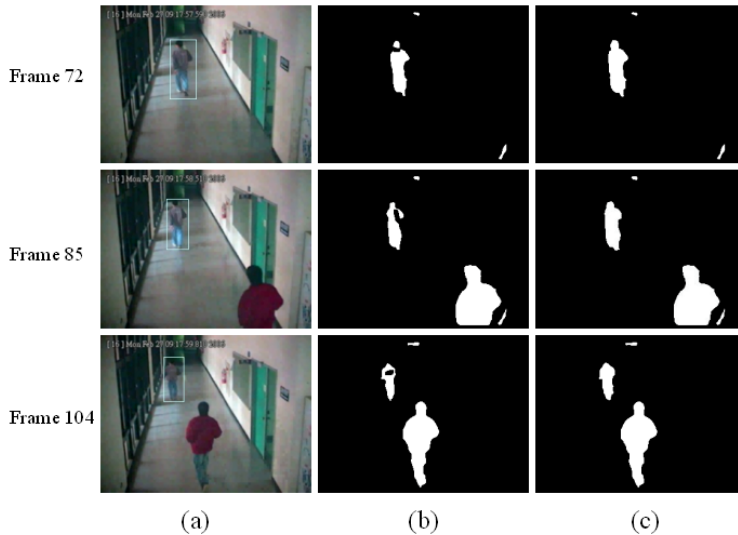
其中， $N(I)$  為影像中所有不屬於  $R$  的影像點數。所以  $E$  的值是越小越好。


實驗是在 Pentium IV 2.0 GHz with 512M RAM 的桌上型電腦上執行。除了  $E$  值的比較數據外，執行的速度亦被列出為參考數值。表一為利用背景相減法以及本計畫方法所得到的結果，其中背景相減法是以方程式(1)來判別前景與背景影像點，而所選用的參數  $thresh(=40)$  是對應到最低的  $E$  值(0.005657)。

Method	E	sec/320*240 image	sec/160*120 image
Background Subtraction	0.00771	180ms	50ms
The Proposed method	0.00362	304ms	78ms

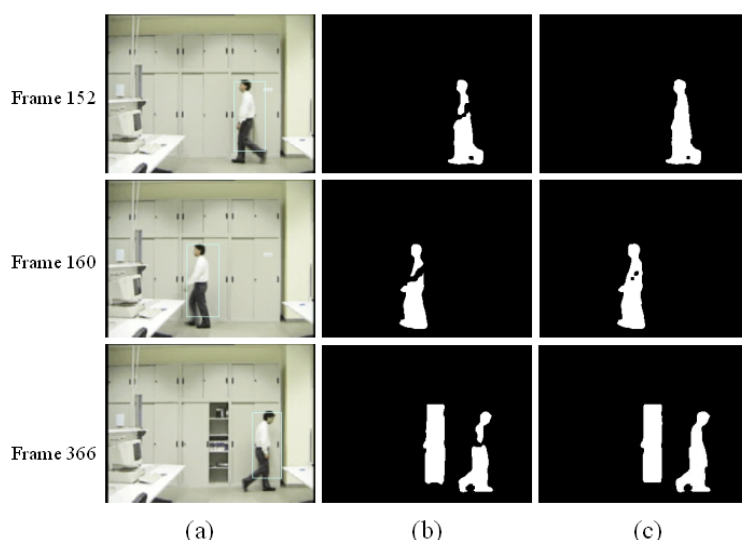
表一、背景相減法以及本計畫方法所得到的前景影像抽取效能比較

圖七即為一視訊影片中三張影像的前景抽取實驗的結果，其中圖七(a)為輸入視訊影像和自動偵測到的人形矩形外框，圖七(b)為傳統前景影像抽取結果，圖七(c)為本計畫的前景影像抽取結果。圖八為另一段視訊影片中三張影像的前景抽取實驗的結果。以上之實驗數據和抽取影像之顯示結果清楚展示本計畫具有抽取較正確與完整的前景影像的能力。



註解 [A2]:   
 註解 [A3]:

圖七、一視訊影片中三張影像的前景抽取實驗的結果



圖八、另一視訊影片中三張影像的前景抽取實驗的結果

### 結論：

本計畫提出一種新穎性的前景影像抽取方法，它將傳統的前景影像抽取與物件偵測等不同的處理模組進行巧妙的結合，可得到更正確與完整的前景影像抽取結果。首先，以傳統的前景影像抽取模組來對整張影像產生與背景影像模型具有足夠差異量的第一階段畫面式前景影像點，再對具有足夠前景影像點的區域進行判斷其是否存在某些特定類別的物件(如人形或車形等)；接著，針對被判斷為具有特定類別物件的區域進一步的抽取其第二階段的區域式前景影像點；最後，將前後兩次(即第一階段和第二階段)所抽取的前景影像點進行整合而得出最後的前景偵測輸出影像。本計畫主要擁有四方面的處理優點：(1)對陰影具有較大的容忍能力，(2)陰影所產生的影響主要被限制於存在前景物件的影像區域範圍內，(3)對與背景顏色相近的前景物件具有較佳的抽取能力，(4)可得到較佳的背景影像更新模型，以及(5)可得到較正確與完整的前景影像。本專利方法容易實施，具有高度實用性。

本計畫成果已發表在 2007 International Conference on Artificial Intelligence and Pattern Recognition (2007/7/9~2007/7/12)，論文名稱為”Local Structure Based Foreground Object Extraction”(附錄一)

### 計畫成果自評部份：

本計畫研究內容與原計畫之規畫相符，採用以物件為導向的觀點，也設計出

一套能更正確和更完整的抽取前景物件影像的方法，已達成當初所預期的計畫目標。本計畫的研究成果除了已發表在國際性會議論文外，我們將會繼續做更完整了實驗，希望能將計畫成果往學術期刊發表。

# Local Structure Based Foreground Object Extraction

Yea-Shuan Huang

*Department of Computer Science & Information Engineering of Chung Hua University, Taiwan*

*Email: yeashuan@chu.edu.tw*

Cheng-Yuan Tang

*Department of Information Management of Huaan University, Taiwan*

## Abstract

*This paper proposes a novel foreground object extraction approach, which combines two kinds of feature information: image color and local image structure. Due to the highly complementary characteristic of the two used features, the proposed approach has a fairly good foreground object extraction ability when being applied to various situations. Compared with other commonly-used methods (such as color-based background extraction, temporal difference and optical flow), this method has shown two major advantages: (1) it can extract rather correct and complete object image even when the foreground objects and their background present similar colors; and (2) it performs much robust under the influence of lighting variation and shadow. Especially, this method is easy to implement and has a real-time execution performance. Consequentially, this method has a high practicability to various applications.*

## 1. Introduction

Foreground extraction from video sequence is important in many tasks, such as video surveillance, face recognizing, object tracking...etc. For all of these applications, foreground extraction is the first step. If the foreground objects can be extracted successfully, it can significantly help the following processes such as object tracking or object recognition and achieve more correct and robust results.

Over the past several decades, many approaches have been developed for foreground extraction. Three most general approaches are background subtraction [1,2,3], temporal difference [4] and optical flow [5]. The background subtraction approach is to subtract a current image from the referenced background image. It is in principle divided into two stages: the learning stage and the testing stage. In the learning stage, the background image is constructed from several pre-collected images without any moving object. Then in the test stage, the color difference is calculated from the input image and the background image pixel by pixel. It is needed to determine a suitable threshold, so pixels will be classified

as foreground or background. Similarly, temporal difference also computes the difference of image pixels from temporally succeeding input images. If the absolute value of difference is above the threshold, the pixel is attributed as foreground. The optical flow approach is to calculate the movement of every pixel between frames. If there exists object movement, some obvious changes of optical flow happen.

However, there are some problems in these approaches. The computation complexity of optical flow approach is large, so it is not suitable for real-time systems. When using background subtraction approach, shadow and light changes may cause the image difference significant enough to make wrong decision. Also when the color of foreground object is close to that of background, the object will be hard to extract correctly and completely. The light changing effect is reduced in temporal difference approach but if the object is static it can't be extracted at all.

We therefore proposed a new method of foreground extraction that uses a designed image structure and color as features. We define the "contrast of neighboring pixels" which represents the relation between neighboring pixels, and combine several kind of relations to represent image structure. We call it "local image structure". Experiments have shown evidently that the results of the proposed foreground extraction performs much better than all the commonly-used three approaches.

The outline of this paper is as follows. In section 2 we describe how to construct image structure and perform matching. A foreground extraction approach by using both image color and structure information is proposed in section 3 and some experiment results are presented in section 4. Finally, in section 5 we summarize our conclusions.

## 2. Local Image Structure and Matching

Based on the imaging optics, a contrast operator is designed to reflect the constituent relation (larger than or not larger than) between two image pixels. This relation records a rough but stable image formation of two pixels. Combining the relation information among one pixel and

a set of its selectively neighboring pixels, a texture-like feature can be derived which is called the local image structure (LIS) of this pixel. Apparently, LIS is a useful image feature description. This section will first introduce LIS in detail, and an effective matching algorithm based on LIS will also be presented.

## 2.1. Local Image Structure

Let  $x$  be an image pixel,  $I(x)$ ,  $R(x)$  and  $L(x)$  be individually the image intensity, the reflectance vector and the illumination vector of  $x$ . From the imaging optics, it exists  $I(x) = R(x) \cdot L(x)$ . This equation describes the image intensity is just the product of image reflectance and its illumination vectors. For another image pixel  $y$ , it becomes  $I(y) = R(y) \cdot L(y)$ . Therefore,

$$\frac{I(x)}{I(y)} = \frac{R(x) \cdot L(x)}{R(y) \cdot L(y)}. \quad (1)$$

Suppose pixels  $x$  and  $y$  are neighbors, being close to each other, so  $L(x)$  and  $L(y)$  are assumed to be similar or even the same. Then

$$\frac{I(x)}{I(y)} \approx \frac{R(x)}{R(y)}. \quad (2)$$

This equation shows that the ratio of two neighboring pixels is almost independent of illumination vector, and it approximately keeps constant if only the background illumination changes and all other factors remain unchanged. However, besides illumination there are other factors (such as surface normal and noises) which can affect the image intensity. Therefore, it is not proper to directly take the image ratio as features. Instead the contrast relationship (larger than or not larger than) is a better and more stable feature description in representing the relation among image pixels. For any two pixels  $x$  and  $y$ , a contrast relationship is defined as

$$\zeta(I(x), I(y)) = \begin{cases} 0, & \text{if } I(x) \geq I(y); \\ 1, & \text{otherwise.} \end{cases} \quad (3)$$

The relationship obtains a value 0 if pixel  $x$  is not darker than pixel  $y$ , otherwise it obtains a value 1. Let  $\Phi(x) = \{P_0, P_1, \dots, P_N\}$  be an  $N$ -element set and each element correspond to a chosen neighboring pixel of  $x$ . This set is called interestedly selected pixel set, which specifies how many and what pixels are chosen to derive the local image structure. It is not difficult to realize that by integrating a set of relationship  $\zeta(I(x), I(p_n))$ ,  $1 \leq n \leq N$  among pixel  $x$  and its neighboring pixels ( $P_1, \dots$ , and  $P_N$ ) a structure-like information can be constructed. In fact, the structure-like information semantically represent one kind of image texture information. With pixel

relationships, the local image structure  $\Gamma(x)$  is designed as

$$\Gamma(x) = \sum_{n=0, p_n \in \Phi(x)}^N 2^n \times \zeta(I(x), I(p_n)) \quad (4)$$

Figure 1 is an example of  $\Phi(x)$  which uses 8 neighbors of  $x$  to construct its local structure. Obviously, with this configuration the corresponding  $\Gamma(x)$  has in total 256 possibilities ranging from 0 to 255, and each possibility corresponds to one specific image structure. For example, when  $\Gamma(x)$  is 0, it means that  $x$  is the brightest pixel among its 8 neighbors. For another example, when  $\Gamma(x)$  is 7, it means that among its 8 neighbors only  $P_0, P_1$  and  $P_2$  are darker than  $x$ . This structure representation is very robust to many kinds of variations, especially the illumination variation. Figure 2 shows an illustrative example of the local structure to the images under various illuminations, which is visualized as index image where the structure index determines the pixel intensity. Figures 2(b)-(e) are generated from Figure 2(a) by using an image processing software with the following parameters: light value  $-75$ , contrast value  $-50$ , gamma value 2.0, and gamma value 0.4, respectively. It reveals that the designed image structure is not affected by the illumination variation.

$P_3$	$P_2$	$P_1$
$P_4$	$X$	$P_0$
$P_5$	$P_6$	$P_7$

Figure 1: One example of the interestedly selected pixel set  $\Phi(X)$ , where 8 neighbors of  $X$ ,  $P_0, P_1, P_2, P_3, P_4, P_5, P_6$  and  $P_7$ , are chosen to construct the local image structure of pixel  $X$ .

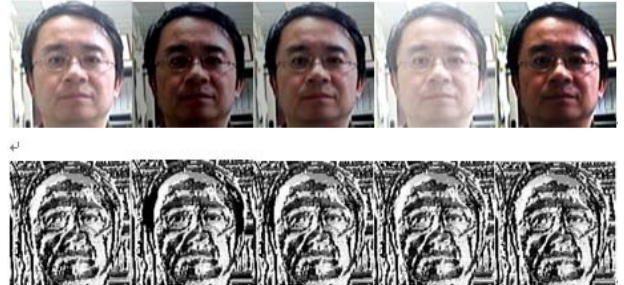


Figure 2: Images with different illumination are shown in the first row, and the corresponding images of derived local structures are shown in the second row.

In order to obtain a richer structure information, not only one but also several interestedly selected pixel sets can be designed. That is M sets ( $\Phi_1, \dots, \Phi_M$ ) are used to construct M kinds of local image structures, which is certainly more powerful in distinguishing different image structures. Figure 3 shows an example with  $M = 3$ , where pixels with indices 1, 2 and 3 are chosen for  $\Phi_1$ ,  $\Phi_2$ , and  $\Phi_3$  respectively. It is clear that the set number of LIS and the constituent pixels of each set can be very flexibly designed in order to produce the best performance for various encountered applications.

2	3	2	3	2
3	1	1	1	3
2	1	X	1	2
3	1	1	1	3
2	3	2	3	2

Figure 3: An example of designing a 3-set local image structures, where pixel  $x$  is the kernel processed pixel, pixels with indices 1, 2 and 3 are chosen for  $\Phi_1$ ,  $\Phi_2$ , and  $\Phi_3$  respectively.

## 2.2. Matching of Local Image Structure

Due to different factors of variations, the image intensity of the monitored background probably will change slightly from time to time. Therefore, for pixel  $x$ , its corresponding LIS may be not unique. For constructing the representative background LIS, a training stage from a set of collected images not containing any foreground objects should be performed. After training for each  $\Phi_m$ , the number  $r$  of appeared image structures, the derived structure values  $\Gamma_{mi} (1 \leq i \leq r)$  and their appeared probabilities  $\pi_{mi} (1 \leq i \leq r)$  can be obtained straightforwardly for each image pixel. Let  $S_m$  denote the trained result of  $\Phi_m$ . Then  $S_m(x) = \{(\Gamma_{mi}, \pi_{mi}) | 1 \leq i \leq r, \text{ and } \pi_{mi} \geq \pi_{m,i+1} \geq 0\}$ , where  $\pi_{mi}$  is the probability of  $\Gamma_{mi}$  and  $\sum_{i=1}^r \pi_{ji} = 1$ . If there are M interestedly selected structure sets, it will certainly generate M training results,  $S_1, \dots, S_M$ .

To extract objects, an input image has to first compute its M kinds of LIS based on specified  $\Phi_1, \dots, \Phi_M$ . Let  $t_1, \dots, t_M$  denote the corresponding computed outcomes, where  $t_m$  is derived from  $\Phi_m$ . Then a measurement  $Q$  to index the degree that a pixel  $x$  belongs to background is computed as

$$Q = \sum_{m=1}^M w_m \left(1 - \frac{G(S_m, t_m)}{N_m}\right) \quad (5)$$

where  $w_m$  and  $N_m$  denote respectively  $\Phi_m$ 's contribution weight and element number, and  $G(S_m, t_m)$  computes the least number of bit difference between  $t_m$  and  $\Gamma_{mi}$  with  $1 \leq i \leq r$ , that is

$$G(S_m, t_m) = \min_{i=1}^r \text{BitCount}(\Gamma_{mi} \oplus t_m) \quad (6)$$

where  $\oplus$  is the bit exclusive OR operation. Obviously, by definition  $\text{BitCount}(y)$  computes the number that the bit of  $y$  is not 0. It is worthwhile to mention that when  $t_m$  belongs to  $S_m$ ,  $t_m$  is then the same as one instance of  $\Gamma_{mi}$  and consequently  $\min_{i=1}^r \text{BitCount}(\Gamma_{mi} \oplus t_m)$  becomes 0.

It is clear that the value of  $G(S_m, t_m)$  is between 0 and  $N_m$ , and to divide  $G(S_m, t_m)$  with  $N_m$  makes the ration between 0 and 1. For an unknown pixel  $x$ , if it is a background pixel, its derived structure values  $t_1, \dots, t_M$  will very probably coincide with some values existed in  $\Gamma_1, \dots, \Gamma_m$  respectively. According to Equation (5),  $Q$  becomes a large value approaching 1.0. On the contrary, if pixel  $x$  belongs to a moving object, its derived structure values  $t_1, \dots, t_M$  will have high possibility to be different to all values existed in  $\Gamma_1, \dots, \Gamma_m$ . Consequentially,  $Q$  becomes a comparatively small value approaching 0. Therefore, from the derived value of  $Q$ , pixel  $x$  can be attributed to background or foreground. In essence, the larger the value of  $Q$  is, the more probable a pixel belongs to background. For avoiding rarely appeared image structures decrease the discrimination ability of foreground and background, only those structures with large enough appearance probabilities are used to compute  $G(S_m, t_m)$ , that is

$$G(S_m, t_m) = \min_{i=1}^{r'} \text{BitCount}(\Gamma_{mi} \oplus t_m) \quad ,$$

where  $\pi_{mi} \geq \text{thresh1}$  (7)

where  $r'$  is the total number that their individual appearance probabilities are larger than  $\text{thresh1}$ . Beneficially, by using only the frequently appeared image structures,  $G(S_m, t_m)$  can be computed faster.



## 2.2. Extraction by Using Image Intensity and Structure

Besides image structure information, image color can be regarded as the most important image feature. Therefore, it is really appropriate to use both color and structure information to extract foreground objects. For color, the background model can be simply constructed by using a mixture Gaussian model (GMM) from a set of collected training background images. Let  $f$  denote the color value of pixel  $x$ ,  $\lambda$  denote the constructed color background model, *i.e.*  $\lambda = \{p_i, \mu_i, \Sigma_i\}$ ,  $i = 1, 2, \dots, C$  where  $C$  is the total number of mixture models, and  $p_i$ ,  $\mu_i$  and  $\Sigma_i$  are the weight, mean and covariance variance of mixture model  $i$  individually. The GMM probability  $p(f | \lambda)$  that pixel  $x$  belongs to background becomes

$$p(f | \lambda) = \sum_{i=1}^C \frac{p_i}{(2\pi)^{N/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(f - \mu_i)' \Sigma_i^{-1} (f - \mu_i)\right\} \quad (8)$$

Combining Equations (5) and (8), an integrated likelihood  $LK(f, T | \lambda, S_1, \dots, S_M)$  indicating  $x$  to be background becomes

$$LK(f, t_1, \dots, t_M | \lambda, S_1, \dots, S_M) = w_c * p(f | \lambda) + w_s * \sum_{m=1}^M w_m * \left(1 - \frac{G(S_m, t_m)}{N_m}\right) \quad (9)$$

where  $f$  is the color value of  $x$ ,

$t_1, \dots, t_M$  are the  $M$  structure values of  $x$ ,

$S_1, \dots, S_M$  are the  $M$  sets of structure statistics of  $x$ ,

$\lambda$  is the color background GMM model,

$N_m$  is the element number of  $\Phi_m$ ,

$w_m$  is the contribution weight of  $\Phi_m$ ,

$w_c$  is the combination weight of color information,

$w_s$  is the combination weight of structure information.

With a suitable threshold  $T$ , the decision of  $x$  becomes

$$D(x) = \begin{cases} 0 & , \text{ when } LK(f, t_1, \dots, t_M | \lambda, S_1, \dots, S_M) \geq T; \\ 1 & , \text{ otherwise.} \end{cases} \quad (10)$$

When  $D(x)=0$ ,  $x$  is assigned to be a background pixel, and when  $D(x)=1$ , then  $x$  is assigned to be a foreground pixel.

## 3. Experiments

In order to find out the extraction ability of the proposed method, two experiments were performed. In the first, videos are taken from an office site and in total there are 350 image frames. Among them, the first 140 frames contain only pure background scene. To construct the respective background models of image color and image structure, the first 120 image frames are used in the training stage. For the color model, a single Gaussian model is adopted. For the structure model, three assignment sets,  $\Phi_1$ ,  $\Phi_2$  and  $\Phi_3$ , are selected to generate three local image structures. The used  $\Phi_1$ ,  $\Phi_2$  and  $\Phi_3$  are shown in the Figure 3. Therefore,  $\Phi_1$ ,  $\Phi_2$  and  $\Phi_3$  contain the same number of elements, and each has 8 pixel elements. So, each local structure  $\Gamma_{mi}$  can be represented effectively with just one byte. Because the pixels of  $\Phi_1$  are closer to the kernel  $x$  than those of  $\Phi_2$  and  $\Phi_3$ , the value of  $\Phi_1$  will be more representative for image structure than those of  $\Phi_2$  and  $\Phi_3$ . As a result, their contribution weights will be set differently. In this experiment,  $w_1$ ,  $w_2$  and  $w_3$  are set to be 0.4, 0.3 and 0.3 respectively. To speed up the process and reduce the ill effect of rarely appeared structures, Equation (6) are used to compute  $G(S_m, t_m)$  and *thresh1* is set to 0.1. Finally, the combination weights  $w_c$  and  $w_s$ , and the decision threshold  $T$  are set to 0.4, 0.6 and 0.6 respectively. Figure 4 shows the extracted results with different foreground extraction approaches which contain color-only background subtraction, temporal difference, structure-only background subtraction, and the proposed approach. Visually, extraction by temporal difference is usually broken and the extracted image mainly focus on both object boundary and image edges. Although both color-only and structure-only background subtraction approaches obtain better results, but they still contain a few broken areas inside the foreground objects. Interestingly, the broken areas of the two approaches are seldom coincided with each other. This phenomenon displays color and structure information are mutually independent so that their extraction results have a complementary effect. This results in that to combine both of them can produces much complete and correct extracted objects as shown in Figure 4(f). Figure 5 shows the second experiment with videos taken from outdoor hallway. In this video, there is one person who approaches the picture taken camera and another person who lingers around at a distance. Again, extraction by using both color and structure information produces the best result. From the above two experiments, the effectiveness of the proposed approach has been clearly demonstrated.

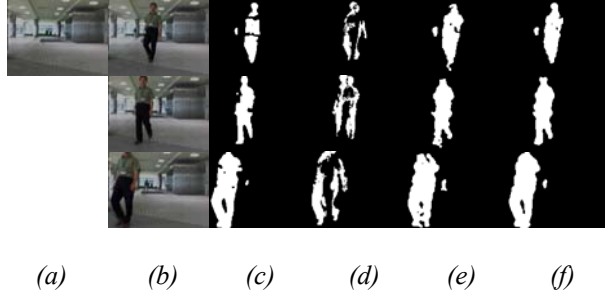


Figure 4: Extracted foreground results of videos taken at an office site by using several approaches, (a) is a background image, (b) shows three video frames, (c) shows the corresponding results of color-only background subtraction, (d) shows the corresponding results of temporal difference, (e) shows the corresponding results of structure-only background subtraction, (f) shows the integrated results of using both color and structure background subtraction.

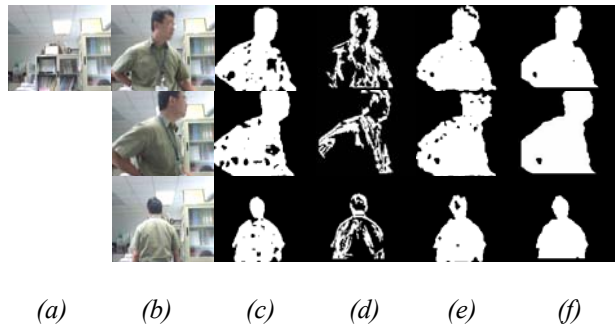


Figure 5: Extracted foreground results of videos taken at an outdoor hallway site by using several approaches, (a), (b), (c), (d), (e) and (f) denote the same image types as Figure 4.

#### 4. Conclusion

An novel foreground object extraction method has been proposed in this paper which makes use of both kinds of image features, image color and image texture, to attribute one pixel into foreground or background. In the past, color is often applied alone to extract the foreground objects. But color is sensitive to the illumination variation and shadow. If a moving object presents similar color to the background image, color indeed is not discriminative enough to separate the object in the integrity. However, structure feature could be quite different among color-alike image regions. With this understanding, the proposed method combines both of them, color and structure, to extract foreground objects. The designed image structure feature can be computed very efficient in the sense of processing speed and memory requirement. Experiments on videos taken from an office and an outdoor hallway sites have shown that both features

perform complementary to each other. The extracted results of our method indeed can obtain much complete and correct foreground object images even when there exists illumination variation, shadow and color approximate between objects and background.

#### 5. Acknowledge

This work is supported by a Taiwan NSC research grant (Project Code: NSC 95 - 2218 - E - 216 - 004).

#### 6. References

- [1] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin. "A System for Video Surveillance and Monitoring." Tech. Rep. The Robotics Institute, Carnegie Mellon University, 2000. CMU-RI-TR-00-12.
- [2] I. Haritaoglu, D. Harwood and L. S. Davis. "W4: Who? When? Where? What? A Real-Time System for Detecting and Tracking People." Proc. International Conference on Face and Gesture Recognition, April, pp. 14-16, 1998.
- [3] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-Time Tracking of the Human Body." IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 19, no. 7, pp. 780-785, 1997.
- [4] C. Anderson, P. Burt, and G. V. D. Wal, "Change Detection and Tracking Using Pyramid Transformation Techniques," In Proc. of SPIE Intelligent Robics and Computer Vision, Vol. 579, pp. 72-78, 1985..
- [5] T. Brodsky and Y. T. Lin, "Linking tracked objects that undergo temporary occlusion," US. Pat., 2004.
- [6] S. C. Cheung and C. Kamath, "Robust Background Subtraction with Foreground Validation for Urban Traffic Video," EURASIP Journal on Applied Signal Processing, vol.14, pp. 2330-2340, 2005.
- [7] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian Detection of Independent Motion in Crowds," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 594-601, 2006.
- [8] B. Chen and Y. Lei, "Indoor and Outdoor People Detection and Shadow Suppression by Exploiting HSV Color Information," Proceedings of the Fourth International Conference on Computer and Information Technology, pp. 137-142, 2004.
- [9] A. Leone, C. Distanto and F. Buccolieri, "A Shadow Elimination Approach in Video-Surveillance Context," Pattern Recognition Letters, vol. 27, pp. 345-355, 2006.
- [10] M. Heikkila and M. Pietikainen, "A Texture-Based Method for Modeling the Background and Detecting Moving Objects," PAMI, vol. 28, pp. 657-662 2006.

## 出席國際學術會議心得報告

計畫編號	NSC 95-2218-E-216-004
計畫名稱	物件式前景物影像抽取技術
出國人員姓名 服務機關及職稱	黃雅軒 中華大學 資訊工程系 副教授
會議時間地點	2007/7/9~2007/7/12, 美國弗羅里達州的奧蘭多市
會議名稱	International Conference on Artificial Intelligence and Pattern Recognition)
發表論文題目	Local Structure Based Foreground Object Extraction

### 一、參加會議經過

此會議舉辦時間為從 2007/7/9 至 2007/7/12，由於 2007/6/25~2007/6/28 我在拉斯維加斯參加另一個會議，因此在洛杉磯停留大約十天後，再於 7/8 日前往奧蘭多市。除了聽取國際其他研究機構在圖形識別領域的研發成果外，我也在 7/10 的下午發表我們的論文「Local Structure Based Foreground Object Extraction」。

### 二、與會心得

AIPR (International Conference on Artificial Intelligence and Pattern Recognition) 為電腦視覺領域一個新的國際性學術會議，著重新穎性的技術和應用的發表，目的為展示此領域技術的發展趨勢。為了進行跨領域技術的分享和整合，AIPR 和其他 3 個會議於 7/9 至 7/12 在美國弗羅里達州的奧蘭多市(Orlando) 同時舉行，這 3 個會議分別為

1. International Conference on High Performance Computing, Networking and Communication Systems
2. International Conference on Automation, Robotics and Control Systems
3. International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics

第一屆 AIPR 論文投稿相當踴躍，論文接受率只有 30%，而我們的論文 Local Structure Based Foreground Object Extraction 很榮幸的被接受。本論文提出一種新穎性的前景影像抽取

方法，它將傳統的前景影像抽取與物件偵測等不同的處理模組進行巧妙的結合，可得到更正確與完整的前景影像抽取結果。首先，以傳統的前景影像抽取模組來對整張影像產生與背景影像模型具有足夠差異量的第一階段畫面式前景影像點，再對具有足夠前景影像點的區域進行判斷其是否存在某些特定類別的物件(如人形或車形等)；接著，針對被判斷為具有特定類別物件的區域進一步的抽取其第二階段的區域式前景影像點；最後，將前後兩次(即第一階段和第二階段)所抽取的前景影像點進行整合而得出最後的前景偵測輸出影像。此方法容易實施，具有高度實用性。

此次出國，除了在會議上發表論文外，也透過分享與討論，了解各國研究單位的技術和發展方向，對未來計畫的研發有很大的幫助。