

以支援向量機辨識不同病理型態之 超薄子宮頸抹片細胞影像

詹永寬	陳榮靜	黃宇睿	黃博琪
國立中興大學	朝陽科技大學	朝陽科技大學	行政院衛生署台中醫院
資管系副教授	資管系副教授	資管系研究生	病理科主任
ykchan@nchu.edu.tw	crching@cyut.edu.tw	s9414636@cyut.edu.tw	pchuanglin.tw@yahoo.com.tw

摘要

子宮頸抹片是公認一項簡單、便宜又極方便的篩檢方法，自從 1990 年台灣開始使用子宮頸抹片細胞篩檢來防範侵犯性子宮頸癌，已使得子宮頸癌的盛行率大大的降低。利用子宮頸抹片細胞篩選仍有一些盲點來自於採檢、篩檢以及判讀的誤差，以及人為篩檢的一些限制，例如主觀意識、疲勞、及時效性等，始終難以完全克服，綜合以上種種需求，設計出更好的科學方法來改善細胞篩檢品質，已成為近年來國際間爭相研究的主題之一。本研究採用醫院所製作的新柏氏電腦超薄抹片，配合醫院的病歷資料，利用高解析之數位相機及研究用顯微鏡挑選出正常及不同程度異常的子宮頸細胞型態，將其中有效細胞影像(超過 500 個影像)數位化，再配合醫師觀點從數位化影像處理以及電腦視覺中找出最佳的十三類特徵項目來分析子宮頸細胞，之後將各特徵項目之值交給支援向量機作訓練，利用訓練出來的類別模型，對細胞進行辨識分類。實驗證明，我們的方法可以獲得 94.43% 的正確率。

關鍵詞：子宮頸癌、新柏氏電腦超薄抹片、細胞影像分析、種子迴旋成長法、支援向量機。

1. 前言

子宮頸抹片是公認一項簡單、便宜又極方便的篩檢方法，自從 1992 年台灣開始使用子宮頸抹片細胞篩檢來防範侵犯性子宮頸癌，已使得子宮頸癌的盛行率大大的降低[5][14]。一般而言，傳統抹片檢查的敏感度約是 80-90% 左右，檢驗本身會有一定程度的偽陰性或偽陽性，我國衛生署針對檢驗也有相當多的品質監控及改善政策：包括採檢品質之專業訓練、檢驗單位之認證標準化、各項檢驗人員之培訓、

規定細胞醫檢師及醫師合理工作量、實驗室流程、建立全國檢驗子宮頸癌資料登錄之數位資訊化系統、各項流行病學研究、外派醫療及公共衛生人員於國外參訪、各項檢驗品質指標之統計，以及舉辦各種研討會及繼續再教育的學分課程等等，以提升我國子宮頸癌之細胞抹片篩檢能力及檢查單位之篩檢品質。

即便是如此嚴密把關，一些盲點來自於採檢、篩檢以及判讀的誤差，使得敏感度或精準度下降，例如採檢品質不良而造成的偽陰性可高達 2/3。還有人為篩檢的一些限制，例如主觀意識、疲勞、及時效性等，始終難以完全克服，綜合以上種種需求，設計出更好的科學方法來改善細胞篩檢品質，已成為近年來國際間爭相研究的主題之一[10]。

細胞質(cytoplasm)與細胞核(nucleus)之大小、形狀、顏色深淺與內部紋路(texture)和子宮頸細胞正常與否，有著密切不可分的關係。故要能夠正確地提供子宮頸細胞抹片篩檢結果，首要條件在於能否精確地從子宮頸細胞抹片影像中切割出細胞質與細胞核的輪廓。本研究的主要目的在於利用電腦來判斷細胞的病理特徵，如細胞的形狀、細胞核與細胞質之面積比例(核質比)、核膜之規則性、染色質之含量及分布、核仁明顯與否等，判斷是否為淺層鱗狀上皮細胞(Superficial squamous cell, S Cell)、中層鱗狀上皮細胞(Intermediate squamous cell, I Cell)、側基底鱗狀上皮細胞(Parabasal squamous cell, P Cell)或低(高)危險度鱗狀表皮內病灶(Low(High)grade squamous intraepithelial lesion cell, (L(H)SIL Cell)，以供醫師作為診療之參考。

要透過電腦來辨識子宮頸細胞影像時，首先需要區分細胞影像的背景、細胞質與細胞核的部分，才能再進一步分析，N. A. Mat-Isa1 等

人[12]利用設定門檻值(threshold)以及利用種子區域成長特徵的方法(seeded region growing features extraction, SRGFE)分別定義細胞核與細胞質的區域,之後利用該面積比值的大小來做簡單的細胞判定,但設定門檻值還是需要人類根據不同細胞而給予不同的數值。本國雖然在生物科技與醫學工程上各有發展,但是有關這方面技術卻是十分有限,相關的研究也很少[4][1][2][3],所使用的細胞影像是取自於傳統抹片[4],或是實驗樣品不足或是僅限於實驗室之初期開發概念尚未能經過醫學的評估[2][3],而[1]主要為細胞影像切割的研究。

由於目前電腦切割細胞的技術多半需要人工加以輔助設定門檻值或者圈選的範圍節點設定[12],以及電腦切割解析度高的細胞影像的所需的運算時間過長,且目前利用電腦切割細胞邊界時,容易受到如拍攝光線、色溫、解析度與細胞重疊、玻片內多餘液體、血絲過多與染色劑等不同因素與程度的干擾,故本研究將採取人工直接繪製細胞遮罩的方式來定位細胞在影像中精確的位置,將細胞從背景中切割分離出來,以求大量分析細胞影像的過程中不會因為切割的部分的誤差而分析錯誤的細胞位置,導致最後的病理判讀錯誤。因為即使採用新的薄層抹片技術下的異常細胞特性仍是成群結塊的黏在一起,故電腦切割細胞尚待更精良的技術或者方法來完成。

本研究目的在於利用電腦判斷病理的部分,著重找出能描述細胞病理的特徵項目讓電腦學習如何藉由特徵項目辨識細胞的病理型態。

透過初步文獻收集、醫師的判讀經驗與實驗結果,我們找出其中十三個能描述子宮頸細胞特徵科學計算特徵項目,分別為實際週長(μm)、實際面積(μm^2)、實際長軸 L (μm)、實際短軸 W (μm)、NCR(面積比)、重心與軸心分別到邊界最大值與平均值、灰階共伴矩陣(Co-occurrence Matrix)的熵值(Entropy)與對比值(Contrast)以及粗糙度(coarseness)、對比度(contrast),我們將這些細胞的特徵值取出後,透過支援向量機(Support Vector Machine, SVM)訓練後產生模型(model),則支援向量機即可以透過模型進行細胞的分類辨識。

剩餘的章節介紹如下,第二節介紹本研究的實驗架構與流程。第三節介紹子宮頸細胞分析系統如何利用細胞遮罩擷取細胞特徵,以及十三項細胞特徵項目介紹。在第四節介紹實驗結果與討論以及第五節結論。

2.研究架構與影像前處理

本研究架構主要分為七個階段進行,各階段內容如下:

- (1) 階段一:首先藉由已通過衛生署子宮頸抹片檢查單位審查之區域教學醫院之病理科利用電腦製作的新柏氏超薄抹片,配合高解析度數位相機及研究用顯微鏡,伴隨著病理醫師一同學習如何挑選出正常及不同程度異常型態的子宮頸細胞影像。圖 2-1 中(1)階段
- (2) 階段二:根據陪同病理醫師學習辨識細胞的經驗,調查匯整歷年數位影像處理領域中探討形狀與材質相關特徵值之論文,整理公式以及其物理意義與計算方法。
- (3) 階段三:設計一套子宮頸細胞分析系統,以自動化的方式分別取得細胞影像中,不同特徵項目的特徵參數,以便未來大量處理細胞影像時能有效節省人工介入處理的時間,此系統結合細胞、遮罩與尺規影像的載入、尺規長度運算、細胞座標定位、特徵項目運算及資料輸出的功能。圖 2-1 中(4)階段
- (4) 階段四:將細胞檔名利用病歷號碼與病理醫師的診斷結果,正規化為唯一檔名,目的在可以從細胞影像檔名直接辨識與瞭解該細胞的完整特性,且細胞檔名也不會重複。圖 2-1 中(1)階段
- (5) 階段五:利用影像繪圖軟體,將階段四的每一張細胞影像處理影像色彩並裁切粗略細胞影像區域後,做出同細胞大小的影像遮罩,並另存為不同格式但檔名相同的檔案。圖 2-1 中(2)階段
- (6) 階段六:利用階段三所設計出的子宮頸細胞整合分析系統,配合階段五的細胞影像與其遮罩影像檔案,分別將單一檔案中的單一細胞特徵值取出量化成數據。圖 2-1 中(3)階段
- (7) 階段七:將階段六的特徵值依照極值正規化,分訓練組與測試組後,經由支援向量機訓練成模型,並利用擁有較佳結果之模型做其他細胞的辨識實驗,以獲得實驗結果。圖 2-1 中(5)-(7)階段

下面的章節將分別介紹子宮頸細胞分析系統與支援向量機的實驗流程與方法。

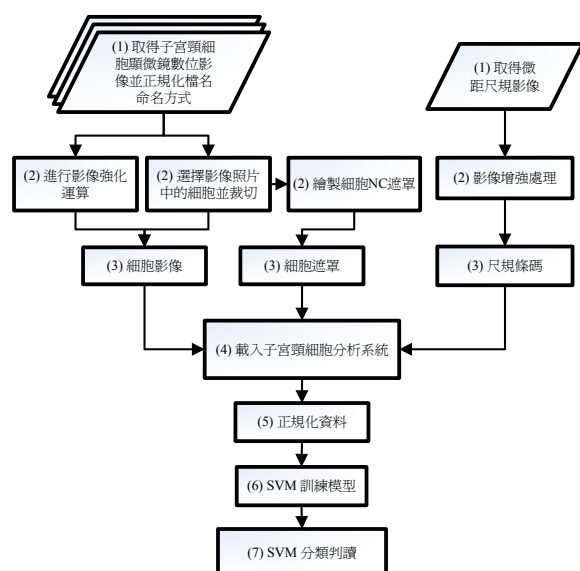


圖 2-1 研究架構流程圖

2.1 影像檔案名稱處理

為了避免大量的細胞影像造成我們管理資料的困難，所以我們設計一種檔名命名的規則，讓每一張細胞影像都有唯一且不重複的名稱，並可以從檔名中直接瞭解該細胞的病理、病歷等資訊，命名規則如表 2-1 所示。

表 2-1 子宮頸檔案命名規則

欄位	(1)	(2)	(3)	(4)	(5)
範例	S	1	4	CE206-236	1
欄位介紹	細胞型態	單顆細胞之流水號	放大倍率	病例號碼	診斷編號

- (1) 細胞型態：共有五種細胞型態代號，為 S、I、P、LSIL 與 HSIL 代表五類細胞，分別如 Superficial squamous cell、Intermediate squamous cell、parabasal squamous cell 與 Low (High) grade squamous intraepithelial lesion。
- (2) 單顆細胞流水號：當拍攝的焦點都為同一顆細胞時，該流水號不會增加，單一流水號可能會有多种放大倍率。
- (3) 放大倍率：我們在拍攝子宮頸細胞的時候，會針對單一顆細胞以 100X、200X 與 400X 的倍率拍攝同一顆細胞，除非第二

顆細胞緊鄰著前一顆細胞，則可能會省略拍攝第二顆細胞的 100X 照片，甚至同時省略 100X 與 200X 的照片，因為我們的研究過程中，用來分析的照片都是 400X 的倍率，低越低倍率的照片，所包含的細胞數量越多，而 400X 的細胞照片通常只有單一顆主題細胞，故低倍率的拍攝目的是讓病理醫師或細胞醫檢師能輕易比對主題細胞與相對其他細胞的大小，進而檢查細胞影像病理判斷的正確性。

- (4) 病歷號碼：行政院衛生署立台中醫院所給予每一位病人之病歷號碼。

- (5) 診斷編號：為衛生署根據 TBS system 統一標準分類原則，請參考附錄表(a)。

2.2 影像前處理及建立細胞遮罩

首先我們先對細胞影像做直方圖正規化 (histogram normalization) 與直方圖均等化 (histogram equalization) 影像增強處理，可以強化並還原成當初在顯微鏡中看到的影像品質，圖 2-2 (a) 為原始拍攝的影像，該背景與細胞由於色溫與光線的影響，所以導致背景與主題之間有點模糊難以分辨，但經過影像強化處理後的影像如圖 2-2 (b) 的影像，背景與細胞已經可以很清晰的被分開來，如同我們當時在顯微鏡下用肉眼看到的結果，可以增加我們繪製細胞遮罩影像的效率。

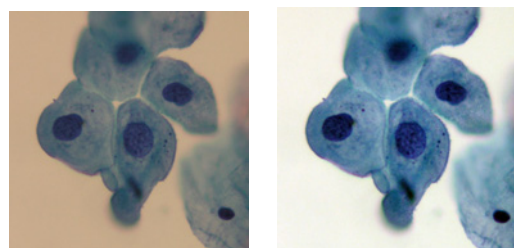


圖 2-2 細胞影像前處理

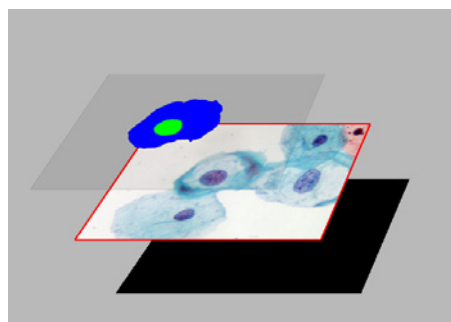


圖 2-3 細胞遮罩與細胞影像關係

我們分別預設將細胞質遮罩設定為藍色

(RGB:0,0,255)，細胞核部分設定為綠色(RGB:0,255,0)如圖 2-3，並將細胞遮罩影像另外存檔為 GIF(Graphics Interchange Format)檔案格式的圖形檔。

2.3 測微尺玻片影像

測微尺玻片(圖 2-4)是用來輔助衡量該細胞在實體世界中的大小，本實驗加入尺規來計算實際細胞大小，有助於醫生可直接利用判讀經驗直接判斷數據是否準確與符合邏輯，讓實驗初步即可讓醫師根據判讀經驗校對，使得爾後分析所取出的數值對於細胞的分辨更增加權威的證據與說服力。影像中利用尺歸來做相對尺度大小運算是一項非常重要的參考數據，也是電腦辨識細胞病理的最重要的依據。我們參考醫師判讀可疑細胞的技巧，當醫師操作顯微鏡進行閱片時，若發現可疑的細胞的時候，可以自行移動抹片，尋找抹片中其他的正常細胞(如發炎細胞)來做相對大小比較，故醫生可以輕易的在抹片中判斷病理，但電腦中只有相對比例的細胞與數據而已，若不利用相同焦距環境所拍攝的尺規，將細胞大小的特徵正規化成絕對長度，則電腦無法分辨特徵相似，但大小有天壤之別的細胞。

我們的實驗將測微尺玻片的影像直接處理為二值化影像，使子宮頸細胞分析系統能利用類似掃描條碼的過程記錄刻度間所代表的像素大小。(圖 2-1 (3)階段)

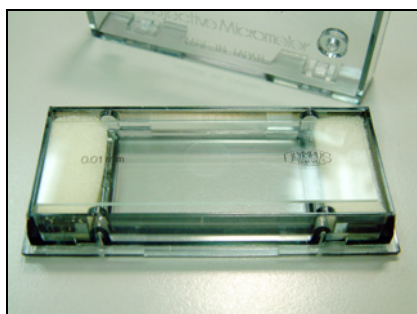


圖 2-4 測微尺玻片

2.4 子宮頸細胞之特徵項目分析

此步驟將處理之後的細胞影像、細胞遮罩影像以及尺規二值化影像輸入到子宮頸細胞分析系統中處理，子宮頸細胞分析系統將在第三節中詳細介紹。

2.5 特徵值正規化

每一個特徵項目的特徵值都有自己所擁有的單位，例如：長度、面積等，所以特徵項目之間的值域差距較大，故我們必須將每個特徵項目中的特徵值正規化，讓全部特徵項的

值域都相同，避免 SVM 訓練時，SVM 誤把將權重放在值域較突兀的特徵項目上。

我們正規化每組特徵項目中的特徵值，在每一組特徵項目中，各別找出特徵值的極值，並利用極值來做特徵數的正規化。正規化的公式為：

$$n' = \frac{X(n) - \text{Min}}{\text{Max} - \text{Min}}$$

其中 X 為第 n 個參數值， n' 為正規化之後的結果，範圍在 $[0,1]$ 。

2.6 支援向量機 (Support Vector Machines, SVM)

支援向量機(Support Vector Machine, SVM)[6][17]是建立在機器學習理論的結構風險最小化原則之上，其主要思想是針對二元分類問題，在高維度空間中尋找一個超平面作為二類的分割，以保證最小的分類錯誤率，而且 SVM 一個重要的優點就是能處理線性不可分的情況。

支援向量機是近代統計學習理論(statistical learning theory)一個新的應用，它是屬於監督式學習法(supervised learning)，監督式學習法代表的是，在分類時的輸入與輸出值都需事先知道，例如我們事先賦予不同類別細胞特定參數，而 SVM 輸出階段會將分類結果利用我們所賦予的參數，註記原始資料被分類的結果。

支援向量機近幾年來被廣泛的被應用於網頁分類、影像辨識、生物科技等分類議題上，都有傑出的成就。支援向量機的主要議題有核心函數(Kernel Function)及超平面(Hyperplane)，這兩個議題將是影響分類優劣的主要因素。

支援向量機最初是設計於處理兩種不同類別(binary classification)的資料集[13]，但由於在實際的情況下，分類並非僅僅只有兩類，也可以是多類[11][19]，因此支援向量機可分為 C-SVC、NU-SVC、OC-SVM、Epsilon-SVR、NU-SVR[15]...等。

3. 子宮頸細胞分析系統

為了能大量讓電腦批次處理子宮頸細胞我們設計了一套能自動細胞定位、特徵項目運算與資料整合輸出的系統，該系統能根據細胞遮罩找到相對座標的細胞點，並針對細胞的形狀與紋理進行分析。

3.1 種子迴旋成長法 (Seeded Circle Region Growing)

為了讓系統能自動根據細胞遮罩影像來讀取細胞的輪廓，並將數據建立成座標點矩陣，我們提出了『種子迴旋成長法 (Seeded Circle Region Growing)』，該演算法主要作法為在種子點周圍 3x3 的範圍，不斷以順時針的方向環繞，且在該範圍像素點中，挑選符合條件的一個點作為成長點(Growing Point)，每旋轉一圈只挑選一個點為成長點，並將該成長點設定為下一個新的種子點，並持續尋找，依序將整個形狀輪廓點座標紀錄下來，直到回到最初的第一點為止，或者新的成長點當兩點距離為零時停止，而種子點周圍的點位 (1 - 8) 將成為尋找下一點的範圍。種子點為中間上色的的位置，與他周圍點的順序位置如圖 3-1。

首先，我們必須找到第一顆起始種子點，方法為根據細胞遮罩影像由左到右，由上到下找尋遮罩中，找到細胞質或細胞核的定義顏色相同的即屬於第一點起始點，也稱為種子點 (Seed Point)。

1	2	3
8		4
7	6	5

圖 3-1 點位順序介紹

原始輸入到程式中的遮罩如圖 3-2 (a)，每一個格子代表一個像素(Pixel)。以下，將以圖 3-2 介紹種子迴旋成長法的後續運作方式。

圖 3-2 中的格子共使用四種灰階程度不同的顏色來表示不同狀態，其中白色部分代表空白背景，其他三種顏色為淺灰色、灰色與黑色，分別代表細胞遮罩影像、曾經找到的輪廓點與種子點位置，其中黑色的種子點，我們會在該點周圍格子中標示點位，便於我們對於圖例的解說。

當第一點種子點被找到的時如圖 3-2(b)標示灰色的位置，以該點當成起始點，由該點開始利用種子迴旋成長法找下一點。

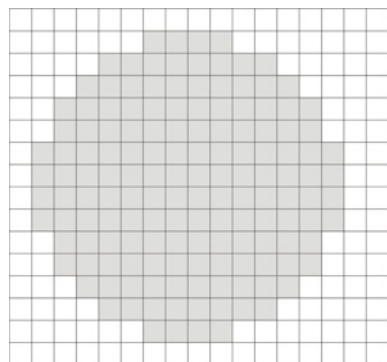


圖 3-2 (a)

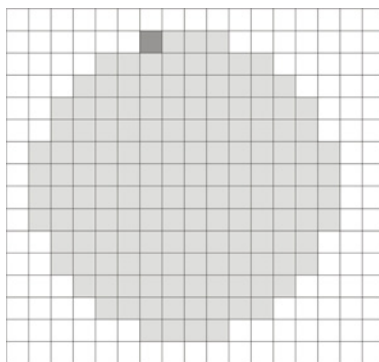


圖 3-2 (b)

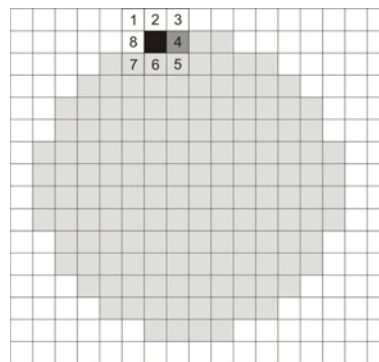


圖 3-2 (c)

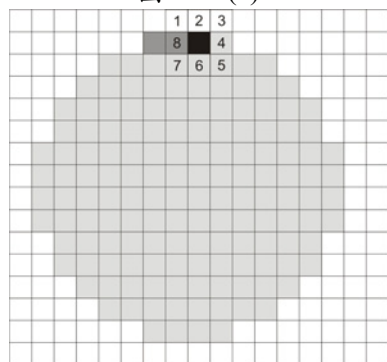


圖 3-2 (d)

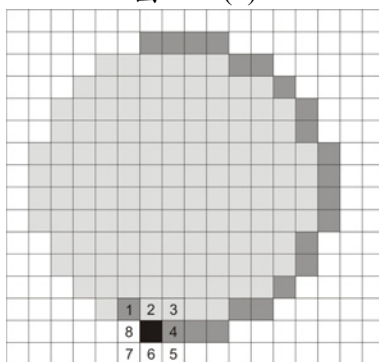


圖 3-2 (e)

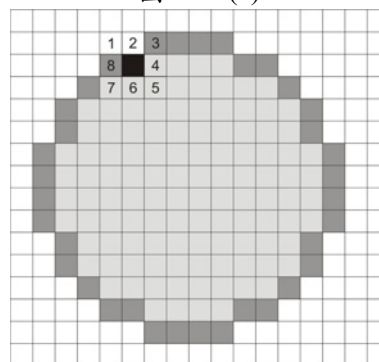


圖 3-2 (f)

圖 3-2 種子迴旋成長法

再來要選擇啟始尋找的位置由於尋找第一點的方向為水平方向尋找，故圖 3-2 (c)水平 8 的位置一定不是新點的位置，故找尋下一點的時候，需從位置 1 開始，由此可知，每一個下一種子點的啟始尋找位置，都為之前所找到種子點的對稱點加 1 的位置，即 $8+1 \rightarrow 1$ (大於範圍的點位就歸回 1 的位置)。

例如圖 3-2 (c)圖中，順著種子點周圍 1,2,3 的位置尋找，最後會找到 4 的位置為下一點新的種子點，而當種子點要繼位到新的種子點如圖 3-2 (d)圖中時，新的種子點的啟始尋找的位置，將是舊的種子點發現新種子點位置 4 的對稱位置 8，根據上述的法則，真正的啟始尋找點為 1 的位置，這樣可以避免一開始就找到舊的種子點而停在迴圈之中。

再以圖 3-2 (e)為例，舊的種子點之對稱點為 4，所以從 5 的位置開始尋找直到 1 的位置才找到新的種子點。

圖 3-2 (e)表示最後找到的點跟第一點吻合，故停止。

3.2 利用遮罩相對應座標的抓取細胞影像

我們同時由左到右，由上到下掃描細胞影像與遮罩影像，掃描時若在影像座標上發現遮罩影像上的點符合細胞質或者細胞核的預設值，則分別將相同座標的細胞影像像素值取出並存放於該點所屬的陣列中，細胞核與細胞質分開存放於兩個陣列。

3.3 細胞特徵項目分析

子宮頸癌細胞在病理學上分為多種不同型態，所表現的特徵也不竟相同，因此我們經過初步實驗後，挑出下面十三個能描述子宮頸細胞的特徵項目對每一顆細胞作分析，每一顆細胞均分細胞質與細胞核兩區域進行獨立分析，分析後的特徵值將匯集成資料表，一顆細胞將同時擁有細胞質與細胞核的十三個特徵值，共二十六個特徵項目。

經過初步的統計評估，本研究之初步實驗只採用細胞核部分的特徵項目來做 SVM 實驗，因為細胞質之紋理特徵的參考價值不高，且易受到其他細胞重疊、黏膜覆蓋等其他因素的干擾，但辨識細胞需要細胞質的面積來計算徵項目之一核質比(NCR)，故我們只利用細胞質之其中一項面積特徵來計算。

各特徵項目如表 3-1 中所示：

表 3-1 十三個特徵項目

標號	特徵項目	標號	特徵項目
1	實際週長(μm)	8	重心到邊界最大值
2	實際面積(μm^2)	9	重心到邊界平均值
3	實際長軸 L (μm)	10	灰階伴隨矩陣熵值
4	實際短軸 W (μm)	11	灰階伴隨矩陣對比值
5	NCR(面積比)	12	粗糙度
6	軸心到邊界最大值	13	對比度
7	軸心到邊界平均值		

3.3.1 形狀特徵項目[8]

形狀特徵是直接可以度量的一種特徵，也是最接近視覺所觀察到的一種描述細胞的重要方法，我們整理各特徵如下：

1. 實際週長(μm)：圍繞著細胞核或者細胞質一圈的細胞輪廓。我們分別獨立取得細胞核與細胞質之輪廓長度值做運算。
2. 實際面積(μm^2)：細胞遮罩所覆蓋的細胞範圍，單一顆細胞將分別獲取細胞質與細胞核的面積兩項獨立數據。
3. 實際長軸 L (μm)：利用全部的輪廓座標點，兩兩點互相利用歐幾理德距離公式計算兩點之間的距離，並不斷更新與紀錄運算過程中的極值項目，將距離最長者設定為為長軸 (L)。

$$L = \sum_{\Delta X} \sum_{\Delta Y} \text{Maximum}(\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2})$$

4. 實際短軸 W (μm)：以上述所建構之長軸之兩點端點為分界，分為左半邊與右半邊輪廓，左右兩邊輪廓點之間所構成的直線中，既垂直長軸且長度最長者，設定為短軸(W)。我們利用向量內積的公式做角度的運算，當左邊一點與右邊的兩點帶入向量公式後，角度為 90° 者，將其座標點與歐幾理德距離記錄起來，並在運算過程中記錄極值，找出最長的設定為為短軸 (W)。

$$\text{向量夾角公式：} \cos \theta = \frac{a \cdot b}{|a||b|}$$

5. NCR(面積比):細胞核面積除以細胞質面積的比值,當核質比越大,則越有可能為異常細胞,特徵項目

3.3.2 重心與軸心運算

本研究採用重心(Gravity center)與軸心(Axle center)的運算來瞭解細胞形狀的特徵,其中軸心(Axle center)為長軸L與短軸W兩垂直軸之交點,由於重心的位置位於所有的週長點的中心位置,所以重心到週長的距離可以一一的顯示出細胞形狀的變化,而當細胞的形狀為對稱的形狀時,軸心會交於重心之上,例如橢圓型、圓形...等,當細胞的形狀為不規則與不對稱時,軸心會與重心相錯開一段距離如圖3-3所示。

我們分別計算從重心與軸心到達細胞週長中每一點的距離,並計算這些特徵值的平均距離值與最大距離值,距離平均值可以與最大距離值互相比較可以瞭解細胞是否有特異的變形,由於重心到週長的距離變化一般來說都會比較平滑,因為該點的位置是位於所有週長點的正中央的位置,反觀軸心點來看,該點取決於短軸的落點位置,故該點到週長的距離長度變化會比利用重心所計算出來的值還要大。這部分的計算,分別為項目6到項目9。

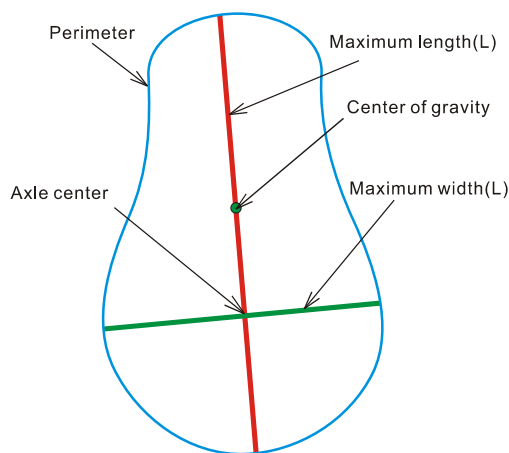


圖 3-3 各形狀特徵項目示意圖

重心(Gra):將所座標點的方量點座標 x, y 分別加總,再各別除以全部有的週長的座標點即可得到形狀的重心點。

$$\text{重心公式: } Gra = \left(\sum_{xi} \frac{x}{N}, \sum_y \frac{y}{N} \right)$$

3.3.3 紋理特徵項目

紋理分析為一種描述區域的重要方法,我們利用以下性質來分別分析細胞質與細胞核

之紋理特徵,用來探討細胞影像細緻與粗糙程度、對比度及規則性。

3.3.3.1 灰階伴隨矩陣計算

一個灰階伴隨矩陣(Co-occurrence matrix)矩陣[18][7]是一個二維直方圖用來估計一對灰階程度的統計資料,我們使用灰階伴隨矩陣來分別檢查細胞影像中的紋理。計算抹片影像的揮度伴隨矩陣中每一個元素值,如果影像灰度伴隨矩陣之熵值很高,則影像很均勻,意味著類似的灰度伴隨矩陣形狀安排,沒有特殊的紋理。同樣的,也可求出影像灰度伴隨矩陣之能量值、對比值與均勻度,我們分別選擇對於細胞鑑別能力較佳的熵值(Entropy)與對比值(Contrast),做為材質因素之一,其公式如下:

10. 熵值公式:

$$\text{Entropy} = - \sum_i \sum_j P[i, j] \log P[i, j]$$

11. 對比值公式:

$$\text{Contrast} = \sum_i \sum_j (i - j)^2 P[i, j]$$

3.3.3.2 Tamura 紋理特徵

基於人類對紋理的視覺感知的心理學的研究中,Tamura等人提出了紋理特徵的表達方式[9]。Tamura紋理特徵的六個分量對應於心理學角度上紋理特徵的六種屬性,分別是粗糙度(coarseness)、對比度(contrast)、方向度(directionality)、線像度(linelikeness)、規整度(regularity)和粗略度(roughness)。我們利用其中兩種分量做特徵項目,分別是粗糙度(coarseness)與對比度(contrast)。

12. 粗糙度(coarseness) 粗糙度的計算可以分為以下幾個步驟進行。首先,計算圖像中大小為 $2^k \times 2^k$ 個像素的活動視窗中像素的平均強度值,即有

$$A_k(x, y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} \frac{g(i, j)}{2^{2k}}$$

其中 $k=0, 1, \dots, 3$, 在這裡,我們為了增加運算速度,所以將 K 的最大值改成 3 (文獻值為 5) 而 $g(i, j)$ 是位於 (i, j) 的像素強度值。然後,對於每個像素,分別計算它在水準和垂直方向上互不重疊的視窗之間的平均強度差。

$$E_{k,h}(x,y) = \left| A_k(x+2^{k-1},y) - A_k(x-2^{k-1},y) \right|$$

$$E_{k,v}(x,y) = \left| A_k(x,y+2^{k-1}) - A_k(x,y-2^{k-1}) \right|$$

其中對於每個像素，能使 E 值達到最大（任何方向）的 k 值用來設置最佳尺寸 $S_{best}(x,y) = 2^k$ 。最後，粗糙度可以通過計算整幅圖像中 S_{best} 的平均值來得到，表示為

$$F_{crs} = \frac{1}{m \times n} \sum_{i=1}^m \sum_{j=1}^n S_{Best}(i,j)$$

粗糙度特徵的另一種該進形式是採用直方圖來描述 S_{best} 的分佈，而不是像上述方法一樣簡單地計算 S_{best} 的平均值。這種改進後的粗糙度特徵能夠表達具有多種不同紋理特徵的圖像或區域，因此對圖像檢索更為有利。

13. 對比度(contrast)。對比度是通過對像素強度分佈情況的統計得到的。確切地說，

它是通過 $\alpha_4 = \frac{\mu_4}{\sigma^4}$ 來定義的，其中 μ_4

是四次矩而 σ^2 是方差。對比度是通過如下公式衡量的：

$$f_{con} = \frac{\sigma}{\left(\frac{\mu_4}{\sigma^4} \right)^{\frac{1}{4}}}$$

該值給出了整個圖像或區域中對比度的全局度量。

4. 實驗結果與討論

本研究是採用 LIBSVM 軟體訓練子宮頸細胞特徵值，並利用該軟體所產生的模型辨識細胞類型，此套軟體除了可以處理單類別支援向量機之外，還可以處理多類別的支援向量機，此套軟體可在 www.csie.ntu.edu.tw/~cjlin/libsvm 網址中取得。

在訓練(training)的階段中，訓練會接受特定格式的輸入值，此處特定的格式表示法為：

[label] [index1]:[value1] [index2]:[value2] ...

一行一筆資料如下所示

+1 1:0.708 2:0.987...

- (1) label : 或說是 class，要分類的種類，通常為一些整數。

- (2) index : 是有順序的索引，通常是放

連續的整數，也是特徵值的編號順序。

- (3) value : 就是用來 train 的資料，通常是一群實數。

我們把不同類型的細胞以不同整數作為分類，標籤的設定數值由使用者自行決定，我們以間隔方式設定標籤，其目的為方便閱讀結果，也間接減少人為誤判錯誤的機會，本研究共分為四類，正常細胞分類為 1、3 與 5 三類，異常細胞單獨分類為一類為 -4，如下表所示：

表 4-1 SVM 標籤值設定

label	細胞種類
1	Superficial squamous cell, S Cell
3	Intermediate squamous cell, I Cell
5	parabasal squamous cell, P Cell
-4	Low(high) grade squamous intraepithelial lesion, L(H)SIL Cell

完成訓練後即產生一個 Model 檔，這個檔案我們可以想成是支援向量機的內部分類資料，而在訓練時我們利用參數的改變來達到最佳的辨識結果。

4.1 細胞影像資料庫

- (1) 我們採用新柏氏電腦超薄子宮頸抹片 (Thin-Prep) 細胞，分別拍攝 100X、200X 與 400X 不同倍率的影像檔。
- (2) 電腦辨識的時候，我們統一採用 400X 的影像作為判斷的主要依據，我們共取得的有效細胞數量四種如表 4-2 所示：

表 4-2 細胞有效顆數數量

正常細胞	病例	有效影像數
S Cell	13 個	139 個
I Cell	10 個	178 個
P Cell	10 個	128 個
異常細胞		
L(H)SIL	25 個	58 個

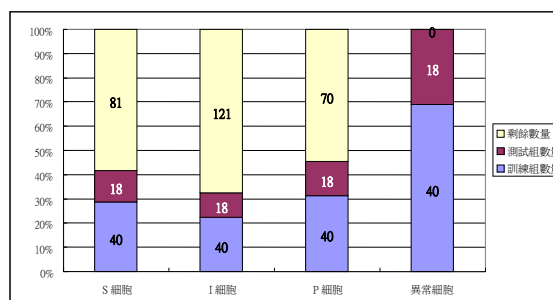


圖 4-1 細胞的數量以及分組使用的情形

實驗分成訓練集、測試集與全部的資料三部分，訓練集的內容是取全部四種的前 40 個細胞，共 160 個細胞資料作訓練的檔案，測試集的部分是取四種細胞的第 40 個以後的 18 個細胞資料，作為測試集，共 72 個。

測試集與訓練集內細胞資料是獨立存在的，也就是說測試與訓練階段，分別採用不同的兩組細胞。而全部的資料則是將全部的有效細胞做整體的測試，共 503 個細胞資料，此資料集的內容包含了前面的訓練集與測試集以及一些全新且尚未測試的剩餘資料集。藉此來做大規模的判斷我們所訓練出來的模型(Model)是否能正確的判斷細胞，圖 4-1 表示細胞的數量以及被使用的情形。

4.2 各項細胞參數的探討

我們在此節討論部分特徵項目的分類結果統計圖。我們也利用此種統計圖的分析方式，將實驗後不具代表性的特徵項目給淘汰，故我們最後萃取僅十三個特徵項目做 SVM 研究。

每一張統計圖中，左邊數列代表正規化後的值域，範圍為[0,1]，底部的數列為細胞個數，我們探討的是每個特徵項目對於細胞整體的表現。

圖 4-2 為四類細胞的實際週長的數值分析結果，由統計圖表中可以看出該特徵項目可以辨別 L(H)SIL 細胞、S 細胞、I 與 P 細胞三群，由圖中可知到 L(H)SIL 細胞之細胞核實際週長對於整體細胞來說屬於最大的一群，S 細胞則是最小的一群，I 與 P 細胞的細胞核大小差不多，故不能單從圖 4-2 中直接分辨。

NCR (面積比) 的分析結果於圖 4-3 中，是醫生最長用來判斷細胞好壞的重要依據，由圖中我們可以看到 P 細胞可以完整的跟其他正常細胞做分離，即使它在 NCR (面積比) 中與 L(H)SIL 細胞略微重疊，但 L(H)SIL 細胞可以透過其他特徵項目的輔助被分離。如圖 4-4 的特徵項目，L(H)SIL 細胞所屬的特殊紋理能較清楚的該被分離出來。

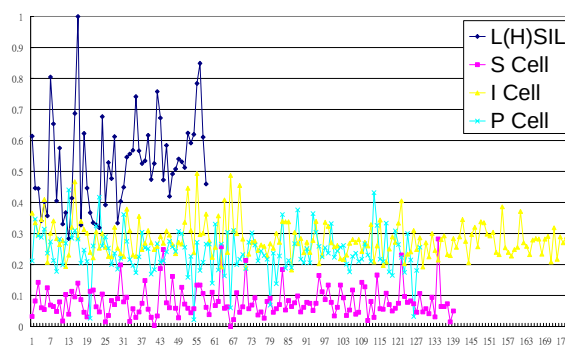


圖 4-2 實際週長(μm)

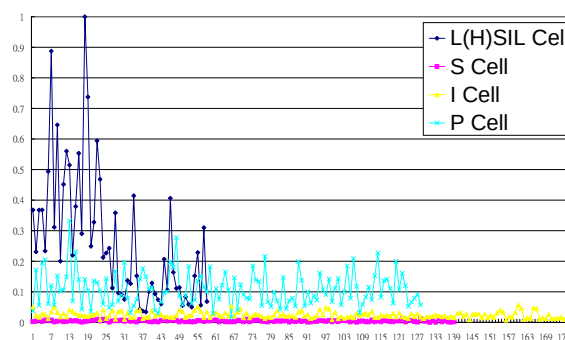


圖 4-3 NCR (面積比)

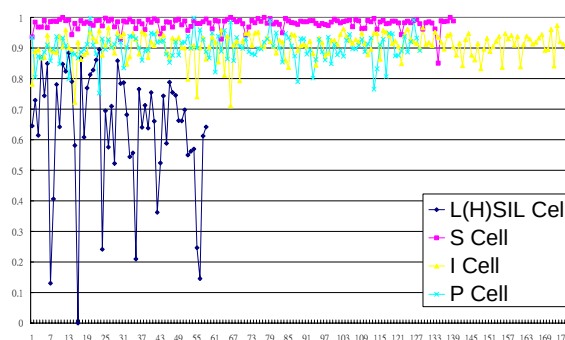


圖 4-4 Co-occurrence Matrix Entropy

Coarseness 特徵項目於圖 4-5 中，似乎可以清晰的將四種類型細胞做區分，不過由圖中一些異常值的出現可以瞭解，若不配合其他的特徵項目的輔助，還是會存在有很大的誤判機會。

Contrast 特徵項目於圖 4-6 中，由於 L(H)SIL 細胞佔大多數的數值範圍空間，其他類型的細胞似乎都重疊在一起，故我們將範圍縮小到 0-0.1 之間來觀察如圖 4-7 所示，由較小的尺度範圍我們可以發現，不同種類的細胞仍能被清楚的鑑別。

其他幾項特徵項目如”實際面積(μm^2)”、“實際長軸 L(μm)”、“實際短軸 W(μm)”、“

軸心到邊界最大值與平均值”、“重心到邊界最大值與平均值”等，所呈現出來的結果跟圖 4-2 有點相似，故不在此討論。

但不同類型細胞所產生的範圍分佈即使肉眼看似相似，但多個特徵參數的結合運算時，也相當重要，因為單一顆細胞是以十三個特徵項目同時描述，故各特徵參數之間有如互助的性質，能間接將各類細胞分辨出來。

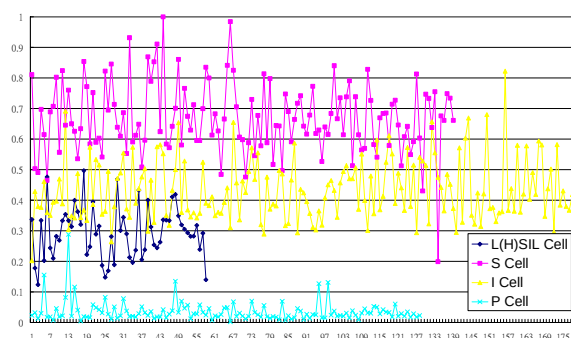


圖 4-5 Coarseness

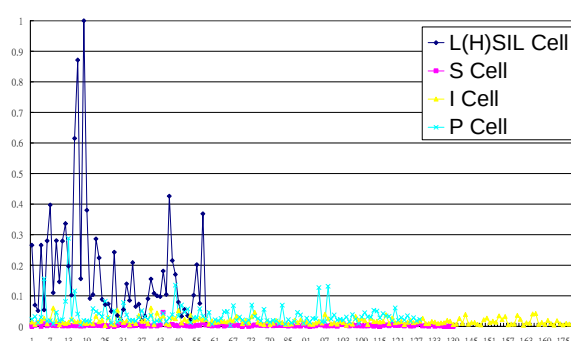


圖 4-6 Contrast

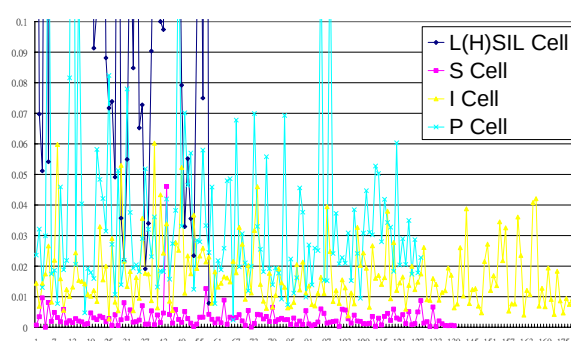


圖 4-7 Contrast 以 0 - 0.1 之範圍觀察

4.3 實驗結果

我們訓練模型的過程方式是利用訓練組的資料訓練出模型(Model)後，再利用相同的訓練組資料集模擬測試，目的是要方便我們對於 C (Cost) 參數的調整，成本(cost)參數由使用者自行決定，當愈大時表示錯誤發生時對目標函數的影響愈大，反之則影響愈小。

$$Accuracy = \frac{\text{辨識後正確的細胞數量}}{\text{全部的細胞數量}} \times 100\%$$

訓練其結果如以下所示。

表 4-3 訓練集所辨識的結果

訓練組資料集 (每組 40 個, 共 160 個)
Accuracy = 96.25% (154/160) (classification)
Mean squared error = 1.39375 (regression)
Squared correlation coefficient = 0.88314 (regression)

表 4-4 測試集所辨識的結果

測試組資料集 (每組 18 個, 共 72 個)
Accuracy = 95.8333% (69/72) (classification)
Mean squared error = 0.791667 (regression)
Squared correlation coefficient = 0.93253 (regression)

我們另外利用上述所測試的模型(Model)來做全部的細胞資料的辨識，以確定我們的模型(Model)是否能正確的辨識細胞的種類。全部的細胞資料共 503，我們一起測試之後，分開探討各種細胞被誤判的情形結果如表 4-5 所示。

表 4-5 全部細胞資料所辨識的結果

細胞種類	正確率 (Accuracy) (%)	誤判情形
S 細胞	96.4029% (134/139)	4 個誤判 I 細胞 1 個誤判為 P 細胞 全部數量：139
I 細胞	91.573% (163/178)	5 個誤判為癌症細胞 4 個誤判為 S 細胞 6 個誤判為 P 細胞 全部數量：178
P 細胞	93.75% (120/128)	2 個誤判為癌症細胞 3 個誤判為 S 細胞 1 個誤判為 I 細胞 全部數量：128
L(H)SIL 細胞	100% (58/58)	0 個誤判 全部數量：58

5. 結論

本研究配合醫師判別細胞的觀點，找出有代表性的特徵項目，並經過初步實驗後，找出對細胞辨識有主要有影響的十三個細胞特徵項目，並將細胞核的特徵項目運用支援向量機訓練模型，進行細胞辨識與分類。經由實驗證明支援向量機對細胞的分析與辨識能力頗高，正確率(accuracy) 為 94.43%。我們研究使用改良後的薄層抹片後發現可使得核重疊，固定不良的問題減少。在影像處理上，細胞與背景的對比強烈，且細胞邊緣明顯，在處理遮罩的部分比起傳統抹片能更容易瞭解細胞邊界所在範圍，但癌症細胞的特性都是成群結隊的互黏在一起，這種現象在薄層抹片技術仍無法改善，故重疊的癌症細胞仍須由醫師等人員依照經驗來分辨單顆癌細胞所屬範圍，而本研究方法可以適用於人工或者電腦自動切割細胞範圍後的辨識工作。在未來的研究方面將增加異常細胞影像的數量，讓實驗的結果能更具說服力之外，才能有足夠的資料將異常細胞分開為 LSIL 與 HSIL 兩組測試。我們也將試著找尋新的且鑑別力佳的特徵項目，讓細胞辨識的準確度能更加提升。

參考文獻

- [1] 陳淑嬌(2004)，“梯度向量流主動輪廓模型於子宮頸抹片細胞邊界偵測之應用”，中興大學電機工程研究所碩士論文
- [2] 馮齡儀(2005)，“電腦輔助子宮頸抹片異常細胞辨識之初期研究”，中原大學醫學工程學系研究所碩士論文
- [3] 黃俊魁(2006)，子宮頸抹片細胞之電腦輔助診斷系統”，中原大學/醫學工程研究所碩士論文
- [4] 黃博琪(2004)，“子宮頸細胞自動化電腦輔助影像分析系統之設計研發”，中興大學生物醫學研究所碩士論文
- [5] Cancer Registry Annual Report(1998), Republic of China,. Executive, Yuan, ROC, Department of Health
- [6] Corinna C., and Vladimir V. (1995), “Support-vector networks,” **Machine Learning**, Vol. 20, pp. 273-297.
- [7] E. L. Chen , P. C. Chung , C. L. Chen , H. M. Tsai , C. I. Chang , (1998), “An automatic diagnostic system for CT liver image classification,” **IEEE 83 Transactions on Biomedical Engineering**, Vol.456 , p 783 - 794.
- [8] H. C. Lin, C. Y. Chiu, S. N. Yang(2003), “Finding Textures by Textual Descriptions, Visual Examples, and Relevance Feedbacks,” **Pattern Recognition Letters**, Vol. 24, pp. 2255-2267.
- [9] H. Tamura, S. Mori, and T. Yamawaki(1978), “Texture Features Corresponding to Visual Perception,” **IEEE Transactions on Systems, Man, and Cybernetics**, Vol. Smc-8, No. 6, pp.460-473.
- [10] Hartmann KE, Nanda K, Hall S, Myers E(2001), “Technologic Advances for Evaluation of Cervical Cytology, Is Newer Better?”, **Obstetrical & Gynecological Survey**, 56(12), 765-774.
- [11] L.Bottou, C.Cortes, J. Denker, H.Drucker, I.Guyon, L.Jackel, Y.LeCun, U.Muller,E. Sackinger, P.Simard, and V. Vapnik(1994), “Comparison of classifier methods,a case study in handwriter digitl recognition,” **In International Conference Pattern Recognition, IEEE Computer Society Press**, pp.77-78.
- [12] N. A. Mat-Isa, M. Y. Mashor, and N. H. Othman(2005), “Seeded Region Growing Features Extraction Algorithm,” Its Potential Use in Improving Screening for Cervical Cancer”, **International Journal of The Computer, the Internet and Management**, Vol.13#1 January – April, pp 61 -70.
- [13] N. Cristianini, J. Shawf-Taylor.(2000) “An Introduction to Support Vector Machines and other kernel-based learning methods,” **Cambridge University Press**.
- [14] Paley PJ(2001), “Screening for the major malignancies affecting women, Current guidelines,” **American Journal of Obstetrics & Gynecology**. , 184(5), 1021-1030.
- [15] R.-E. Fan, P.-H. Chen, and C.-J. Lin.(2005), “Working set selection using second order information for training SVM,” **Journal of Machine Learning Research** 6, 1889-1918.
- [16] S.Yonekawa, N. Sakai, O. Kitani(1996), “Identification of Idealized Leaf Types Using Simple Dimensionless Shape Factors by Image Analysis”, **Transactions of the ASAE**, Vol. 39, No. 4, pp. 1525-1533.
- [17] Steve G. (1998), “Support vector machines for classification and regression”, **ISIS technical report**, Image speech and

intelligent systems group of University of Southampton.

- [18] T. Hermes, C. Klauck, J. KreyW, J. Zhang(1995), "Image Retrieval For Information Systems, Proceedings of SPIE", **Storage and Retrieval for Image and Video Databases III**, Vol. 2420, p. 394-405.
- [19] Vapnik, V. and Chervonenkis(1964), A., "A Note on one Class of Perceptrons," **Automation and Remote Control**, vol.25.

附錄 A

附錄表(a) 衛生署根據 TBS system 統一標準分類原則

編號	報告(英文)		報告(中文)
1	Within normal limit		正常
2	Reactive changes: inflammation, repair, radiation, and others		良性反應-發炎、修補或電療
3	Atrophy with inflammation		荷爾蒙缺乏萎縮
4	Atypical squamous cells (ASC-US)		意義不明的非典型鱗狀上皮細胞
5	Atypical glandular cells		非典型腺體上皮細胞
6	LSIL	Mild dysplasia (CIN I) with koilocytes	輕度細胞變性併空洞細胞
7		Mild dysplasia (CIN I) without koilocytes	輕度細胞變性,無空洞細胞
8	HSIL	Moderate dysplasia	中度細胞變性
9		Severe dysplasia	重度細胞變性
10		Carcinoma in situ	原位癌
11	Squamous cell carcinoma		鱗狀上皮癌
12	Adenocarcinoma		腺癌
13	Other malignant neoplasm_____		其他惡性腫瘤
14	Other_____		其他
15	Atypical glandular cells favor neoplasm		非典型腺體上皮細胞，疑似腫瘤
16	Atypical squamous cells cannot exclude HSIL		非典型鱗狀上皮細胞，無法排除 HSIL