

國立中央大學

資訊工程研究所

碩士論文

以相機取像之中文文件辨識前處理系統

Camera based Preprocessing System for Chinese
Document Image Recognition

研究生：黃自達

指導教授：范國清 博士

溫敏淦 博士

中華民國九十六年七月



國立中央大學圖書館

碩博士論文電子檔授權書

(95 年 7 月最新修正版)

本授權書所授權之論文全文電子檔(不包含紙本、詳備註 1 說明)，為本人於國立中央大學，撰寫之碩/博士學位論文。(以下請擇一勾選)

(☒) 同意 (立即開放)

(☐) 同意 (一年後開放)，原因是：_____

(☐) 同意 (二年後開放)，原因是：_____

(☐) 不同意，原因是：_____

以非專屬、無償授權國立中央大學圖書館與國家圖書館，基於推動「資源共享、互惠合作」之理念，於回饋社會與學術研究之目的，得不限地域、時間與次數，以紙本、微縮、光碟及其它各種方法將上列論文收錄、重製、公開陳列、與發行，或再授權他人以各種方法重製與利用，並得將數位化之上列論文與論文電子檔以上載網路方式，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

研究生簽名：_____ 黃 自 達 _____ 學號：_____ 945202062 _____

論文名稱：_____ 以相機取像之中文文件辨識前處理系統 _____

指導教授姓名：_____ 范 國 清 _____

系所：_____ 資 訊 工 程 _____ 所 ☐ 博士班 ☒ 碩士班

日期：民國 96 年 07 月 12 日

備註：

1. 本授權書之授權範圍僅限電子檔，紙本論文部分依著作權法第 15 條第 3 款之規定，採推定原則即預設同意圖書館得公開上架閱覽，如您有申請專利或投稿等考量，不同意紙本上架陳列，須另行加填聲明書，詳細說明與紙本聲明書請至 <http://blog.lib.ncu.edu.tw/plog/碩博士論文專區> 查閱下載。
2. 本授權書請填寫並親筆簽名後，裝訂於各紙本論文封面後之次頁（全文電子檔內之授權書簽名，可用電腦打字代替）。
3. 請加印一份單張之授權書，填寫並親筆簽名後，於辦理離校時交圖書館（以統一代轉寄給國家圖書館）。
4. 讀者基於個人非營利性質之線上檢索、閱覽、下載或列印上列論文，應依著作權法相關規定辦理。

中文摘要

文件傳達許多重要的資訊，如何將文件影像數位化並擷取文字訊息這個議題，隨著數位相機的普及而逐漸受到重視。為了獲得正確的文字辨識結果，一個以相機取像的中文文件辨識前處理系統，必須能處理不同排版格式、多種字體大小的行列、與使用者拍攝產生的文件輕微歪斜等問題，使得擷取出的文字區塊不會產生嚴重的謬誤。

中文文件與英文文件前處理最大的不同，是中文字由多個相連元件組成，如何將組成中文字的相連元件正確合併來進行中文字切割，是擷取中文字訊息最重要的步驟。本文提出的中文字的行列串連演算法與中文字部件合併判斷法則，可以克服影像小角度傾斜與中英字元混雜時的中文字合併問題，並能提供文字區塊讀序的資訊。另外論文中提出兩個文字訊息的保護機制：第一個是反白字區域偵測，通常反白字於相連元件抽取時會被視為背景雜訊過濾掉，為了保護資料完整性，有必要對反白字組成原件另外偵測；第二個是非正向文件的偵測，為了相機取像的便利性，以及使文件內容有較清晰的入鏡範圍，經常有將文件平面垂直光軸旋轉的需要，通常拍攝出的文件影像為接近正矩形的影像，但若要呈現正向的文字內容，可能仍需要旋轉 0° 、 90° 、 180° 、 270° 四種情形的校正。本研究提出一個以統計為基礎的方法，透過分析中文字筆劃的輪廓像素的方向性與文章中的中文字垂直投影波型，總合判斷中文文件的旋轉方向，提供文字識別模組，一個自動化的方向判斷機制。

本論文以名片測試本研究提出的辨識前處理系統，結果文字區塊正

確切割擷取文字影像的成功率可達到 98%，足以證明前處理系統設計方法的正確性。

Abstract

As we know, Chinese documents convey a lot of meaningful and useful information. Due to the popularization of digital cameras, it is convenient to take picture and retrieve important text information from the digitalized Chinese document images. A successful camera-based Chinese document processing system should overcome the problems resulted from various document formats, font sizes, and document skewing to extract correct text block without generating erroneous results.

The major difference between Chinese documents and English documents is that Chinese characters are mainly composed of multiple connected components. The most important step in obtaining the message of the existence of Chinese documents is to merge connected components with correct combining and produce complete Chinese character blocks. In this thesis, we propose a method to link Chinese characters into text line and develop a rule to discriminate the merging condition of ordering connected components to hypothesize the existence of skewing documents. Two mechanisms are developed in the thesis. The first mechanism is the detection of inversed text blocks which may be filtered out as oversize noise blocks in the preprocessing. The second mechanism is the detection of document images laid in incorrect direction because sometimes people will rotate camera 90° or 270° to capture document images. A two pass statistical method is proposed to automatically determine the rotating degree of documents images(0° 、 90° 、 180° 、 270°). The first step is devised by using the phenomenon that horizontal strokes appear more frequently than vertical strokes in Chinese characters. The second step is devised by analyzing the vertical projection histogram of each text block and defining keywords that assist in deciding the rotating degree.

In the experiments, business cards as adopted as the testing samples to verify the performance of the proposed system. The results show that the correct accuracy rate of character segmentation is 98%. It reveals that the proposed system is feasible and effective in accomplishing the goals.

誌謝

本論文承蒙指導教授 范國清博士與傑出學長 溫敏淦博士兩年來在學業上之悉心指導與督促，無論是研究計畫中的實務問題探討，或是論文研究方向，論文的撰寫及修改，兩位老師不吝惜時間、辛勞，給與寶貴的意見與指導，使論文得以順利完成，在此致上最誠摯的謝意。兩位教授不論是在學術研究的成果及待人處世的態度，均讓學生獲益良多。

感謝口試委員林啓芳博士、曾定章博士與鍾國亮博士在口試時給予論文上的寶貴意見與指導，使本論文得以更臻完善。

在此特別感謝圖形識別與人工智慧實驗室博士班的呂信德學長的提攜與照顧，對於學弟妹提出的大小問題，皆樂於傾囊相授，以及謝豐陽、陳志明、莊啟宏、王彥棋、何崗峯學長和碩士班同學：林志瑋、林莉鳳、林準、范聖恩的幫忙，在研究過程中給予寶貴之建議，並陪伴著共同走過兩年之碩士生涯，讓我過得相當充實。亦感謝親切的助理劉佩雯平時幫忙處理實驗室之行政事務，在校園裡為我們奔波勞走。

最後，相當感謝我家人及男友，在這兩年的求學路上不斷的給予我支持與鼓勵，讓我得以專心投入研究而無其他顧慮。

付梓在即，僅以最誠摯的心感謝所有愛護我、關心我的人，以此論文與他們一同分享。

總目錄

中文摘要.....	iii
Abstract.....	v
總目錄.....	viii
圖目錄.....	x
表目錄.....	xiii
第一章.....	1
1.1 研究動機.....	1
1.2 文獻回顧.....	2
1.2.1 文字切割.....	2
1.2.2 文件方向校正.....	4
1.2.3 文件排版分析.....	5
1.3 論文架構.....	6
第二章 文字組成元件偵測.....	8
2.1 彩色至灰階轉換.....	8
2.2 灰階至二值化的轉換.....	9
2.4 反白字區域偵測.....	15
第三章 非正向文件偵測與校正.....	23
3.1 文件 90°或 270°旋轉偵測.....	25
3.2 正反向文件之判別.....	26
3.2.1 垂直投影量統計分析.....	29
3.2.2 輔助方向判讀之關鍵字選取.....	30
3.2.3 文件旋轉方向判讀範例.....	33
第四章 文字行擷取與行中文字切割.....	37
4.1 文字區域群聚切割.....	37
4.1.1 遞迴水平垂直切割方法.....	39
4.2 文字行擷取.....	43
4.3 中文文字行合併.....	45
4.3.1 重疊相連元件行偵測.....	45

4.3.2 傾斜文件的異常合併檢查	46
4.3.3 文字行合併	49
4.4 相連元件語言辨識	50
4.4.1 特徵擷取	51
4.4.2 文字區塊分類器	53
4.5 中文字組成元件合併	55
第五章 實驗結果	58
5.1 中文字合併效能評估	58
5.2 方向判別關鍵字效能評估	65
第六章 結論與未來工作	68
6.1 結論	68
6.2 未來工作	69
參考文獻	70
附錄A	72

圖目錄

圖 1.1 中文文件前處理系統架構圖.....	7
圖 2.1 二值化示意圖.....	10
圖 2.2 以相機取像名片影像光影分布不均圖例	11
圖 2.3 文字組成元件擷取流程圖.....	12
圖 2.4 彩色影像的二值化文字像素擷取過程	13
圖 2.4 八鄰居點示意圖.....	14
圖 2.5 合併類別示意圖.....	15
圖 2.6 反白字區域追蹤起始點偵測.....	18
圖 2.7 人為拍攝瑕疵圖例.....	19
圖 2.8 反白字區域追蹤過程示意圖.....	19
圖 2.2 矩形輪廓追蹤示意圖.....	21
圖 2.3 矩形輪廓追蹤流程圖.....	22
圖 3.1 文件方向判別流程圖.....	24
圖 3.2 Sobel水平濾波器	25
圖 3.3 Sobel垂直濾波器	25
圖 3.4 存在明顯單一波谷類.....	27
圖 3.5 不存在明顯峰谷類.....	27
圖 3.6 存在明顯多個波谷類.....	27
圖 3.7 中文字垂直投影量的統計分布圖	28

圖 3.8 關鍵字擷取流程圖	32
圖 3.9 旋轉 90°的文件影像範例圖	34
圖 3.10 階段一處理後文件影像.....	35
圖 3.11 判斷旋轉方向關鍵字擷取與校正範例.....	36
圖 4.1 名片區塊切割及樹狀結構示意圖	39
圖 4.2 區塊的水平投影及垂直投影量分佈圖	41
圖 4.3 遞迴水平垂直切割流程圖.....	42
圖 4.4 外緣距離示意圖.....	43
圖 4.5 相連元件串連演算法流程圖.....	44
圖 4.6 相連元件串連演算法適用於中文文件的問題	45
圖 4.7 字高異常判斷示意圖.....	48
圖 4.8 文字行類別分裂示意圖.....	48
圖 4.9 文字行合併前.....	49
圖 4.10 文字行合併後.....	49
圖 4.11 方向屬性設定圖	52
圖 4.12 方向屬性設定圖.....	52
圖 4.13 方向屬性設定圖.....	53
圖 4.14 SVM用於二維資料分群示意圖.....	54
圖 4.15 中文字合併流程示意圖.....	57
圖 5.1 無局部區塊語言辨識的中文字合併範例	59
圖 5.2 無反白字區與偵測的中文字合併範例	60

圖 5.3 傾斜文件使用完整流程的中文字合併範例	61
--------------------------------	----

表目錄

表 3.1 組成關鍵字部首表.....	27
表 4.10 混淆字集範例.....	51
表 5.1 中文字合併效能評估(a)無局部區塊語言辨識表.....	62
表 5.2 中文字合併效能評估(b)無反白字區域偵測	63
表 5.3 中文字合併效能評估表(c)完整流程	64
表 5.4 文件方向判斷關鍵字效能評估表	66
表 5.5 加入字頻權重的關鍵字測試表	67

第一章

序論

1.1 研究動機

在這個資訊數位化的時代，人類的生活模式已經高度仰賴數位資訊的傳遞，數位相機取代了傳統相機，行動電話取代了傳統類比式的家用電話，電子郵件取代了傳統信件。而文件中的文字內容數位化，近年也逐漸受到重視，影像中的文字可以表達高度的資訊含量，也較符合人類的閱讀習慣，廣泛可使用在商業交易、社交活動等場合，一個成功的應用實例是將名片與文件影像中的文字註記下來，供後續查詢之用。

一個好的文件前處理系統，可以排除文件格式的限制、擷取完整的文字影像，供給辨識核心模組進行有效率的辨識，這是文字影像辨識非常重要的一環。尤其在以相機取像的文字辨識系統，有許多需要克服的問題，都需要在前處理程序中獲得解決。中文文件影像常見的辨識前處理的問題，例如：中文文件影像文字偵測時，變動的字體大小或中英文字元混雜時對中文字切割的影響；由於相機取像的便利性，中文文件經常旋轉拍攝，造成影像內容上下顛倒或左右傾倒等問題，似乎目前仍無研究自動判斷的機制。於此強調拍攝文件影像與一般路標、招牌等辨識系統存在的問題不同，通常拍攝室外文字(Scence Text)相機取像角度無法讓相機光軸與文字平面垂直，所以擷取的文字影像

必須進行仿射(Algin)校正，但要將相機光軸垂直文件平面拍攝並不困難，故於此不對使用者刻意拍攝透視變形文字影像情形進行討論。

本研究針對目前中文辨識時存在的前處理問題，嘗試提出對應的解決方案，以降低後續文字辨識核心演算法設計的複雜度及提升辨識的效能。

1.2 文獻回顧

文件影像的文字切割、方向校正、排版分析，是文件辨識處理系統中，十分重要的幾項議題。本節將相關文獻中關於文字切割、文件方向校正、文件排版分析三個研究議題提出的觀點及演算方法做整理討論。

1.2.1 文字切割

R. G. Casey[3]提出文字切割是一種將文字串列組成的影像，分解成單獨符號影像區域的程序。在光學文字辨識(OCR)系統中，這是一個決定性的步驟，它決定需要進行文字辨識的區域影像範圍。文字切割的方法，可依照「切割」與「分群」在整個處理過程中的互動來分類。第一類是根據文字的特性，將影像中有意義的組成原件切割出來；第二類則是以辨識為基礎的文字切割，系統會搜尋影像中與符號系統對應到的影像區塊，輔助切割的進行；第三類方法需搭配辨識整個英文單字的辨識系統，故前處理系統只需切割出整個英文單字(Holistic method)，可避免進行字元切割判斷，

降低設計演算法的複雜度。

與本研究相關的文字切割方法屬於直接分割字元法(Dissection Directly Into Characters)，也屬於剛剛提及的第一類方法。此法有三種實作技巧。第一種以白色空白間隙作為切割切割文字的依據，缺點是此技術高度依賴輸入影像的品質，印刷字體太輕或者過重皆會產生錯誤的切割結果，也無法適用於變動的字體寬度；第二種是以投影分析判斷區塊切割，缺點是垂直投影結果對於傾斜文字比較敏感，或者字元間的相連情形與字筆劃有相同厚度時，也無法採用此法；第三種是相連元件抽取，利用矩形環繞方塊(bounding box)進行分析，此法切割非連字的效率比投影分析法快四倍之多[3]，但對於英文連字則必須進行相連元件的分裂處理。

S. Zhao 等[12]提到過度切割是中文字切割的一個主要問題，中文字組成常有兩個以上的字根(radius)而且多數的字根本身也可表示成獨立的中文字，有左右側字根結構是中文字相當常見的一個特徵，例如：「明」有兩個字根「日」、「月」，在做印刷體中文字切割時，可利用字距大於字根間距的特徵將中文字切割。

由於本研究探討的是一般中文文件影像，通常這類文件背景較單純，所以文字切割的方法選用抽取相連元件的類型。由於中文字本身具有多個相連元件組成的性質，故只能初步將組成中文字的相連元件偵測出來，後續整合中文字的動作則需要文件分析提供的局部區域資訊，使得完整中文字區塊可以被正確的擷取。

1.2.2 文件方向校正

文件方向校正，可以分成兩種觀點探討，一種是幾何意義上，將傾斜的文件影像、或者透視變形(perspective distortion)的文件影像，轉換成正向(fronto-parallel view)的平面影像。另一種是以文件分析的觀點，分析文件內容傳達的資訊，判斷文件影像需要旋轉的角度。

傾斜的文件影像偵測，討論的是掃描文件影像因人為誤差所需做的復原動作，可使影像減少傾斜造成的後續處理干擾。Y. Cao 等[2]提出分類方法，主要有五個群組：霍氏轉換(Hough transform)、相互關係函式(cross correlation)、投影分析(projection profile)、Fourier transformation 與 k 鄰居群聚演算法(k-NN clustering)。其中共通的觀念是利用文字行的線性結構，來尋找相連元件最佳的對齊排列，而達到傾斜校正的效果。

透視變形的文件影像校正，是以相機擷取的影像較常發生的問題，當影像平面法向量與相機的光軸存在夾角的時候，需用仿射校正轉換成的正向影像。L. Jagannathan 等[5]提出依照文件影像內容而論的綜合性校正方法，例如：文件影像邊緣可以對投影校正有很大的助益，若有清晰的文件四個矩形邊緣，可直接利用仿射校正成正向影像。但若僅有文件局部區域影像，仍有機會利用排版資訊、規律的文字排列、圖塊、表格等元件，抽取出文件影像中水平方向、垂直方向的線性資訊，計算出水平方向及垂直方向的消失

點，將影像轉換成正向平面影像。

文件的內容通常由較多個數的中文字所組成，要拍攝影像中的文字到具有辨識能力，必須要近距離拍攝，此時相機光軸要垂直文件平面拍攝，對使用者並不困難，故仿射校正的問題在本文中不多做討論。反而是以相機拍攝文件影像，由於相機解析度的緣故，經常需要將文件旋轉拍攝，讓影像涵蓋需要辨識的文字範圍，所以後續章節將介紹的文件方向校正是以文件分析觀點，偵測文件中的文字內容是否為正向，並進行校正。

1.2.3 文件排版分析

林家禎於[14]提到文件切割的技術主要可分為兩種:一種是由上往下(top-down)的切割方法，另一種是由下往上(bottom-up)的切割方法(例如: Region growing)，此兩種方法各有其優缺點。在由上往下的切割方法中，較常遇到的問題是，當文件排版為擁擠、複雜的情形時，不同類型的資料可能無法被切割開來，而影響後續動作的效能。使用由下往上合併的切割方法，合併過程太過繁複且很容易將不同類型的資料合併在一塊，導致合併結果不太理想。

前述的方法僅能使用於曼哈頓排版(Manhattan layout)，定義為文字、圖、網版印刷區域皆可被垂直線及水平線分離的排版方式，而對於非曼哈頓排版的文件影像 H. M. Sun [10]提出了選擇性的 CRLA (Selective Constraint Run-Length Algorithm)，由兩個回合組成，第一回合擷取文字主體，其參數設定考量為避免文字區域與圖形區域連

結在一起；第二回合之參數設定讓較遠間距的文字例如標題也可以串連在一起，則可分離非曼哈頓排版的文件影像不同資料屬性區域。

1.3 論文架構

本篇論文共由六個章節所組成，第二章先介紹文字組成元件偵測，其中包括反白字的組成元件偵測，目的是將所有組成中文字的相連元件擷取出來，以供後續中文字切割時正確的整併；第三章內容是非正向文件的偵測與校正，分為前後兩個階段，階段一使用中文字水平筆劃多於垂直筆劃的性質，判斷是否為垂直走向(90° 或 270° 的傾倒)的文件，當此性質為真時，先逆時鐘旋轉 90° ，使得後續處理文件影像只有正向與反向兩種情形；階段二則為正向與顛倒文件的判別，使用中文字常具有左右偏旁部首的性質，進行垂直投影像分析，可以成功分辨正向與顛倒文件並校正之；第四章內容是文字行擷取與行中文字切割，首要步驟是利用文字區域群聚的性質進行文件區域切割，取得各區域的文字大小以及行距屬性，作為文字行串連、中文字切割的判斷依據，但由於中文文件有少數英、數字元出現的疑慮，在進行中文字切割前，需要進行中文字部件區塊及英、數文字區塊語言屬性的判斷，做為中文部件區塊合併的依據，以達到中文字正確切割及文字行擷取的目的；第五章是一些實驗結果評估；最後，第六章則為結論以及未來研究方向討論。

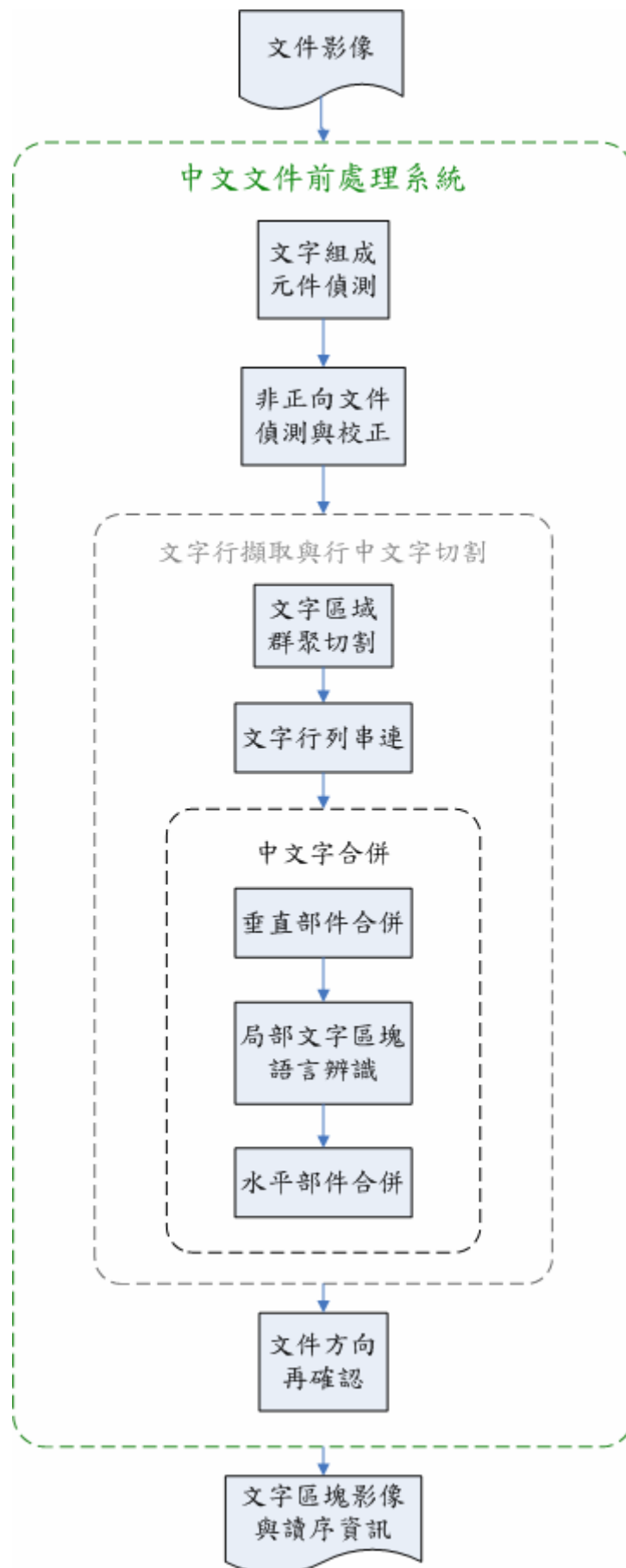


圖 1.1 中文文件前處理系統架構圖

第二章

文字組成元件偵測

通常在文件分析處理系統中，為了偵測文字區塊，會設計一些影像處理流程，化簡資料處理量，並維持原資料於後續系統可用之資訊，增加後續處理的有效性與正確性。本系統使用到的有彩色影像轉灰階、灰階影像二值化、相連元件(connected component)抽取、反白字區域偵測。

2.1 彩色至灰階轉換

相機擷取的彩色文件影像，每一個像素(pixel)都由三個位元組所組成，其成分表示紅(red)、綠(green)、藍(blue)三種色彩資訊。為了減少影像處理所需要的資料量，並且加速相對的處理速度，本文件處理系統需漸進轉換彩色影像至二值化影像，本節將介紹彩色影像轉為灰階影像的方法，灰階影像轉為二值化影像的方法將在下節介紹。將影像的色彩資訊由 RGB 三維色彩空間轉換成一維的灰階(Gray)色彩空間轉換公式定義如下：

$$Gray = \frac{1}{3}(R + G + B) \quad (2.1-1)$$

2.2 灰階至二值化的轉換

將灰階影像(X)轉換至二值化影像(Y)的方法，首先求得影像的灰階統計分布圖(histogram)，再採用以分群方法為基礎的 Otsu 演算法，決定一個二值化影像的閾值(threshold) T。P. S. Liao 等[7]描述 Otsu 的演算法，一張灰階影像可以表示成 2D 灰階值函式，包含 N 個像素點，灰階值介於 K 至 L，具有共同灰階值 i 的像素點個數為 f_i ，則影像中灰階值 i 的出現機率可表示為 2.1-2 式，在二階(bi-level)二值化情形下，像素點將分成兩個類別，分別為 C_1 具有 $[K, \dots, t]$ 的灰階值像素， C_2 具有 $[t+1, \dots, L]$ 的灰階值像素，則兩個類別的灰階值機率分佈可表示為 2.1-3 式，其中 $\omega_1(t)$ 表示灰階值 K 至 t 的機率值加總， $\omega_2(t)$ 表示灰階值 t+1 至 L 的機率值加總， C_1 與 C_2 與整張影像的灰階平均值以 μ_1 、 μ_2 、 μ_T 代表，計算公式如 2.1-4、2.1-5 式，Otsu 演算法定義的群間變異數(between-class variance)，可使用 2.1-6 式求值。Otsu 演算法取得的二值化閾值最佳解，必須要在影像灰階值 K 至 L 之間選擇一個參數 t^* 產生最大群間變異數 σ_B^2 (如式 2.1-7)，則 t^* 為 Otsu 演算法輸出的二值化閾值。

$$p_i = f_i / N \quad (2.1-2)$$

$$\begin{aligned} C_1 : & p_K / \omega_1(t), \dots, p_t / \omega_1(t) \quad \text{and} \\ C_2 : & p_{t+1} / \omega_2(t), p_{t+2} / \omega_2(t), \dots, p_L / \omega_2(t) \end{aligned}$$

where $\omega_1(t) = \sum_{i=K}^t p_i, \quad \omega_2(t) = \sum_{i=t+1}^L p_i$ (2.1-3)

$$\mu_1 = \sum_{i=K}^t i * p_i / \omega_1(t), \quad \mu_2 = \sum_{i=t+1}^L i * p_i / \omega_2(t) \quad (2.1-4)$$

$$\omega_1\mu_1 + \omega_2\mu_2 = \mu_T \quad (2.1-5)$$

$$\sigma_B^2 = \omega_1(\mu_1 - \mu_T)^2 + \omega_2(\mu_2 - \mu_T)^2 \quad (2.1-6)$$

$$t^* = \underset{K < t < L}{\text{Arg}} \quad \text{Max} \{ \sigma_B^2(t) \} \quad (2.1-7)$$

以 Otsu 演算法取得的閾值為標準，將影像像素(pixel)灰階值大於閾值者設其值為 255(表示白色)，小於閾值者設其值為 0(表示黑色)，如式：

$$Y(i, j) = \begin{cases} 0 & (\text{object}) \quad \text{if } X(i, j) \leq T \\ 255 & (\text{background}) \quad \text{if } X(i, j) \geq T \end{cases} \quad (2.1-8)$$

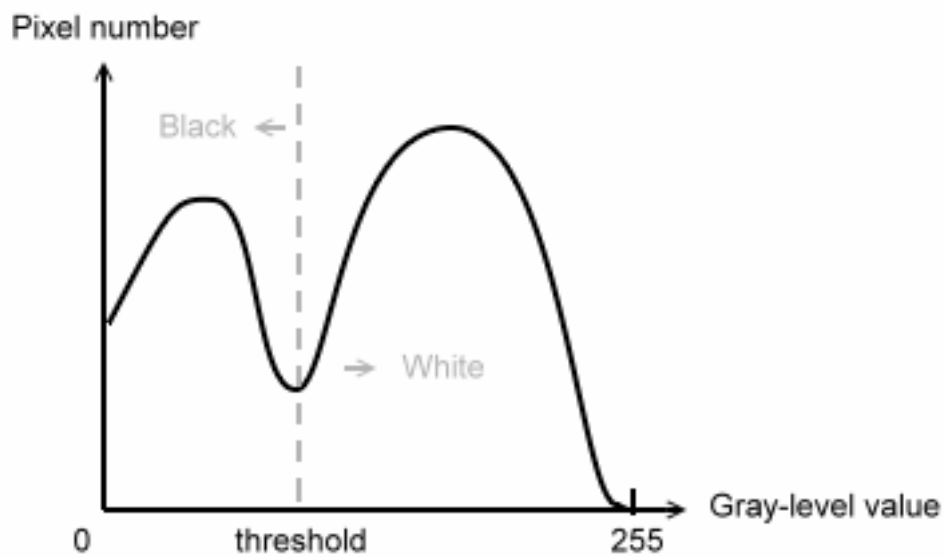


圖 2.1 二值化示意圖



圖 2.2 以相機取像名片影像光影分布不均圖例

以相機取像時，受限於相機鏡頭取像的物理限制，一般極易產生同心圓式的中間亮邊緣暗的不均勻光影效果(如圖 2.2 所示，本圖使用灰階的 256 色降為 16 色處理)。由於 Otsu 演算法考慮全域影像的灰階統計分布(histogram)，當輸入影像的亮度(intensity)分佈較不均勻時，全域 Otsu 取得的臨界值進行局部區域二值化時常造成不必要的雜訊。為了使取得的二值化影像，保留較清楚的文字影像資訊，我們採用 Otsu 演算法的變形。在流程上使用兩階段的灰階二值化動作，第一階段進行全域灰階影像 Otsu 二值化，產生初步的二值化文字影像像素並進行文字相連元件的抽取；接下來，第二階段使用區域範圍進行灰階影像二值化，可利用下節介紹的相連元件(connected component)抽取方法，取得二值化影像中各個相連元件環繞矩形範圍，以每個相連

元件環繞矩形為範圍，再進行區域灰階影像的 Otsu 二值化，好處是此時計算二值化閾值所參考鄰近區域像素點是最少的，可將影像亮度分布不均勻時的影響最小化，並且中文字字體筆劃較粗，使用小範圍的區域二值化，較不容易產生筆畫連接的問題。因此相機取得的影像在本節提出的二值化處理當中，仍然可以將相機物理性質造成的光影問題影響減小，使較正確的二值化文字像素得以保留下來。

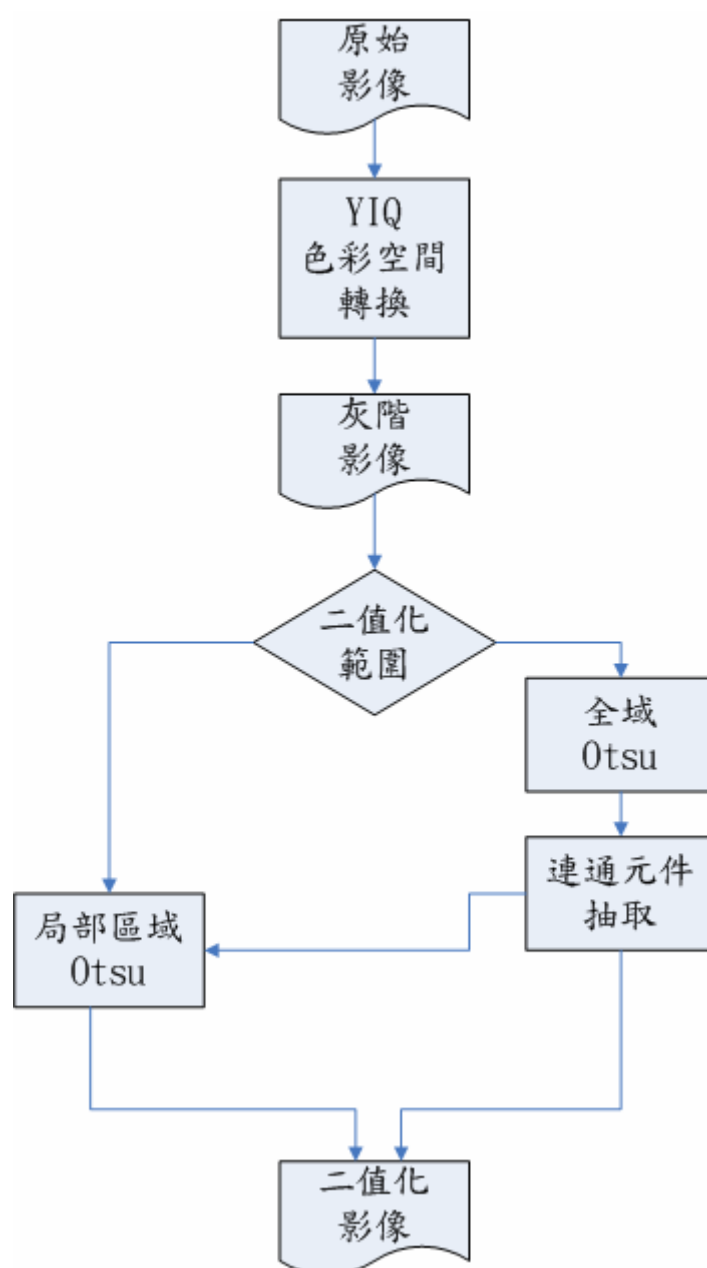


圖 2.3 文字組成元件擷取流程圖

(a)彩色影像	 <p>威播科技股份有限公司 新竹市建功路 26 號 3 樓之1 統一編號：70648971 T e l: 886-3-5744855 F a x: 886-3-5735893 E-mail: alex@broadweb.com.tw U R L: www.broadweb.com.tw</p>
(b)灰階影像	 <p>威播科技股份有限公司 新竹市建功路 26 號 3 樓之1 統一編號：70648971 T e l: 886-3-5744855 F a x: 886-3-5735893 E-mail: alex@broadweb.com.tw U R L: www.broadweb.com.tw</p>
(c)全域 Otsu 影像	 <p>威播科技股份有限公司 新竹市建功路 26 號 3 樓之1 統一編號：70648971 T e l: 886-3-5744855 F a x: 886-3-5735893 E-mail: alex@broadweb.com.tw U R L: www.broadweb.com.tw</p>
(d)區域 Otsu 影像	 <p>威播科技股份有限公司 新竹市建功路 26 號 3 樓之1 統一編號：70648971 T e l: 886-3-5744855 F a x: 886-3-5735893 E-mail: alex@broadweb.com.tw U R L: www.broadweb.com.tw</p>

圖 2.4 彩色影像的二值化文字像素擷取過程

2.3 相連元件的抽取

在二值化影像中擷取相連元件的方法，採用傳統的 8 個相連元件標記演算法(8-connected component labeling algorithm)，[18]有兩個步驟：

步驟一、掃描(scanning)

掃描影像中的像素，由左而右、由上而下，假設我們遇到第一個黑像素(pixel)，設定一個新標記(label)，此標記表示該像素已經被掃描過，以順時針方向檢查該像素點的鄰居像素是否為黑像素，如果結果為真，鄰居中的黑像素設定成相同的標記，並且循環的檢查新的 8 個鄰居點。圖 2.4 顯示一個像素的 8 個相連元件。如果一個某個黑像素的 8 鄰居點皆為白色，或者皆已標記，退回上一個黑像素檢查，循環追蹤的原則是沒有標記過的像素點才會被納入考慮，所有的像素點都需被追蹤一次。

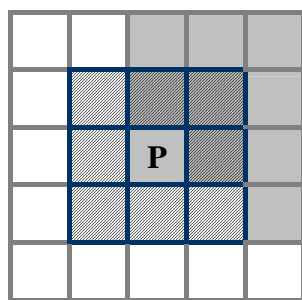


圖 2.4 八鄰居點示意圖

其中斜線區域表示 p 像素點的八個鄰居點

步驟二、合併類別(merging class)

考慮標記合併時，需再做一次由左而右、由上而下掃描，若存在兩相鄰黑像素所屬標記不同時，表示這兩個標記屬於等價類別(如圖 2.5(a))。所有等價類別中的標記需要取其一作為代表(如圖 2.5(b))，則影像中所有的連通單元可被成功的標記出來。

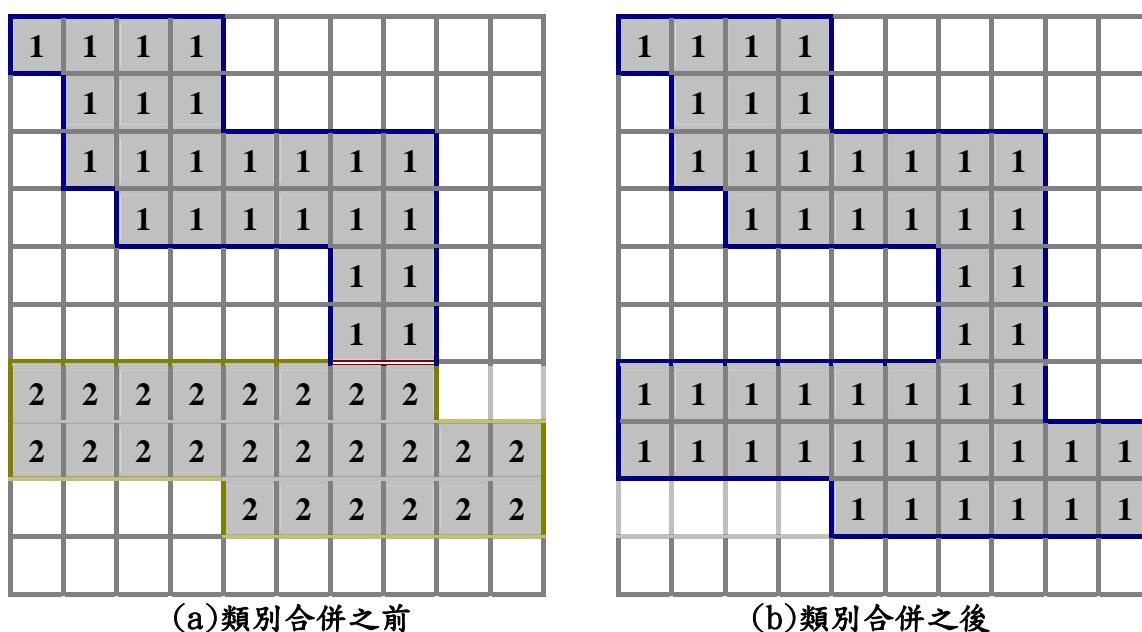


圖 2.5 合併類別示意圖

2.4 反白字區域偵測

名片或報章雜誌中，有時會有以反白字區塊所呈現的標題字樣。使用一般相連元件抽取的方法擷取文字區塊，往往會喪失反白字區塊內的重要訊息，所以本節利用文件中常出現的一些反白字區塊特徵，例如面積、密度、文字包含於四邊形區域內...等等，將 2.3 節中相連元件抽取

的方法擴充，當抽取出的相連元件偵測為反白字區塊時，將二值化影像 $B(i,j)$ 中該區塊範圍像素反白(黑白像素轉換如 2.4-1 式)，並且對反白字區域重新抽取相連元件。

每個相連元件皆需進行反白字區域偵測，首先利用反白字區域的形狀及密度作為篩選特徵，由於正常文件影像背景像素點與文字像素點比例會相當的懸殊，文字與背景像素反白後依然具有這樣的性質，故設定反白字區域的黑點密度 d 滿足 $\alpha_1 < d < \alpha_2$ ，規範了反白字區域中合理的文字、背景像素點比例；再者，本文偵測反白字區域，是為了增加資訊傳達的完整性，所以反白字區域的範圍，必須可以包含足以傳達文字意涵數量的文字區塊，則反白字區域的寬(w)、高(h)必須滿足 $w \in \{w > \beta_1 * ccw\}$ 或 $h \in \{h > \beta_2 * cch\}$ 的條件。(平均相連元件寬度以 ccw 、高度以 cch 表示)。由以上兩個特徵條件篩選完，仍滿足反白字候選區塊性質者，將利用矩形輪廓追蹤，檢查是否具有近似矩形輪廓的反白字區塊特質。

$$B'(i, j) = \begin{cases} 0 & \text{if } B(i, j) = 255 \\ 255 & \text{if } B(i, j) = 0 \end{cases} \quad (2.4-1)$$

由觀察發現，常見於名片及文件中的反白字，大多包含於封閉的四邊形區域內。為提供解決普遍情形的反白字偵測的演算法，本論文研究假設反白字僅會出現於近似矩形的相連元件範圍中，假設反白字矩形區塊的左側、下緣、右側、上緣四個區段，以 Bry_1 、 Bry_2 、 Bry_3 、 Bry_4 為代表。

尋找反白字區域的第一個步驟是尋找追蹤起始點。反白字矩形候選區塊的四個區段的邊緣上，以相連元件環繞矩形深度 20 pixels，作為尋找反白字區塊輪廓(contour)追蹤起始點範圍，分別為左側(R_1)、下緣(R_2)、右側(R_3)、上緣(R_4)四個搜尋範圍，定義 $p(i, j)$ 為影像中二維座標 (i, j) 位置的點， $R_{k=\{1,2,3,4\}}(i, j)$ 表示二值化影像位置 (i, j) 的值。由公式 2.4-2 至公式 2.4-9，可求得四個邊緣上的追蹤起始點 $S_{k \in \{1,2,3,4\}}$ 、終止點 $E_{k \in \{1,2,3,4\}}$ 。

$$S_1 = \left\{ \min_i p(i, j) \mid \min_j (R_1(i, j) = 0 \ \& \ p(i, j) \in R_1) \right\} \quad (2.4-2)$$

$$S_2 = \left\{ \max_j p(i, j) \mid \min_i (R_2(i, j) = 0 \ \& \ p(i, j) \in R_2) \right\} \quad (2.4-3)$$

$$S_3 = \left\{ \max_i p(i, j) \mid \min_j (R_3(i, j) = 0 \ \& \ p(i, j) \in R_3) \right\} \quad (2.4-4)$$

$$S_4 = \left\{ \min_j p(i, j) \mid \min_i (R_4(i, j) = 0 \ \& \ p(i, j) \in R_4) \right\} \quad (2.4-5)$$

$$E_1 = \left\{ \min_i p(i, j) \mid \max_j (R_1(i, j) = 0 \ \& \ p(i, j) \in R_1) \right\} \quad (2.4-6)$$

$$E_2 = \left\{ \max_j p(i, j) \mid \max_i (R_2(i, j) = 0 \ \& \ p(i, j) \in R_2) \right\} \quad (2.4-7)$$

$$E_3 = \left\{ \max_i p(i, j) \mid \max_j (R_3(i, j) = 0 \ \& \ p(i, j) \in R_3) \right\} \quad (2.4-8)$$

$$E_4 = \left\{ \min_j p(i, j) \mid \max_i (R_4(i, j) = 0 \ \& \ p(i, j) \in R_4) \right\} \quad (2.4-9)$$

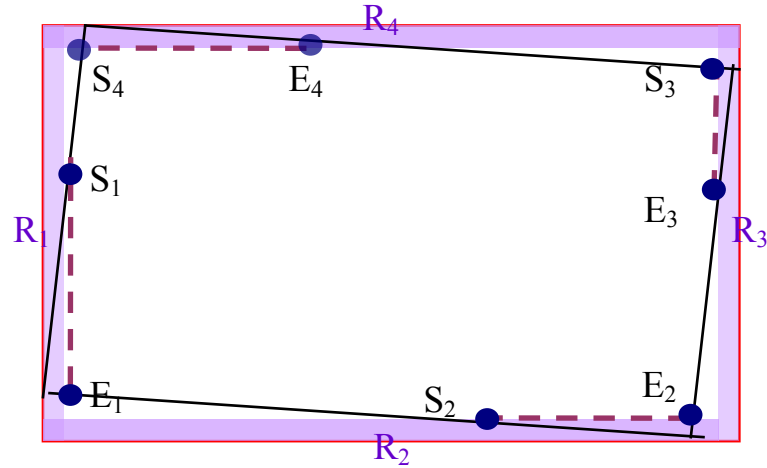


圖 2.6 反白字區域追蹤起始點偵測

第二個步驟是矩形輪廓追蹤(square contour tracing)[20]。先選取某側(預設為 L)為起始追蹤相連元件環繞矩形的邊緣區段，由步驟一求得的 S_k 、 E_k 兩點，以離相連元件環繞矩形轉角點較遠者為追蹤起始點，是為了避免找到的追蹤起始點落於其他側的邊緣區段。先進行順時針方向繞行反白字左側(Bry_1)、上緣(Bry_4)、右側(Bry_3)、下緣(Bry_2)四個邊緣，當經過小區段的邊緣雜點或邊緣殘缺(如圖 2.7)導致的追蹤中斷，嘗試以同一個追蹤起始點，則進行逆時針方向繞行反白字左側(Bry_1)、下緣(Bry_2)、右側(Bry_3)、上緣(Bry_4)四個邊緣，若順時針與逆時針方向搜尋的結果整合，可滿足找到矩形區域四個轉角點的條件，則可以將反白字矩形區域擷取出來。定義起始追蹤像素 s ，為目前的像素 p ，且相對於追蹤方向的下一個像素為 n ，追蹤偏差角為 θ ，探索轉角距離閾值 ε ，預設追蹤方向為順時鐘，clockcycle 設為真。



圖 2.7 人為拍攝瑕疵圖例

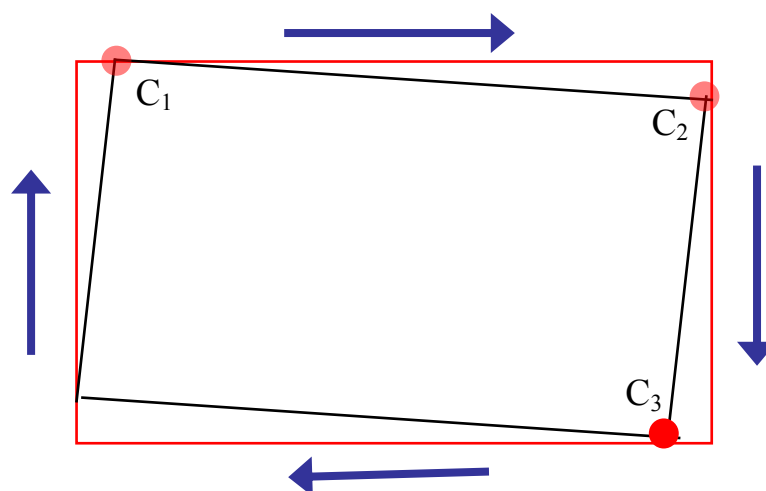


圖 2.8 反白字區域追蹤過程示意圖

反白字區域追蹤演算法：

Phase I - 尋找追蹤起始點

Step1 選擇起始追蹤區段 Bry_k (對應的搜尋區域為 R_k)， $k \in \{1,2,3,4\}$ 。

Step2 R_k 範圍中由上而下，由外往內尋找到的第一個黑像素點為 S_k 。

Step3 R_k 範圍中由下而上，由外往內尋找到的第一個黑像素點為 E_k 。

Step4 S_k 、 E_k 兩者分別計算與相連元件環繞矩形中最近頂點距離，將距離較遠的設為該區段追蹤起始點 S_k ，設定 clockcycle 為真，設定 s、n 像素與 S_k 位置相同。

Phase II - 矩形輪廓追蹤

Step5 若 n 為黑像素則執行 Step6，若 n 為白像素則執行 Step7。

Step6 (1) 若 clockcycle 為真，n 像素往前進方向的左側平移，直到 n 的左鄰像素為白像素，執行 Step8。

(2) 若 clockcycle 為假，n 像素往前進方向的右側平移，直到 n 的右鄰像素為白像素，執行 Step8。

Step7 (1) 若 clockcycle 為真，n 像素往前進方向的右側平移，直到 n 為黑像素。

(2) 若 clockcycle 為假，n 像素往前進方向的左側平移，直到 n 為黑像素。

Step8 若 $n=s$ ，執行 Step10；若 $n \neq s$ 且滿足 (1) $\overline{np} < \varepsilon$ (2) \overline{ns} 與追蹤方向夾角小於 θ ，設 p 為 n 像素，執行 Step5；否則執行 Step9。

Step9 (1) 若 clockcycle 為真，則改追蹤 $Bry_{(k+3)\%4}$ 區段，紀錄轉角點 $C_{(k+3)\%4}$ 為 p 像素，s 為 $C_{(k+3)\%4}$ 像素，回到 Step5。

(2) 若 clockcycle 為假，追蹤 $Bry_{(k+1)\%4}$ 區段，紀錄轉角點 $C_{(k+1)\%4}$ 為 p 像素，s 為 $C_{(k+1)\%4}$ 像素，回到 Step5。

Step10 若追蹤過程中取得的 $\{C_1、C_2、C_3、C_4\}$ 可組成近似矩形則成功；
 若否且 clockcycle 為假，將 s 與 n 設成 S_k ，clockcycle 設為真，回
 到 Step5；以上條件皆不成立時追蹤失敗，結束。

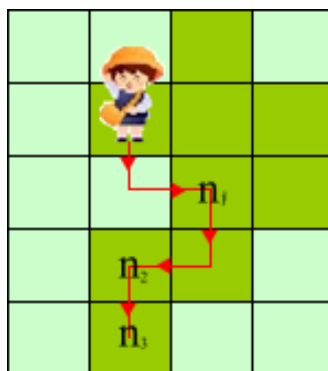


圖 2.2 矩形輪廓追蹤示意圖

每個反白字候選區域經過密度、長、寬、矩形輪廓特質等等篩選，
 將具有以上所有特值的相連元件區塊，判定該環繞方塊範圍，包含著近
 似矩形的反白字區域。之前使用矩形輪廓追蹤演算法取得的四個轉角點
 $\{C_1、C_2、C_3、C_4\}$ ，可用其組成四條代表矩形輪廓邊界的線性方程式，
 將包含於四條線性方程式的二值化像素反白，並將此範圍內的二值化像
 素重新抽取相連元件，則可擷取包含反白字區和其他區域範圍中的所有
 文字組成元件，使得後續流程有更完整的資訊，進行文字行擷取與行中
 文字的切割。實驗中使用的經驗值為 $\alpha_1 = 0.9$ ， $\alpha_2 = 0.7$ ， $\beta = 10$ 。

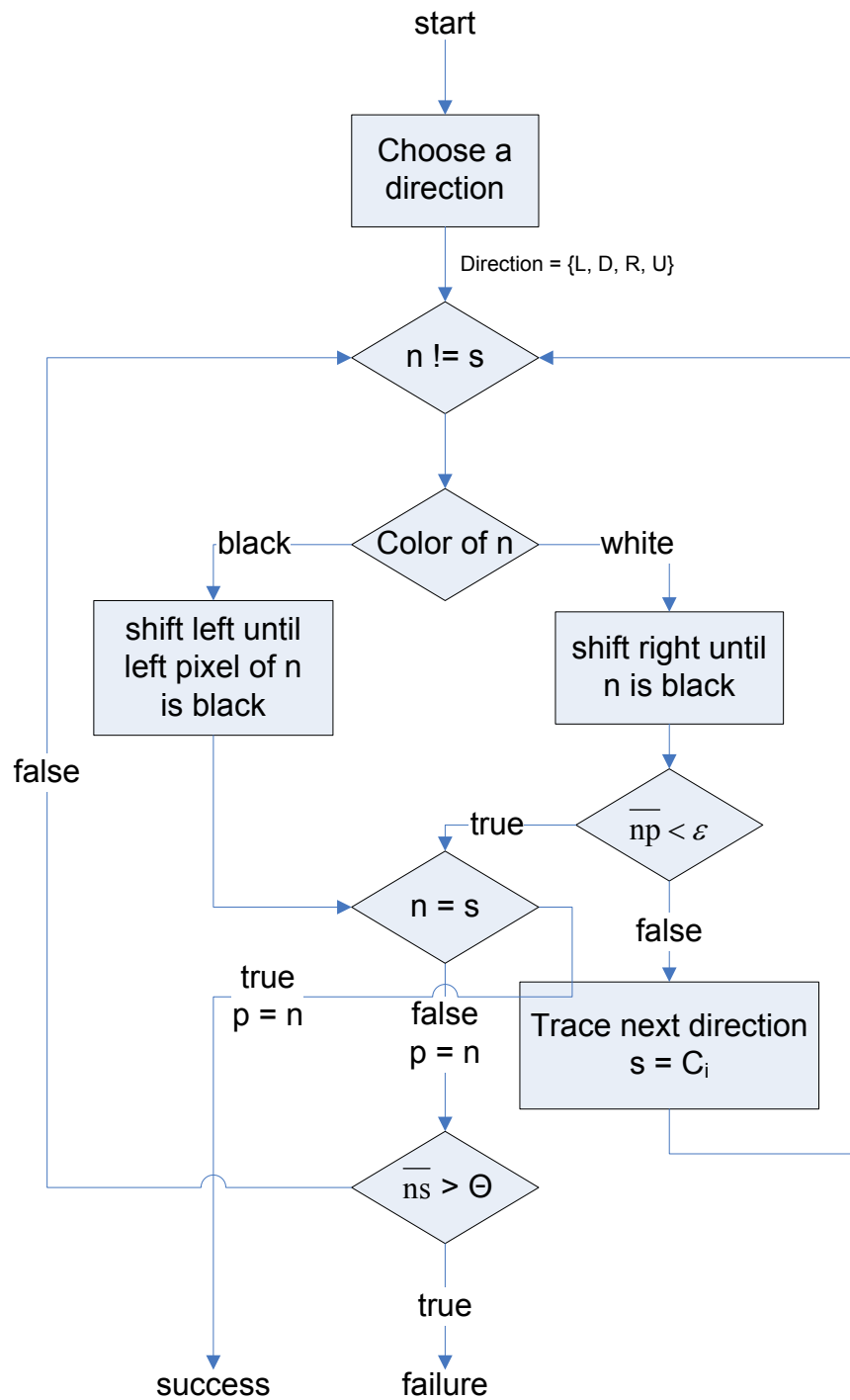


圖 2.3 矩形輪廓追蹤流程圖

第三章

非正向文件偵測與校正

文件的方向校正可以分為兩種類型，第一種為傾斜偵測與校正，經由仿射(affine)校正克服相機取得影像時存在相機視角偏差的問題(perspective distortion)，當影像平面與相機的光軸非垂直關係時，使用仿射校正轉換成正向影像(fronto-parallel view)。另一種類型是以文件分析角度探討，當輸入的是相機拍攝的文件影像，由於相機取像的便利性，通常會將文件放置成水平或是垂直方向，並且將相機光軸垂直文件平面拍攝，使得在有限的相機解析度限制下，拍攝出足夠清晰的文字段落影像，此時文件內容的呈現方向，仍然可能有順時鐘方向旋轉 0° 、 90° 、 180° 、 270° 四種情形，就算是經過仿射校正的路標及招牌影像亦是如此。當有中文文件影像大量批次處理的需要時，若不考慮耗時的人工校正方法，非正向中文文件影像辨識將會造成很高的錯誤率。本章討論的文件方向校正，是以文件分析角度切入，在中文文件影像的前處理流程中，提供一個文件方向確認的機制，使得前處理流程與文字辨識模組的銜接更加的自動化。

本文提出的文件方向校正方法是以統計原理為基礎，共分為兩個階段，第一個階段利用邊緣偵測，取得影像文字輪廓上水平方向像素與垂直方向像素的差異，判斷文件內容是否為垂直文件影像(90° 或 270° 旋轉

後的文件內容影像)。第二個階段，透過分析文件影像內，中文字垂直投影分布，得知中文文件是否需要旋轉 0° 或 180° 角度校正，視為文件方向的再確認。由於中文字垂直投影分析，需要輸入整合好的文字區塊，於文字行擷取與中文字切割之後進行(如圖 3.1 流程圖)，以判斷確認文件方向。

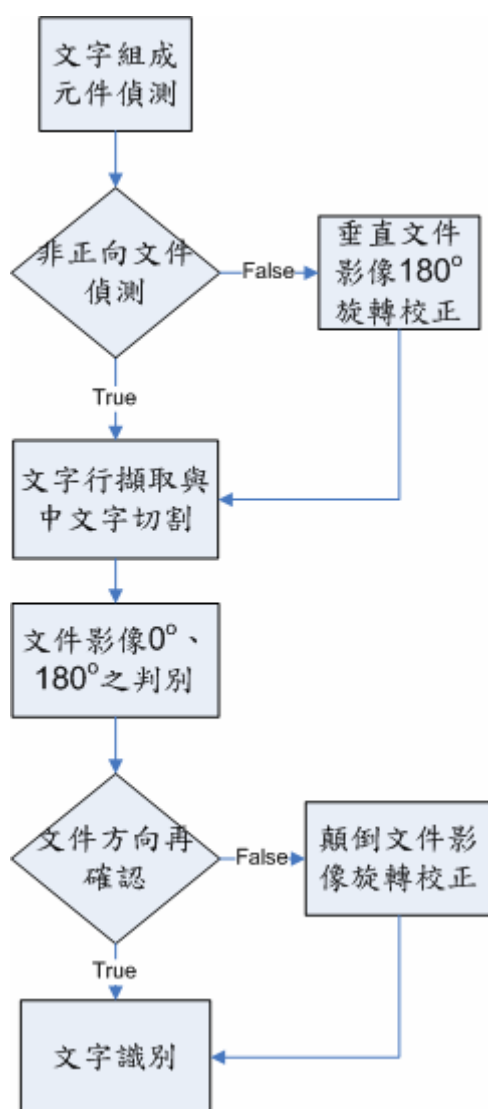


圖 3.1 文件方向判別流程圖

3.1 文件 90° 或 270° 旋轉偵測

中文字的組成具有水平筆劃多於垂直筆劃的特徵，此特徵有利於判斷中文文件內容為垂直走向文件(經 90° 或 270° 旋轉)或是水平走向文件(經 0° 或 180° 旋轉)。為了簡化抽取文字筆畫問題的複雜度，每個中文字影像使用 Sobel[19]水平濾波器(horizontal filter)、垂直濾波器(vertical filter)如圖 3.2 與圖 3.3，取得文字輪廓線上水平方向的像素點及垂直方向的像素點。累計所有垂直方向的像素點(edge pixel)個數 SUM_V ，水平方向的像素點個數 SUM_H ，則此統計上可能有有三種結論，(a)當式 3.1-1 成立，並且 SUM_V 大於 SUM_H ，則有足夠信心此影像中文內容為旋轉 90° 或 270° 的文件影像，為了後續處理的方便，將此影像做 90° 旋轉；(b)若式 3.1-1 成立，並且 SUM_H 大於 SUM_V ，則有足夠信心此影像中文內容為正向或反向旋轉；(c)若式 3.1-1 不成立，影像中水平及垂直方向像素點個數不夠懸殊，則不進行後續文件方向校正流程，以警示訊息提醒使用者需進行人工文件方向校正。式 3.1-1 的物理意義是判斷 SUM_V 、 SUM_H 兩者的懸殊程度，當兩者的懸殊程度越高，旋轉的信心指數越高， η 以經驗值定為 2%。

-1	-2	-1
0	0	0
1	2	1

-1	0	1
-2	0	2
-1	0	1

圖 3.2 Sobel 水平濾波器

圖 3.3 Sobel 垂直濾波器

$$\frac{|SUM_V - SUM_H|}{\max(SUM_V, SUM_H)} > \eta \quad (3.1-1)$$

3.2 正反向文件之判別

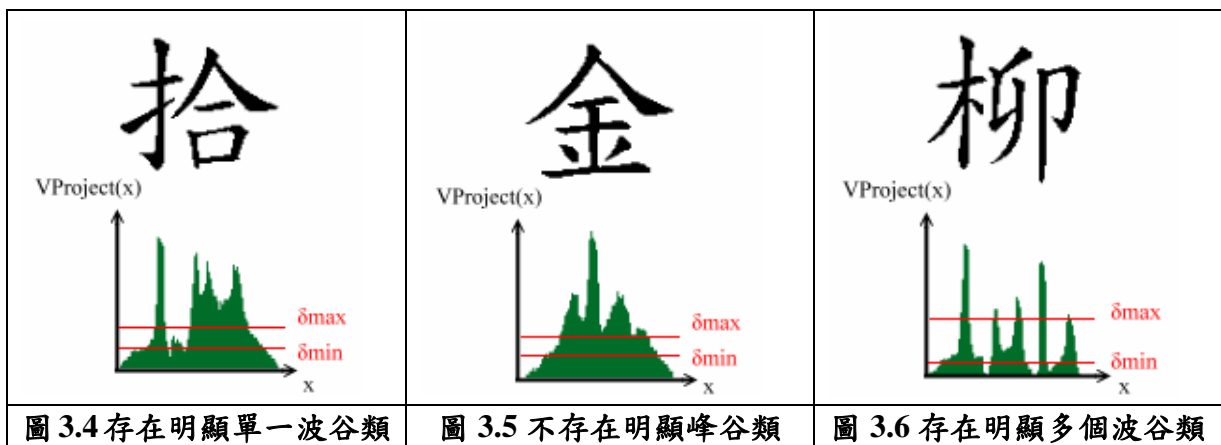
與前一節中所處理的結果，已可得知文件為垂直或水平文件，並在必要的旋轉後成為水平文件。然而水平文件仍有正向(0°)、反向(180°)的旋轉問題有待克服。本節中將利用中文字的部首特徵，進行正反向文件偵測，並進行文件走向之旋轉校正。

中文字的組成有部首的概念，康熙字典定義的部首有 214 個[22]，其中偏旁的部首較容易從中文字影像中切割出來。在左與右側的部首稱為旁，例如：亻、扌、刂、冫；在上者、或外包文字上及左邊者，稱為頭，例如：宀、艹、疒；同時上下夾著文字，稱為框，例如：口；在下面的稱為底，例如：辶。觀察中文字中有不少比例是由左、或左上、左下偏旁部首所組成，故嘗試整理了康熙字典中常用的左、右側偏旁部首共 35 個，如表 3.1 所示，其中常見的右偏旁部首僅有 刂、力、冫、卜，四者，其他 31 個部首皆是瘦長的左偏旁，或是垂直投影量集中於左側的左上(例如：疒)、左下(例如：辶)偏旁，本研究試圖用垂直投影量分布來分析這個現象。

表 3.1 組成關鍵字部首表

	二劃	三劃	四劃	五劃	六劃	七劃	八劃
組成關鍵字部首	イ ン リ 卜 力	口 土 女子 彳 忄 才 シ 尸 广 爿	日 木 月 牛 彡 ネ 辶	禾 米 糸 ネ 玉	虫 耳	走 足 言	金 食

常用中文字，以垂直投影量的統計分布圖(histogram)分析，觀察其特性大約可以分為三類:存在明顯單一波谷類、存在明顯多個峰谷類、不存在明顯波谷類(如圖 3.4、3.5、3.6)。其中我們探討僅存在「明顯單一波谷類型」的中文字，並應用左偏旁中文字的高出現頻率，分析文件之走向。用垂直投影像分布圖中的波谷為參考點(如圖 3.7)，可將文字區塊分割為左右兩個子區塊(其寬度可用 w_L 、 w_R 表示)。



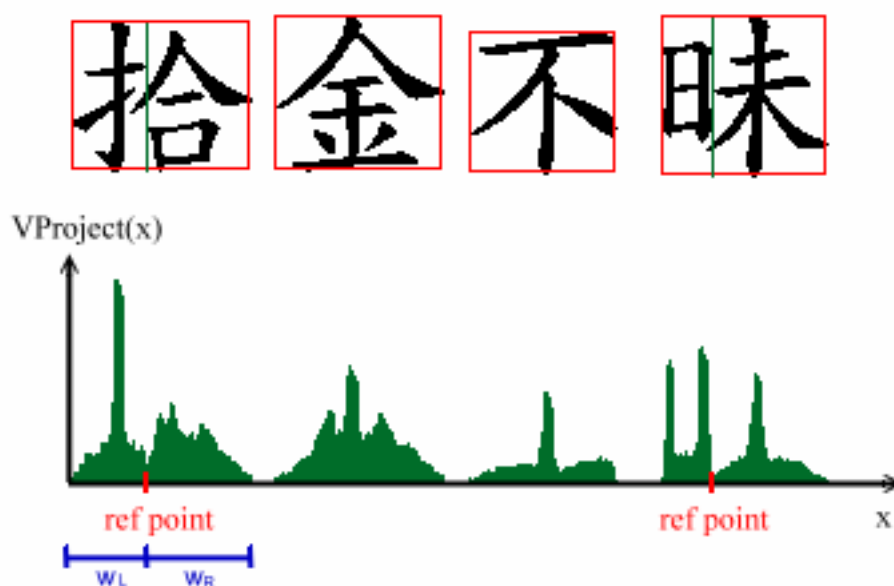


圖 3.7 中文字垂直投影量的統計分布圖

由之前整理中文字特性得到的結論，左側偏旁中文字與右側偏旁中文字相比，比例明顯偏多，故統計文件中存在明顯單一波谷類中文字數，扣除左右側區塊大小接近，容易在統計上造成變異的字數($w_L \approx w_R$ ，例如：的)，正向中文字的 w_L 、 w_R 關係大多是 $w_L > w_R$ ，使用這些左右字集的中文字為關鍵字，即可以統計整張影像各個關鍵字的左右區塊寬度關係，判斷影像為正向(0°)或顛倒影像(180°)。

3.2.1 垂直投影量統計分析

每一個中文字區塊的垂直投影量計算需輸入文字二值化影像 $Y(x, y)$ ，則任意行(column)位置 x 得到的垂直投影量如 3.2-2 式，則每個中文字可以對應到一個垂直投影量分布波形。

$$B(x, y) = \begin{cases} 0 & \text{if } Y(x, y) = 255 \\ 1 & \text{if } Y(x, y) = 0 \end{cases} \quad (3.2-1)$$

$$VProject(x) = \sum_{y=0}^h B(x, y) \quad (3.2-2)$$

本文以統計的法則，對於不同中文字的波形，定義動態的波峰、波谷閾值，由公式 3.2-3、3.2-4 可求得判斷波峰的閾值 δ_{\max} 、與判斷波谷的閾值 δ_{\min} ，以該字的垂直投影波形的平均值(avg)與標準差(std)做為統計量代表的參數。得到波峰、波谷閾值後，可將中文字的波形區分成三個區段， $y \geq \delta_{\max}$ 、 $\delta_{\min} \leq y \leq \delta_{\max}$ 、 $y \leq \delta_{\min}$ (如 3.4、3.5、3.6 圖所示)，當像素點(x,y)的高度 y 值大於 δ_{\max} 表示落於波峰範圍，當高度 y 值小於 δ_{\min} 表示落於波谷的範圍，描述波峰 (peak) 區間、波谷 (trough) 區間的公式如 3.2-5、3.2-6。則可將輸入中文字的垂直投影波形，依波峰、波谷間的關係分類，如圖 3.4 僅存在一個波谷區段在兩個波峰之間，故分類到「存在明顯單一波谷類」；圖 3.5 不存在任何

的波谷區段在兩個波峰之間，故分類到「不存在明顯波谷類」；圖 3.6 存在多個波谷區段在兩個波峰之間，故分類到「存在明顯多個波谷類」。詳細篩選輔助文件方向判斷的關鍵字的方法，將於 3.2.2 節介紹。

$$\delta_{\max} = avg + \alpha * std \quad (3.2-3)$$

$$\delta_{\min} = avg + \beta * std \quad (3.2-4)$$

$$peak(x_i, x_j) = \{[x_i, x_j] \mid x_i \leq x_k \leq x_j \ \& \ VProject(x_k) > \delta_{\max}\} \quad (3.2-5)$$

$$trough(x_i, x_j) = \{[x_i, x_j] \mid x_i \leq x_k \leq x_j \ \& \ VProject(x_k) < \delta_{\min}\} \quad (3.2-6)$$

3.2.2 輔助方向判讀之關鍵字選取

得到每個中文字的垂直投影量的統計分布圖後（如圖 3.7），我們可以利用上節提出的垂直投影量分佈分析，篩選出「明顯單一波谷類型」的關鍵字。

如果存在 $trough(x_i, x_j)$ 介於 $peak(x_t, x_s)$ 、 $peak(x_p, x_q)$ 之間，其中 $trough(x_i, x_j)$ 是唯一具有全域垂直投影量最小值的波谷區間並且 $t < s < i < j < p < q$ ，則以此投影量最小值位置作為參考點(ref point)，將文字區塊分割為左右兩個子區塊(b_L 、 b_R)，當 3.2-7 式成立時，此關鍵字中的參考點才具有文件方向區分力。在 $w_L < w_R$ 的條件下，此

中文字為代表文件為正向的關鍵字，pctr(正向計數器)加一；若是在 $w_L > w_R$ 的條件下，此中文字為代表文件為顛倒方向的关键字，nctr(反向計數器)加一。最後統計正向、反向計數器數值，當 3.2-8 式成立的時候，得到的旋轉方向結論是可信賴的，以經驗值訂定 δ_{rotate} 為 80%。

$$\frac{\min(w_L, w_R)}{\max(w_L, w_R)} \leq \delta_{ratio} \quad (3.2-7)$$

$$\frac{|pctr - nctr|}{\max(pctr, nctr)} \geq \delta_{rotate} \quad (3.2-8)$$

總結，中文文件影像的旋轉判斷可能有三種，(1)當 pctr 大於 nctr 並且帶入 3.2-8 式的結果高於 δ_{rotate} 閾值時，文件為正向，故不做旋轉處理；(2)當 pctr 小於 nctr 並且帶入 3.2-8 式的結果高於 δ_{rotate} 閾值時，文件為逆向，故須做 180° 旋轉；(3)當 3.2-8 式的結果小於 δ_{rotate} 閾值時，由於關鍵字比例不夠懸殊，故不進行後續文件方向校正流程，以警示訊息提醒使用者需進行人工文件方向校正。

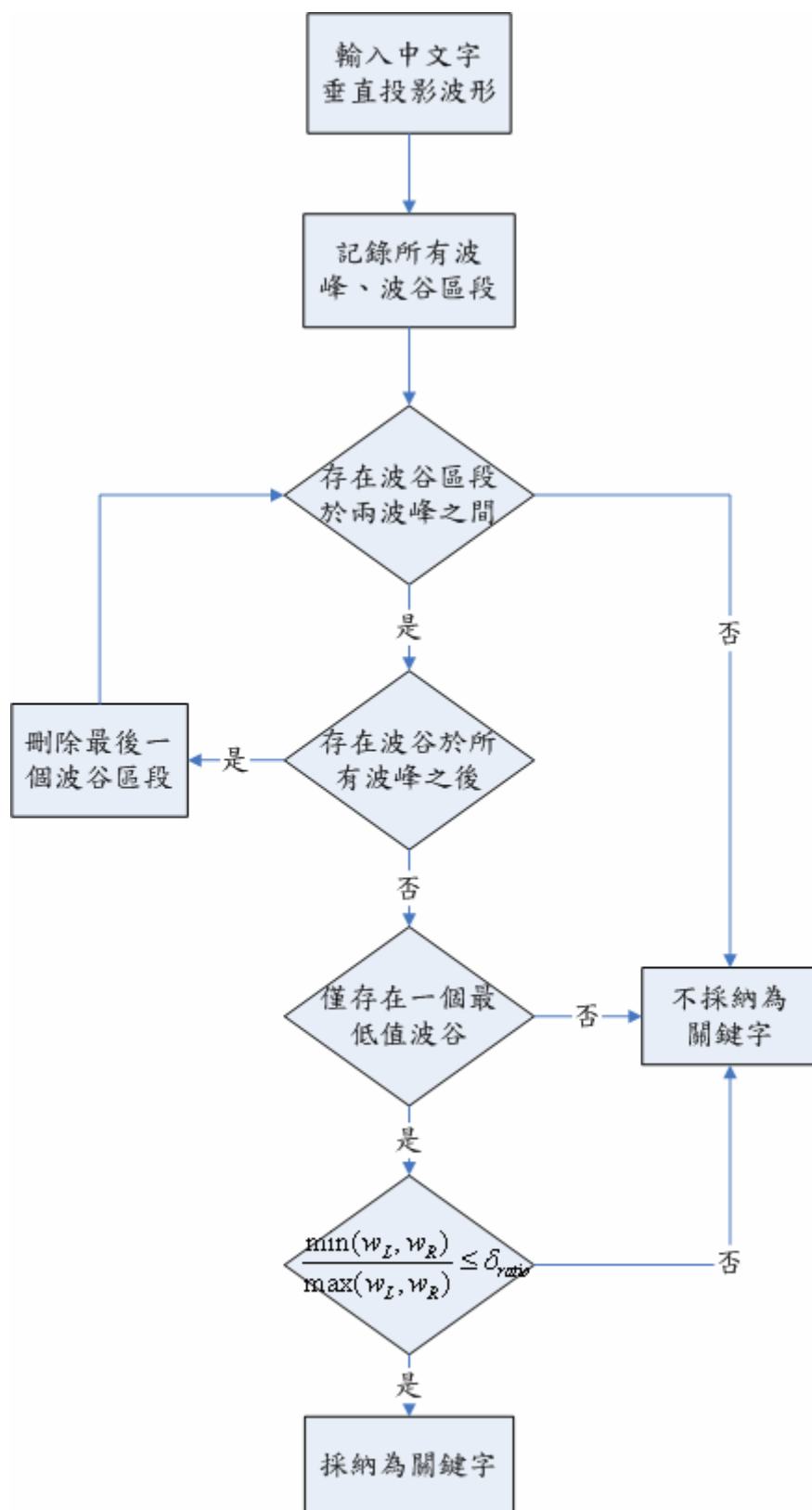


圖 3.8 關鍵字擷取流程圖

3.2.3 文件旋轉方向判讀範例

設定相機取像文件內容可能有被順時鐘旋轉 0° 、 180° 、 90° 、 270° 四種情形，本節範例可分為兩個階段，說明文件的旋轉方向，經由階段一，可化簡成 0° 、 180° 旋轉或是 90° 、 270° 旋轉兩種類型，後者可先將文件影像逆時鐘旋轉 90° ，使得後續處理時，文件影像內容僅可能需要旋轉 0° 或 180° 。將文件中擷取的文字區塊交與文字辨識模組辨識之前，需要進行第二階段文件內容旋轉方向為 0° 或是 180° 的再確認，此時中文字已經切割完成，利用中文字垂直投影量統計分析，可確認文件需要選轉的角度為 0° 或 180° 。由於相機解析度的因素，本節提供的文件方向旋轉判讀範例，是由半張 A4 中文文件所構成被旋轉 180° 的影像，影像大小為 1200×744 像素，內含 242 個中文字，16 個英數字元。

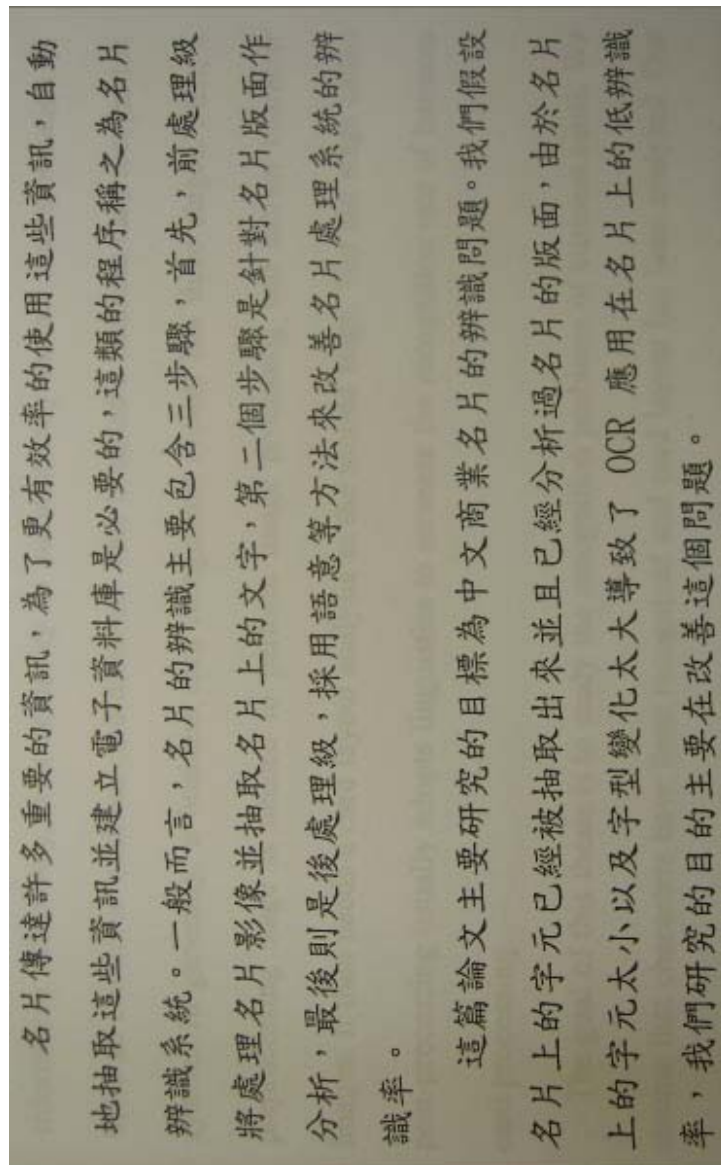


圖 3.9 旋轉 90° 的文件影像範例圖

階段一

1. 對各文字區塊中的二值化影像，使用 Sobel 水平濾波器進行遮罩運算，得到各區塊中的文字輪廓線之水平方向像素點，共有 46225 個像素點，故本範例的 SUM_H 值為 46225。

2. 對各文字區塊中的二值化影像，使用 Sobel 垂直濾波器進行遮罩運算，得到各區塊中的文字輪廓線之垂直方向像素點，共有 47239 個像素點，故本範例的 SUM_V 值為 47239。
3. 由前兩步驟得到的 SUM_H 與 SUM_V 代入 3.1-1 式，可得到文件內容需要做 0° 、 180° 旋轉或是 90° 、 270° 旋轉兩類情形的旋轉信心值，本範例中的旋轉信心值為 2.14% 略大於下限值 2%，並且 SUM_V 大於 SUM_H ，推論文件內容可能需要做 90° 、 270° 旋轉，由於下階段可判斷的文件旋轉角度為 0° 、 180° ，故於此階段此張影像需先逆時鐘旋轉 90° 。

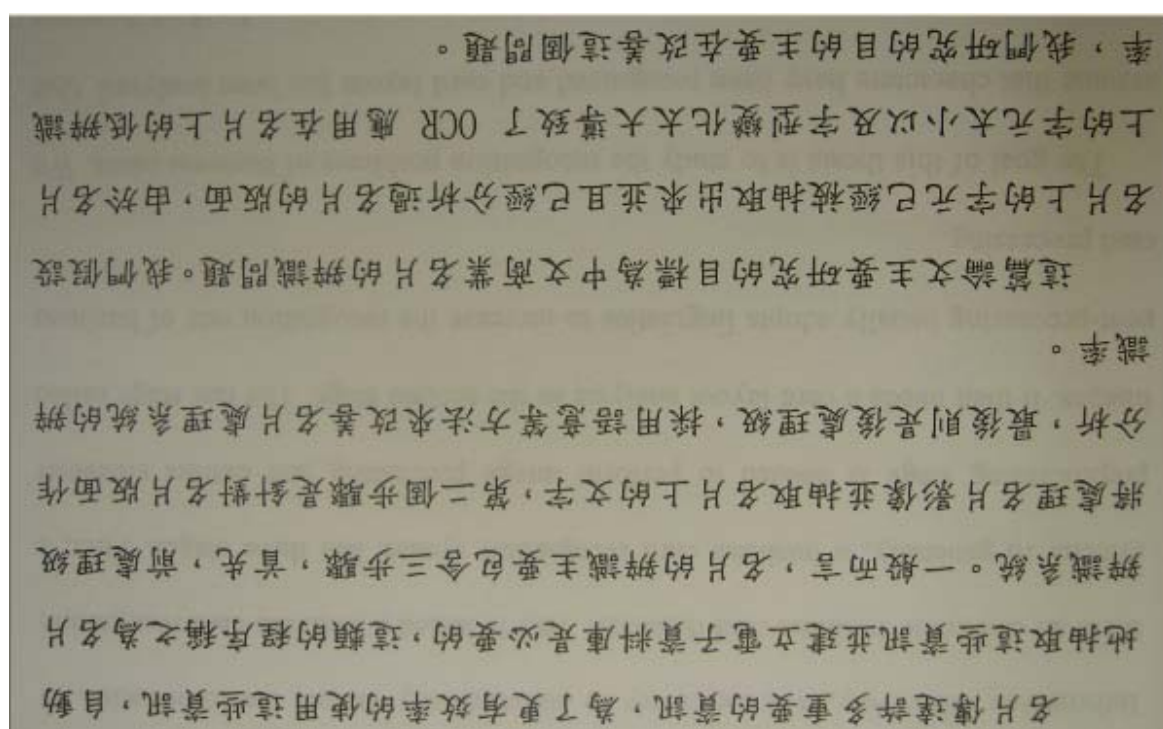


圖 3.10 階段一處理後文件影像

階段二

1. 對各文字區塊進行垂直投影量統計分析，尋找符合「明顯單一波谷類型」的文字區塊(波峰、波谷定義於 3.25 與 3.26 式)，並且滿足 3.2-7 式者視為判斷文件旋轉方向之關鍵字，有 40 個滿足條件的逆向關鍵字，3 個滿足條件的正向關鍵字，故本範例中 nctr 值為 40、pctr 值為 3。
2. 將 nctr、pctr 數值帶入 3.2-8 式，得到旋轉信心值為 92.5% 大於 δ_{rotate} 罰值 80%，則可判斷此文件影像需旋轉 180°。

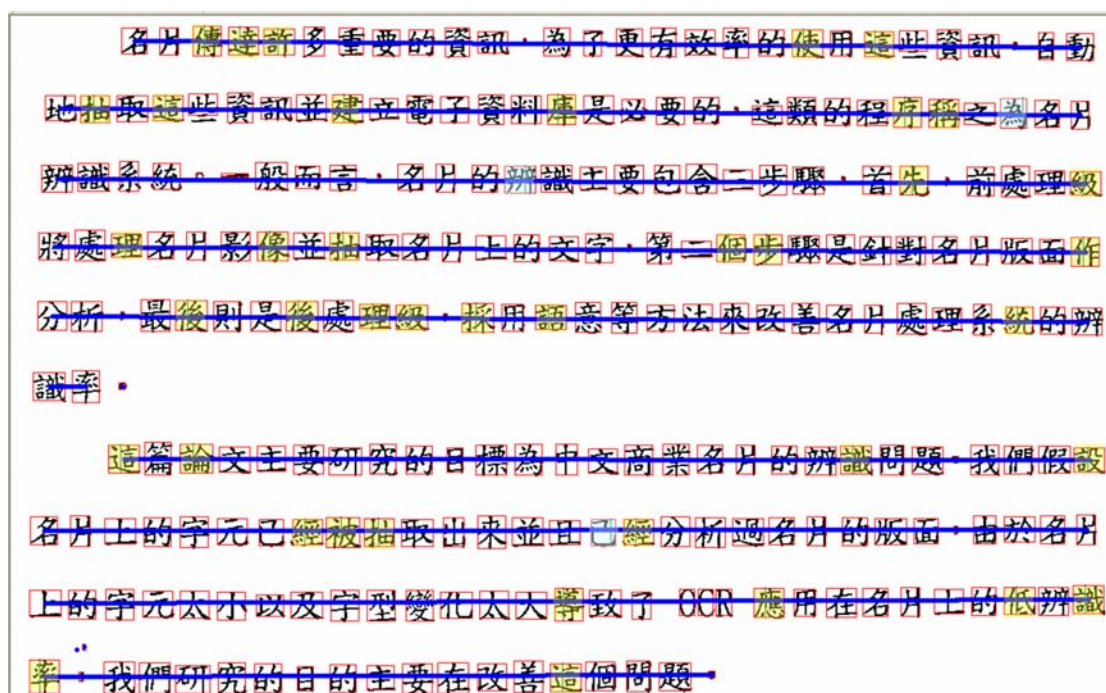


圖 3.11 判斷旋轉方向關鍵字擷取與校正範例

藍色區塊表示旋轉前的正向關鍵字

黃色區塊表示旋轉前的逆向關鍵字

第四章

文字行擷取與行中文字切割

本章主要是探討當中文文件中區域屬性(字體大小、字間距、行間距...)變異較大時，如何將文字行中的像素相連區塊串連起來並進行中文字切割。中文字切割主要採用合併的策略，先將組成中文字的像素相連區塊，合併成類似部首的部件，接著再將水平方向連續中文部件合併，並且克服中英文字區塊混雜時對文字區塊切割的影響。本章將介紹文件區域切割的方法，文字行抽取的流程，以及中文字組成元件的合併判斷法則。其中考量到文件中常有中英文字混雜的情形，使用可分辨中文字局部合併區塊與英數文字類別的 SVM 語言辨識模組，來進行部件分類，以避免英數字元在中文文件字元切割時被錯誤的合併。

4.1 文字區域群聚切割

文件上的文字資料具有群聚的現象，以名片為例。名片中屬於公司名稱的字元會聚集在一起，而屬於通訊資料屬性的字元，例如電話、傳真、郵箱、網址、地址等資料也會群聚在另一個區塊內。以文件分析的觀點討論，通常同行字間距會小於行間距、並且同屬性文字行(text line)的字體大小、行間距通常會相同，不同屬性的文字行間距較大，所以自然形成不同的區域群聚現象。文件切割方法的選擇上，使用以資料間距為主要考量的遞回垂直切割演算法(recursive X-Y cut)，可以有效的將文

件切割成數個區域。

Jaekyu 等[4]提及的遞迴水平垂直切割方法(recursive X-Y cut)，一開始需做影像像素投影，然後用投影出的局部波峰、波谷來決定何處為影像切割點，以遞迴的方式將文件影像分解成一組矩形的區塊。這種結構化的分解方式可以用 X-Y 樹(X-Y tree)這種空間資料結構，來表現文件分解的過程，樹根(root)代表整張影像，每一個節點(node)代表文件中的一個矩形範圍，父節點經由水平與垂直投影，找出兩者投影分佈圖中最大的空白間距位置，若滿足切割條件則進行水平或垂直方向切割，此時父節點可以產生上、下兩個子節點(水平方向切割)或者左右兩個子節點(垂直方向切割)，詳細 recursive X-Y cut 演算法可見 4.1.1 節。由圖 4-1 可知文件影像經由 recursive X-Y cut 演算法切割的過程與其代表的樹狀結構表示法，其中「人名」、「職稱」、「公司名」、「公司地址」、「公司的其他連絡資料」在本名片具有為獨立的屬性。

假設文件的排版並不是擁擠、複雜的情形下，將區域切割完成後，已可粗略將不同區域屬性資料(字體大小)區分開來。這樣的成果，可以解決後續整合中文字行串連演算法時，需要考慮到影像中存在多種字體大小行列的問題。受惠於區域切割已經把不同屬性的字元分群，各區域在進行整合中文字行串連時，可視為僅存在一種字體大小的情形。

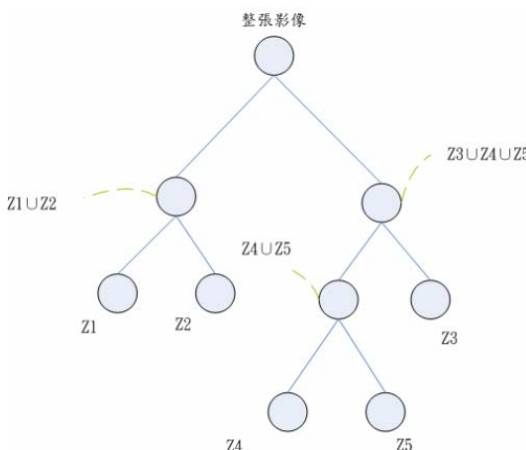
<p>楊 智 傑</p> <p>工程師 研發部 機箱產品事業處</p> <p>晟銘電子科技股份有限公司</p> <p>台北市瑞光路513巷33號2樓</p> <p>TEL: (02) 2797-3999 ext:319 FAX: (02) 2797-2876 Mobile: 0920-112398 E-mail: jj_yang@tw.uneec.com Web site: www.uneec.com 統一編號 04777186</p>	<p>楊 智 傑</p> <p>工程師 研發部 機箱產品事業處</p> <p>晟銘電子科技股份有限公司</p> <p>台北市瑞光路513巷33號2樓</p> <p>TEL: (02) 2797-3999 ext:319 FAX: (02) 2797-2876 Mobile: 0920-112398 E-mail: jj_yang@tw.uneec.com Web site: www.uneec.com 統一編號 04777186</p>
(a)	(b)
<p>楊 智 傑</p> <p>工程師 研發部 機箱產品事業處</p> <p>晟銘電子科技股份有限公司</p> <p>台北市瑞光路513巷33號2樓</p> <p>TEL: (02) 2797-3999 ext:319 FAX: (02) 2797-2876 Mobile: 0920-112398 E-mail: jj_yang@tw.uneec.com Web site: www.uneec.com 統一編號 04777186</p>	
(c)	(d)

圖 4.1 名片區塊切割及樹狀結構示意圖

4.1.1 遞迴水平垂直切割方法

傳統的投影法，是由區塊中影像像素各別進行投影，這是一種廣泛使用在由上往下的切割方法，但不幸的是在計算效能上表現不佳，如果在遞迴水平垂直切割方法中，可改用字元個數當作計算單位，將可大幅提升計算效能，故本研究使用的投影法，以相連元

件環繞方塊個數做為投影量單位，代替傳統 X-Y cut 中的像素投影。當相連單元抽取完成，各相連單元的環繞方塊也一併產生。以圖 4.2 為例，不同高度的水平投影量可表示環繞方塊出現的頻率，可避免字元影像因筆畫多寡、粗細的不同，所造成的投影量權重差異，以提升切割的正確性，並且在效能上達到很大的改進。

水平與垂直投影的方法描述如下，假設我們有一組環繞方塊 $B = \{b_1, b_2, \dots, b_N\}$ ，其中每個 b_i 環繞相連元件區域定義為 $R_{b_i} = \{(x, y) | x_{b_i \min} \leq x \leq x_{b_i \max} \text{ and } y_{b_i \min} \leq y \leq y_{b_i \max}\}$ ，則環繞方塊 b_i 的水平投影量可以用高度函式 H_{b_i} 表示，其定義如 4.1-1 式，其中 $y \in Z$ 。如圖 4.2 所示：

$$H_{b_i}(y) = \begin{cases} 1 & \text{if } y_{b_i \min} \leq y \leq y_{b_i \max} \\ 0 & \text{otherwise} \end{cases} \quad (4.1-1)$$

同樣的，環繞方塊 b_i 的垂直投影量也可以用寬度函式 V_{b_i} 表示，其定義如 4.1-2 式，其中 $x \in Z$ 。

$$V_{b_i}(x) = \begin{cases} 1 & \text{if } x_{b_i \min} \leq x \leq x_{b_i \max} \\ 0 & \text{otherwise} \end{cases} \quad (4.1-2)$$

而整組的環繞方塊 B 的水平投影或垂直投影量分佈圖，被定義為所有個別環繞方塊水平投影或垂直投影量分佈圖的總合，如式和 所定義。

$$H_b = \sum_{i=1}^N H_{b_i} \quad (4.1-3)$$

$$V_b = \sum_{i=1}^N V_{b_i} \quad (4.1-4)$$

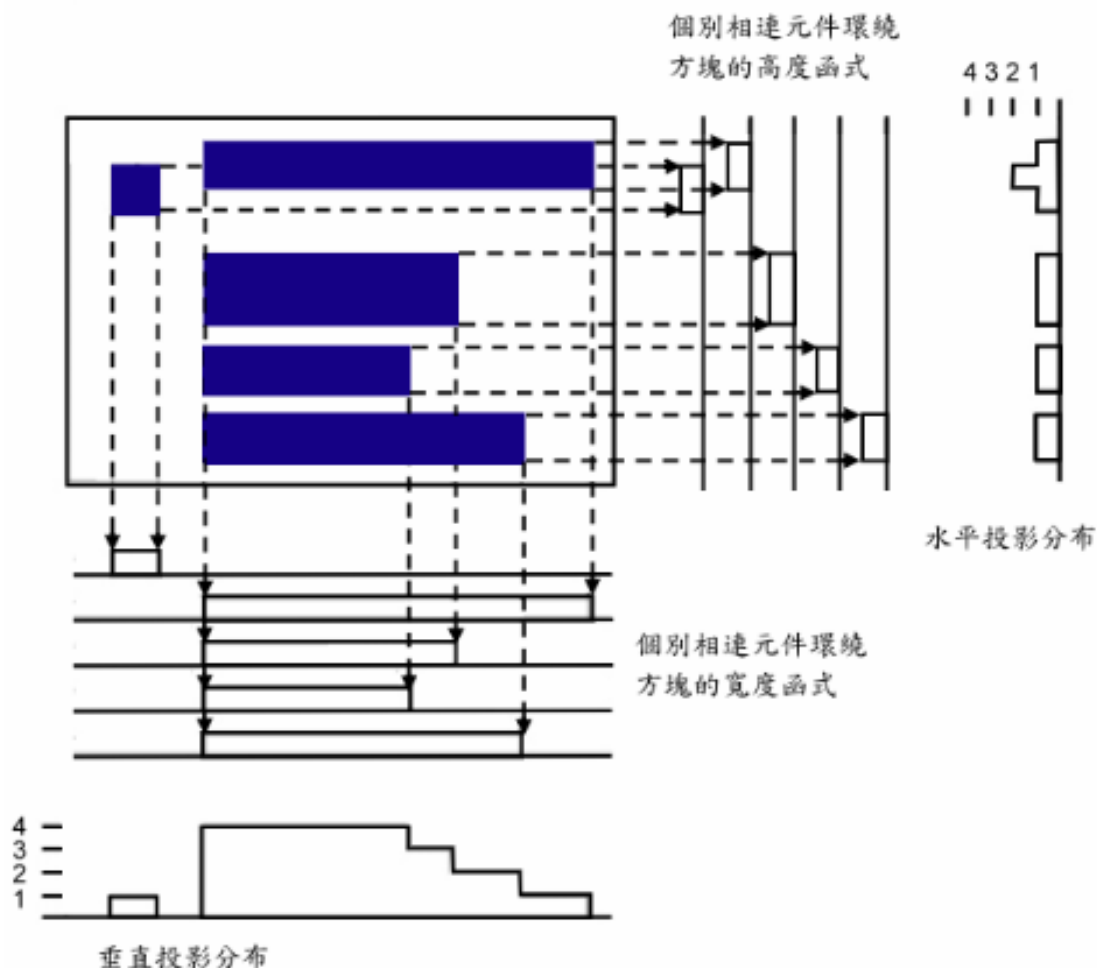


圖 4.2 區塊的水平投影及垂直投影量分佈圖

由圖 4.2 所示，整張影像(root)做完環繞方塊的水平、垂直投影後，掃描水平投影分布及垂直投影分布，找到兩者的最大間隙 gap_{Hmax} 、 gap_{Vmax} ，令最大者為 gap_{max} ，當 $gap_{max} > V_{thd}$ (最小容許切割間隙閥值)時需要進行水平或者垂直方向的影像切割。若 $gap_{Hmax} > gap_{Vmax}$ ，則由該位置將整張影像切割成上下兩張子影像；若 $gap_{Vmax} > gap_{Hmax}$ ，則由該位置將整張影像切割成左右兩張子影像(child node)。

此方法遞迴將子影像視為新的 root，來進行切割條件判斷，直到沒有任何葉節點(leaf node)滿足可分裂的條件，結束區域切割程序。

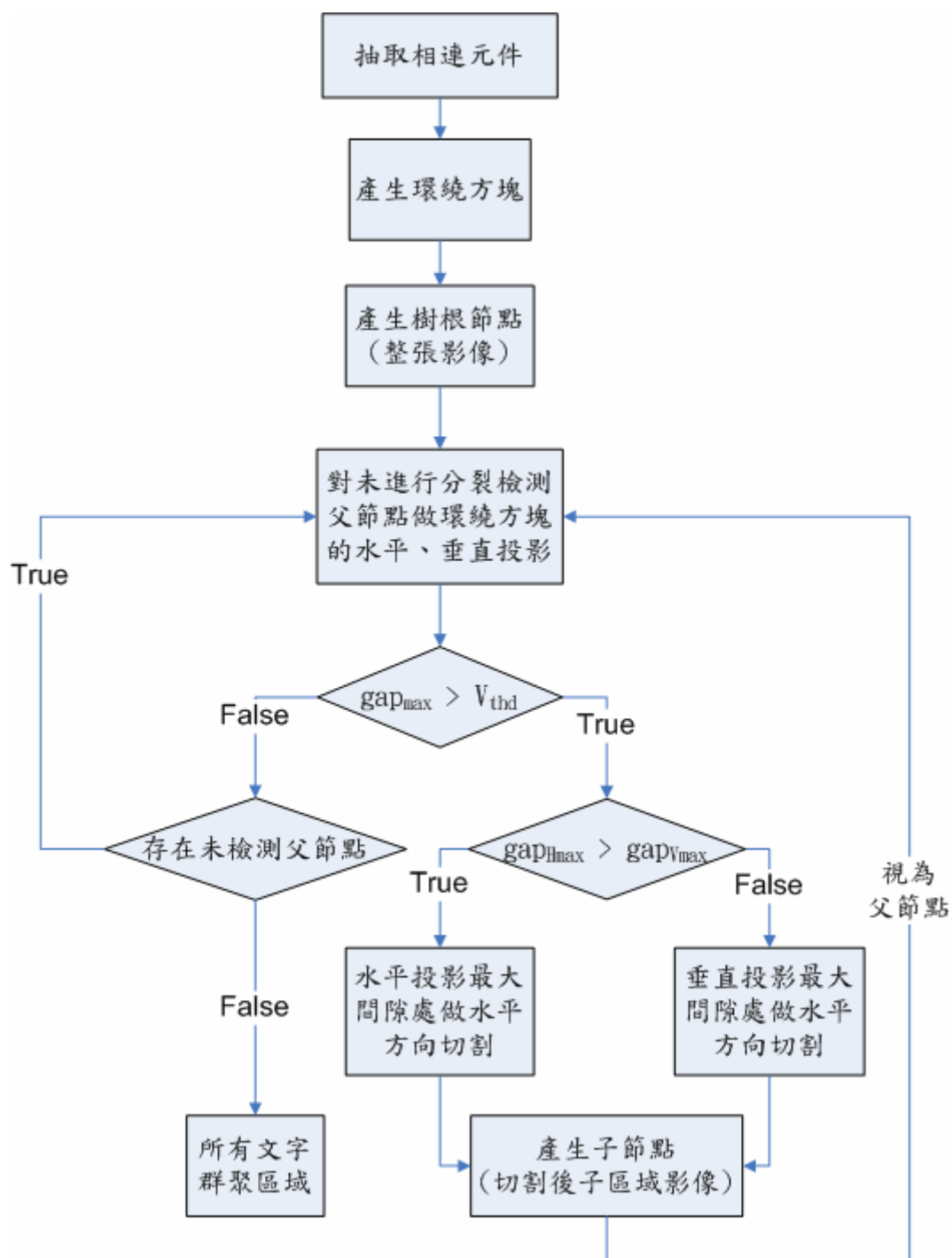


圖 4.3 遞迴水平垂直切割流程圖

4.2 文字行擷取

在印刷體文件影像中，排版類型樣式繁多，但由人眼判斷文字串列方向、讀序，卻是一件十分自然、直覺的事。由於人眼閱讀資料多以線性方向進行，將文字行、列擷取出來的意義，類似在做線狀分群，其中同群相連元件的屬性相近，可以方便簡化資料處理。本節探討相連元件之間的外緣距離，來判斷是否為同行、列的相鄰相連元件，進而將行、列中的相連元件依序串連起來。

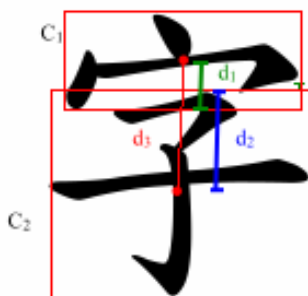


圖 4.4 外緣距離示意圖

假設有 n 個相連元件，每個相連元件(cc)皆與其他 $n-1$ 個相連元件計算距離，採用王亮盛[13]提出的環繞方塊外緣距離計算方法，如圖 4.4 所示 C_1 與 C_2 的外緣距離為中心點連線距離(d_3)扣除包含在 C_1 範圍中的線段距離(d_1)與包含在 C_2 範圍中的線段距 (d_2)，設定 $\text{distOt}(C_1, C_2) = d_3 - d_1 - d_2$ 。定義外緣距離最接近的相連元件稱為 nr_1 、第二接近的相連元件為 nr_2 ；cc 與 nr_1 形成的向量與水平線的夾角稱為 θ_1 、與 nr_2 形成的向量與水平線的夾角稱為 θ_2 ；由所有 cc 的 θ_1 估算出相連元件的主要連接角度 θ ，並以正、負 30° 做為相連元件串連的角度容許範圍，計算出串連的角

度上、下界臨界值為 σ_1 、 σ_2 ，則經由 σ_1 、 σ_2 的限制決定與 cc 串連的鄰居位置。完整的串連演算法，由 4.4 圖表示，連接步驟完成後得到的相連元件串列，僅是局部的文字行，需接續 4.3 節的內容，才能得到由完整中文字區塊所組成的文字行。

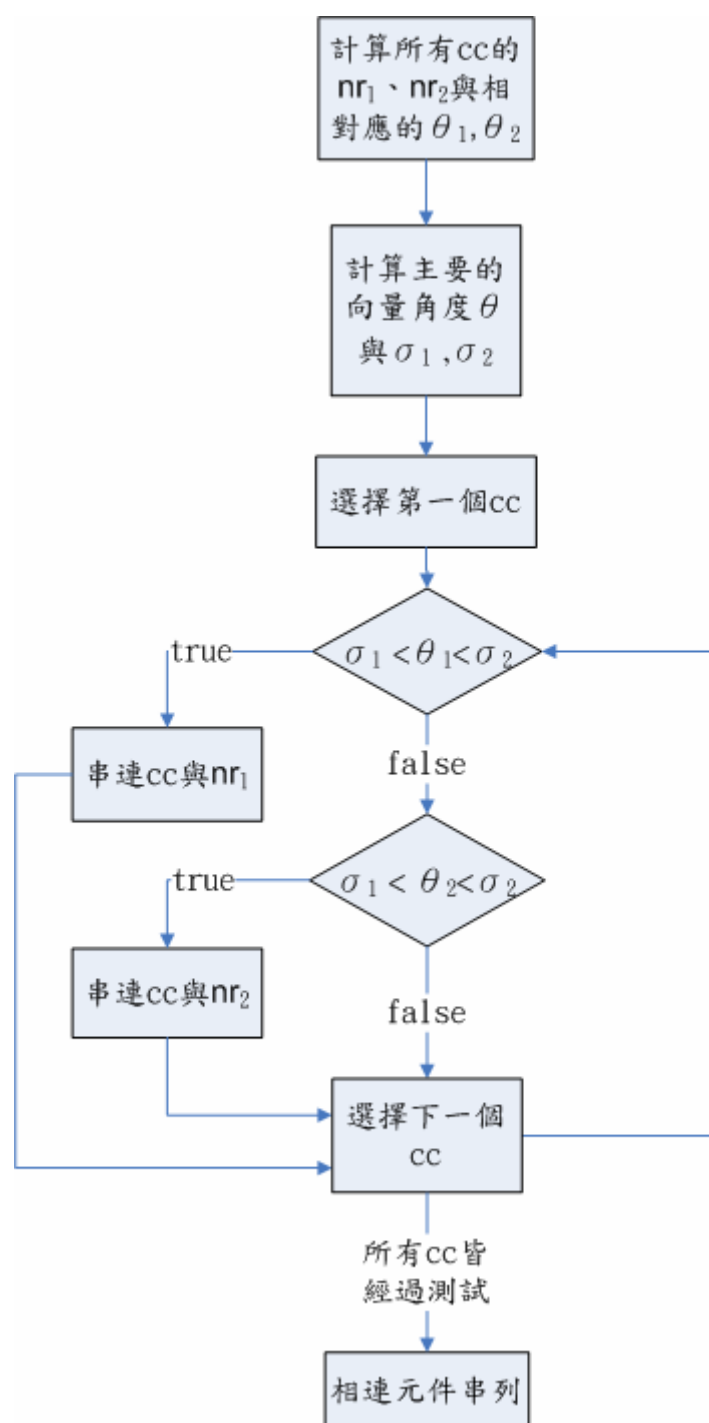


圖 4.5 相連元件串連演算法流程圖

4.3 中文文字行合併

4.2 節使用的相連元件串連演算法，較適用在像英數字元，每一個字元只包含單一相連元件的類型。套用在中文文件上時，則因中文字的組成，常由多個類似部首的相連元件構成單一中文字，造成只能得到局部的文字行。探究其主因，是之前使用的相連元件串連演算法，限定了相連元件合法的連結角度範圍為 $\sigma_1 \sim \sigma_2$ ，而中文字的特性是一個完整字可能由多個相連元件所組成，若其中某個元件與同一文字中其他相連元件連結角度皆超過限定範圍，則會形成額外的局部文字行，如圖 4.5 中「公」的「丿」、「厶」相連元件間連線角度太大，無法將之串連在同一文字行中，故容易將中文字的局部元件串連成額外的文字行，而將一文字行拆解為重疊的兩行。

從觀察中可以發現，普遍在水平走向的行中，上下分離的中文部份元件，無法串連在同一條文字行中。由於取得的文字行過於破碎重覆，將影響後續的資料擷取，故有其必要對文字行做額外的合併處理。

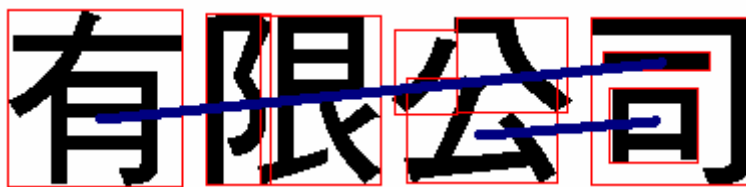


圖 4.6 相連元件串連演算法適用於中文文件的問題

4.3.1 重疊相連元件行偵測

4.1 節的文件區塊切割步驟，將影像分割為數個區域，可視為對相連元件行做初步的分類。由圖 4.1 所示，並排的相連元件行(由未合併中文字部件組成的串列)區塊已被切割開來，此時在同區塊中的相連元件行(必為水平走向)，可以用內部相連元件的 y 座標值範圍進行重疊偵測，重疊的相連元件行將被合併為同一個文字行類別。初始時每個相連元件行，皆為獨立的文字行類別，檢查任兩個相連元件行的 y 值範圍若為重疊，則合併為同一個文字行類別。合併完成後的每一個文字行類別，皆可能合併成同一文字行。

4.3.2 傾斜文件的異常合併檢查

為避免因文件傾斜導致相鄰兩行 y 座標值範圍重疊，而造成合併成同一個文字行類別的錯誤情形發生，本節中進行傾斜文件的異常合併檢查機制。

假設沒有人為刻意製造的傾斜情形，僅考慮一般正常輸入影像，傾斜狀況限制在 $+30^\circ$ 至 -30° 之間，可減少異常偵測執行的頻率。觀察組成中文字的相連元件，有某個比例是具有整個字的字高或整個字的字寬，所以每個行類別 C_i 計算代表的字寬(W_i)、字高(H_i)，先以平均值的作為篩選，減少參考部件寬高造成的誤差，高於平均值以上者再進行眾數計算，取得該行類別中代表的字寬(W_i)、代表的字高(H_i)。由於中文字寬高接近，故取 W_i 與 H_i 較大者記錄 H_{\max} 值。假設原本共有 N 個行類別，在異常合併偵測過程中，分裂出的行類別個

數以參數 t 記錄。

本演算法以幾何的概念，判斷一個文字行類別是否具有異常的文字行高度，其中使用文字行類別中較長的局部文字行斜率來估算傾斜角，視為 C_i 的傾斜角(Θ_i)，以文字行類別中所有相連元件的 x 座標範圍($X_1 \sim X_2$)長度，視為合併完文字行的寬度($X_2 - X_1$)，則可估算寬度為($X_2 - X_1$)的文字行，在傾斜 Θ_i 角度的情形下，合理的高度範圍為($X_2 - X_1$) * $\tan(\Theta) + H_{\max}$ (如圖 4.7 所示)，考慮得到的文字高度資訊(H_{\max})是約略值，設定異常高度的下限為($X_2 - X_1$) * $\tan(\Theta) + H_{\max} * \delta$ ，其中 $1 \leq \delta \leq 1.5$ ，當文字行類別高度($Y_2 - Y_1$)大於異常偵測下限，需進行線性分割分裂判斷；若($Y_2 - Y_1$)低於異常偵測下限時，則不進行分裂判斷。

傾斜文件的異常合併偵測演算法:

Step1、 對每一個文字行類別 C_i

Step2、 對 C_i 進行高度範圍檢測，若 $Split(i) = true$ ，則該群需進行分裂判斷，若否回到步驟一。

Step3、 利用線性分割，檢測 C_i 是否可分裂成兩個文字行類別，可以則分裂，回到步驟一，若不能分裂並且已不存在任何可以分割的文字行類別，結束。

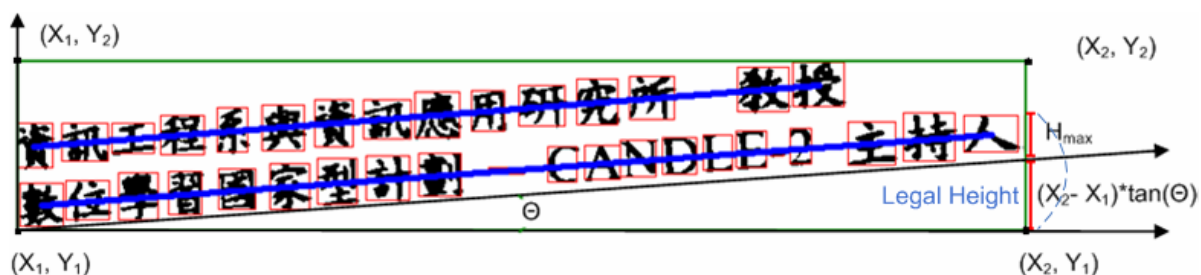


圖 4.7 字高異常判斷示意圖

其中高度範圍檢測使用 Split 函式(如 4.3-1 式)判斷行類別區要分裂與否，如圖 4.6 展示，原理是以估計一條傾斜 Θ_i 度的相連元件行合法的高度範圍，來檢測 C_i 中是否存在其他平行的相連元件行。

$$Split(i) = \begin{cases} true & \text{if } (X_2 - X_1) * \tan(\Theta) + H_{\max} * \delta < (Y_2 - Y_1) \\ false & \text{if } (X_2 - X_1) * \tan(\Theta) + H_{\max} * \delta \geq (Y_2 - Y_1) \end{cases} \quad (4.3-1)$$

對需要分裂的 C_i ，所需做的線性檢測內容如下，測試所有相連元件個數夠多(超過 5 個)的局部文字行 T_j ，紀錄所有相連元件的上緣點 u_k 、下緣點 d_k (文字行為水平方向時)，篩選去除寬與高的差異超過 0.5 倍寬或高的相連元件(其頂點可能造成判斷誤差)，再使用最小平方誤差(least square error)演算法，找出上緣線性方程式 Eq_u 、下緣線性方程式 Eq_d 。若 Eq_u 或 Eq_d 可將 C_i ，分割為 T_j 及 $\{C_i - T_j\}$ ，設定 $C_{N+t+1} = \{C_i - T_j\}$ ， $C_i = \{T_j\}$ ，分裂出的行類別個數增加 1。

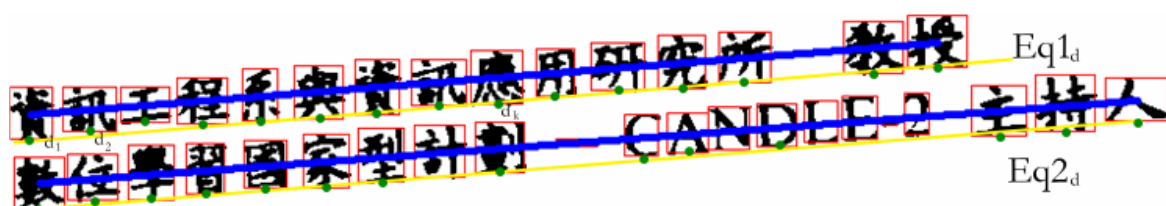
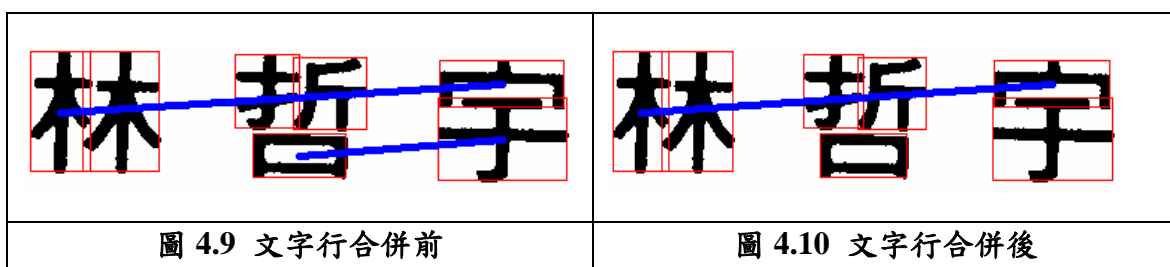


圖 4.8 文字行類別分裂示意圖

4.3.3 文字行合併

經由 4.2.1 節得到所有的文字行類別，高度區間相互重疊的相連元件行列已分在同一類，再者使用 4.2.2 節的異常合併檢查，可將因影像傾斜造成多行合併於同行列的風險去除，此時文字行類別中的局部文字行共有性質是：

1. 局部文字行高度區間相互重疊。
2. 不存在任一條線性方程式，可將兩個局部文字行分隔於 R^2 空間兩側。



假設判斷合併條件的兩個文字行代號為 T_p 和 T_q ， T_p 中的任一相連元件代號 CC_p 、 T_q 中任一相連元件代號 CC_q ，若存在 CC_p 與 CC_q 滿足下列二條件之一，則 T_p 和 T_q 兩個文字行允許合併，其中 dist 函式表示 CC_p 與 CC_q 兩區塊中心點間的距離， distOt 函式則表示兩區塊的外緣距離。

$$1. \text{distOt}(CC_p, CC_q) \leq \rho_1 \quad (4.3-2)$$

$$2. \text{dist}(CC_p, CC_q) \leq \rho_2 \quad (4.3-3)$$

考量兩個文字行中相連元件之間的距離關係，歸納同行中合理的字間距上限，應與字體大小成正比，設定距離參數 $\rho_1 = H_i/7$ ， $\rho_2 = H_i \times 0.6$ 。

4.4 相連元件語言辨識

由文件影像擷取出的文字區塊，可能為英文、數學符號、亦或中文字局部區塊，為使中文字的組成元件重整為完整的區塊，在進行中文字區塊合併之前，必須先行判斷單一相連元件的語言屬性，即其屬於中文或英文。為便於後續中文字區塊合併演算法使用，設計將文字區塊分為三類，第一類：英數、符號字元，第二類中文字部件區塊，及第三類混淆字區塊，其中第三類區塊，由於語言符號特性，無法判定其中、英文語言屬性，(如表 4.10 所示)，接著以 SVM 分類器進行文字區塊分類訓練。因後續進行中文字合併程序時，僅知道文件影像方向為垂直或水平，並已旋轉校正為水平文件，但仍無法判斷其走向是否反向(180°)旋轉，故進行 SVM 訓練時，訓練樣本必須包括三類文字區塊的正向影像，及 180° 旋轉的影像。

表 4.10 混淆字集範例

中文字部件	英文字元或符號
二	= 、 :
十	+
\	\
丿	/
丿	j
E	E
	1 、 l

4.4.1 特徵擷取

中文部件與英文文字構造屬性主要有兩項差異，第一、由曲線筆畫構成的英數文字較多，而鮮少有由曲線筆畫構成的中文字部件；第二、中文字部件中的水平、垂直筆劃較多，並且筆劃密度、複雜度都較英數字元來的高。為了能有效區分中文字部件與英數字元以上兩種筆劃組成的差異，並且克服不同字體的變異性，我們採用 F. Chang 等[1]提出使用於手寫字辨識的特徵擷取方法，應用於部件語言識別的特徵設計。此方法擷取文字輪廓像素點上的方向特徵，恰好適合描述不同語言組成文字

的筆劃特性的差異，包含三個步驟：正規化(normalization)、方向特徵擷取(directional feature extraction)、特徵模糊(feature blurring)。

第一步需要將文字區塊影像正規化成 64x64 的影像大小，接下來正規化後影像中的每一個點，都要設定方向屬性，演算法如下：

1. 擷取文字區塊輪廓(contour)，使用輪廓追蹤演算法(contour tracing algorithm)。
2. 假設 A 為輪廓邊界上的一個點，B 為他的鄰居且也是一個輪廓點，則 A 與 B 的位置相對關係會分別設定 A 與 B 一個方向屬性值，如圖 所示，若 B 在 A 水平方向鄰居位置，A 與 B 皆會獲得一個方向屬性值 1，由於輪廓點只能有兩個鄰居點，所以一個輪廓點至多可獲得兩個方向屬性值。

2	0	3
1	A	1
3	0	2

圖 4.11 方向屬性設定圖

3. 若 P 點非輪廓點，但 p 的鄰居點 A、B 皆為輪廓點，如圖 (a)情形，P 的方向值屬性為{0、1、2}、(b)情形，P 的方向值屬性為{0、1、3}，為了不遺漏 P 可能的方向屬性，將分類模糊的情形都考慮進來。

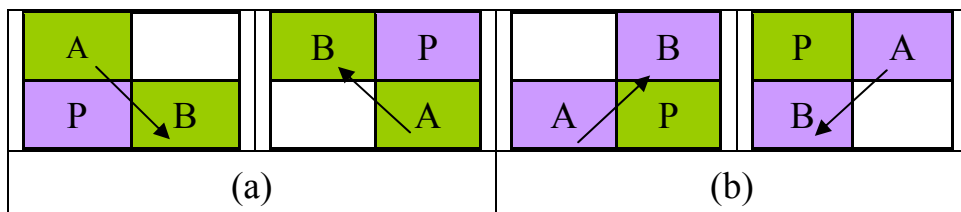


圖 4.12 方向屬性設定圖

經由方向屬性擷取演算法的運算，影像中各點的方向屬性，可由特徵影像 F_i 呈現出來，其中 $i=0,1,2,3$ 。

特徵影像 F_i 需經由特徵模糊的技巧，使得進行文字區塊特徵比對時，可以吸收局部位移造成的比對誤差，我們採用 4×4 的高斯遮罩(Gaussian)作為低通濾波器(如圖 4.11)。在採用模糊遮罩前需要先進行資料簡化的動作，將 4×4 區塊簡化成一個點，其值為 4×4 區塊中 1 的個數，產生特徵影像 G_i ，並在其上、下、左、右各加上一列的 0，產生擴增特徵影像 H_i ，並對 H_i 中每個左上點 (x, y) 都是偶數的 4×4 區塊套用模糊遮罩，此時可由 H_i 運算出 64 個值，全部總共 256 個值，則每個文字區塊影像可以得到 256 維的特徵向量，其中 $\alpha=0.1444$ ， $\beta=0.0456$ ， $\gamma=0.0144$ 。

γ	β	β	γ
β	α	α	β
β	α	α	β
γ	β	β	γ

圖 4.13 方向屬性設定圖

4.4.2 文字區塊分類器

本節採用 SVM(Support Vector Machine)做為文字區塊屬性的分類器，基於 SVM 在解決小樣本、非線性及高維模式識別問題中表現出許多特有的優勢，除了良好的學習能力，從有限訓練樣本得到的決

策規則，對獨立的測試樣本及仍能表現良好的分類效果。

用 SVM 解決兩群的分群問題，簡而言之是要找出一個超平面 (hyperplane)，超平面意指在高維中的平面，使兩個不同集合分開，因實際資料可能是屬於高維度的資料。

以一個二維的例子來說明，假設我們有訓練的向量資料 x_i ，

$i=1, \dots, k$ 。定義向量 $y_i = \begin{cases} 1 & \text{if } x_i \text{ in class 1} \\ -1 & \text{if } x_i \text{ in class 2} \end{cases}$ ，我們希望找一條直線

$f(x) = w^T x - b$ ，其中 w 為超平片的法向量， b 為超平面的截距，使所有 $y_i = -1$ 的點落在 $f(x) < 0$ 的這一邊，因此依據 $f(x)$ 的正負號可以區分這個點是屬於哪一個集合，這條線離這兩個集合的邊界 (margin) 越遠越好，而距離兩邊邊界最大的就稱為 optimal separating hyperplane (OSH) [14]。

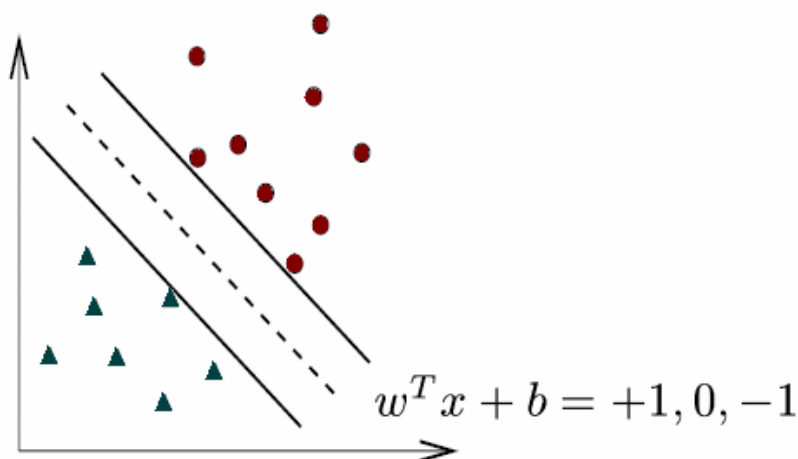


圖 4.14 SVM 用於二維資料分群示意圖

但是當我們處理非線性可分割的資料時，解決方法是把資料投射到更高維度的特徵空間去， ϕ 是映射函數，我們透過 ϕ 把資料映射到特徵空間(如 4.3-5 式)，尋找一個超平面可將資料分成兩群，在本研究中 SVM 的映射函數採用 radial based function[1]，則

$$\phi(x_i)^T \phi(x_j) = \exp\left[\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right], \text{ 決策函數為}$$

$D(x) = w^T \phi(x) - b = \sum_{i=1}^k \alpha_i y_i \phi(x_i)^T \phi(x_i) + b$ ，可以用 $\text{sgn}(D(x))$ 來判斷二值化分群。當實際分群數大於二值化時，可將 SVM 這個二值化分類器擴充到多元分類，對於 k 類分群使用 k 個 SVM，第 m 個 SVM 會將第 m 類與其他 k-1 類分開。

$$x_i^T x_j \rightarrow \phi(x_i)^T \phi(x_j) \quad (4.3-5)$$

4.5 中文字組成元件合併

中文字在尋找完相連元件後，多半有破碎成類似字根部件的問題，本節的目的在整合破碎的中文字，並且克服文字行中有中英字混雜的情形，依然能有效的將破碎的中文字合併起來。本節提出的演算法主要有兩個回合，第一回合進行垂直方向的破碎組成元件整併，產生垂直方向合併完成具有字根特性的局部元件，稱之為「部件」。第二回合則以水平方向掃描，將屬於相同文字區塊的部件合併。此演算法，僅考慮在同一條文字行中，影像範圍以 R^2 空間表示。

中文字組成元件合併演算法:

步驟一、 在同一文字行中，任兩相連元件為上下關係(x 值域範圍重疊)則合併，產生部件。

步驟二、 將部件或者英數文字區塊影像，以 SVM 的語言模組辨識出中文字部件、英數字元區塊或中英混淆字元區塊，分別設為 $type_1$ 、 $type_2$ 、 $type_3$ 。

步驟三、 水平方向尋找連續之中文部件做合併判斷。

在水平方向尋找連續部件做合併判斷時，考慮 $type_1$ 、 $type_3$ 兩種類型，可排除 $type_2$ 區塊在字間距不足時被錯誤合併的情形。假設某文字行以 T_i 表示，水平方向排序後的部件區塊以 b_k 表示， k 表示該區塊在 T_i 中水平排序順位， $type_c$ 代表 $type_1$ 聯集 $type_3$ ， w 表示該行中文字寬。合併法則描述如下：

- 若存在連續三個部件區塊 b_k 、 b_{k+1} 、 b_{k+2} 屬於 $type_c$ ，並且
$$dist(b_k, b_{k+1}) \leq w \text{ 、 } dist(b_{k+1}, b_{k+2}) \leq w \text{ 與 } dist(b_k, b_{k+2}) \geq w$$
同時發生，可以外緣距離決定 b_{k+1} 的合併方向。若 $distOt(b_k, b_{k+1}) \leq distOt(b_{k+1}, b_{k+2})$ 發生，則 b_k 、 b_{k+1} 合併；若否則 b_{k+1} 、 b_{k+2} 合併
- 若存在 $B = \{b_k \mid t \leq k \leq s \text{ and } b_k \in type_c\}$ 並且 $dist(b_t, b_s) \leq w$ ，將 B 集合中的區塊合併。

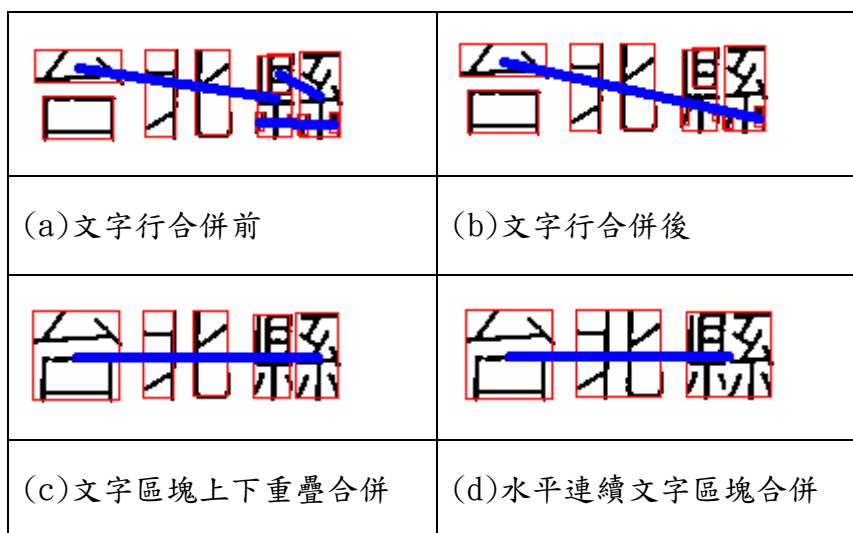


圖 4.15 中文字合併流程示意圖

第五章

實驗結果

本章將針對中文文件中的名片進行中文文件辨識前處理系統效能測試，採用名片影像進行實驗的主因，是名片的排版較複雜文字大小的變化情形也較多，可用來證明中文字合併效能的穩定度。另外文件方向判別，則以中文字庫來測試以關鍵字判斷文件正向或顛倒方向的穩定度。

5.1 中文字合併效能評估

中文字合併測試，使用的硬體環境為解析度 1600*1200，自動對焦的網路攝影機。拍攝的 55 張中文名片做為測試資料集，測試名片中的中文與中英字元混合欄位，共 5099 個中文或英數字元，其中第 51 至第 55 章為具有反白字區域的名片。中文字合併效能評估共分三個部份，(a)無局部區塊語言辨識的中文字合併效能評估，(b)無反白字區域偵測的中文字合併效能評估，(c)完整流程的中文字合併效能評估。以此三個實驗來觀察，反白字區域偵測模組與局部文字區塊語言辨識模組，對整個系統中文字擷取能力的影響。

在評估(a)當中，由於未使用局部區塊語言辨識模組，共 5099 個中文或英數字元中，合併錯誤高達 713 個區塊，由於無法將中文字之外

的區塊排除於合併條件之外，較常出現的錯誤是將間距較小的英數字元合併，錯誤發生的比率與名片使用的文字間距大小有密切關係，正確率僅有 85%。

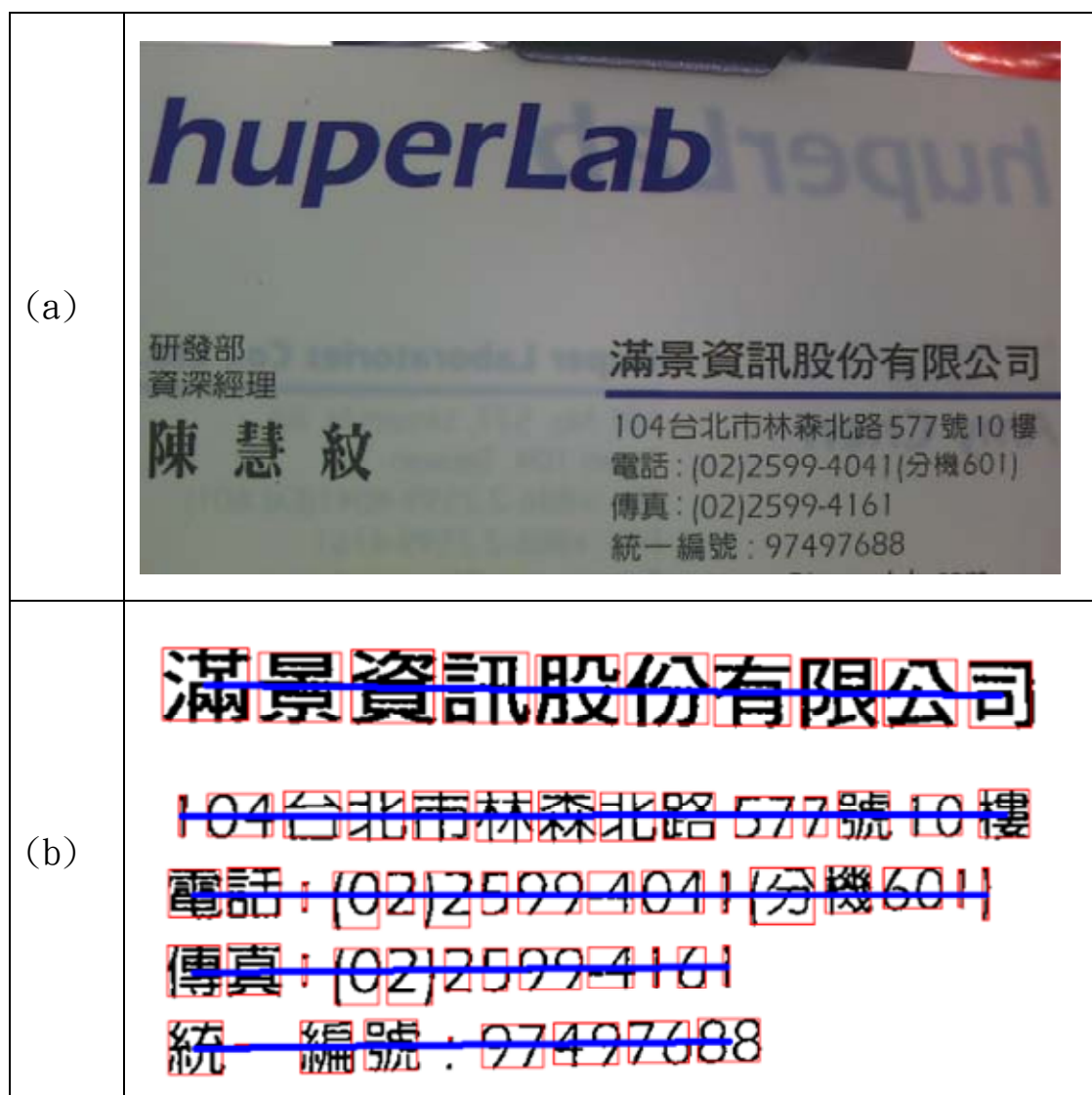


圖 5.1 無局部區塊語言辨識的中文字合併範例

評估(b)僅對第 51 至 55 張存在反白字的區塊的名片作實驗，反白字區塊通常於相連元件抽取完後，會因大小異常而被過濾掉，所以區塊內的文字字訊會完全喪失，以此五張為例，578 個中文或英數字元中，就喪失了 77 個反白文字，正確率僅接近 86%。

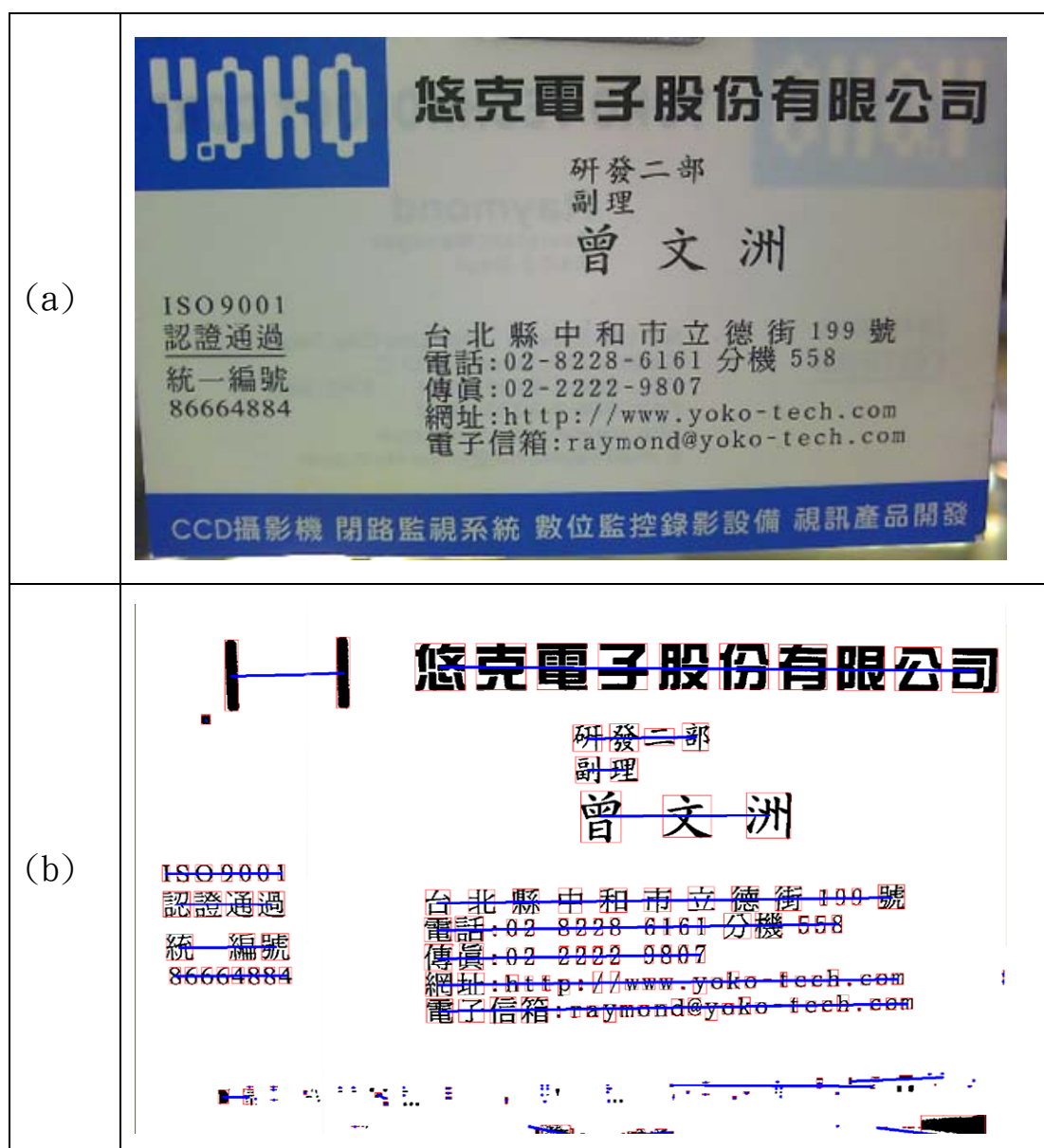


圖 5.2 無反白字區與偵測的中文字合併範例

評估(c)使用完整流程，包括局部區塊語言辨識模組，與反白字區域的偵測。在 55 張中文名片中，未合併的中文局部區塊共 14 個，中文局部區塊與英數錯誤合併共 68 個，正確率高達 98.39%，少數錯誤合併的原因，可能是受到二值化影響，文字區塊喪失文字特徵，亦或是較特殊字體，使得語言辨識模組判斷出錯誤的語言分類，導致合併錯誤。


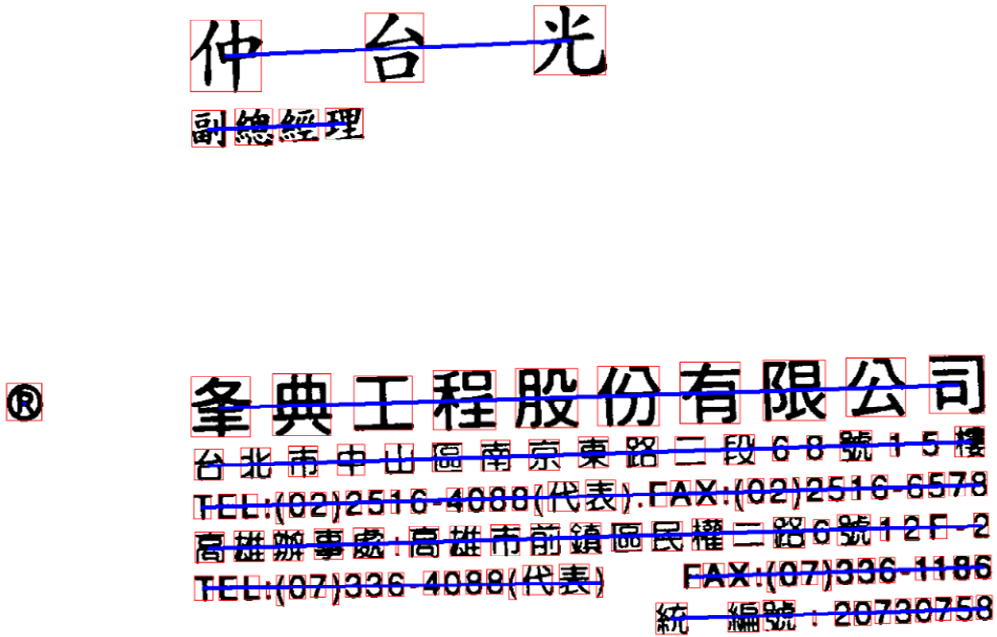
(a)	 <p>仲 台 光 副總經理</p> <p>® 峯典工程股份有限公司 台北市中山區南京東路三段68號15樓 TEL:(02)2516-4088(代表).FAX:(02)2516-6578 高雄辦事處:高雄市前鎮區民權二路6號12F-2 TEL:(07)336-4088(代表) FAX:(07)336-1186 統一編號: 20730758</p>
(b)	 <p>仲 台 光 副總經理</p> <p>® 峯典工程股份有限公司 台北市中山區南京東路三段68號15樓 TEL:(02)2516-4088(代表).FAX:(02)2516-6578 高雄辦事處:高雄市前鎮區民權二路6號12F-2 TEL:(07)336-4088(代表) FAX:(07)336-1186 統一編號: 20730758</p>

圖 5.3 傾斜文件使用完整流程的中文字合併範例

表 5.1 中文字合併效能評估(a)無局部區塊語言辨識表

名片編號	局部區塊未合併	中文區塊合併錯誤	中英文區塊總數	正確率
1	0	13	84	100.00%
2	0	19	102	100.00%
3	0	14	88	100.00%
4	0	11	80	97.50%
5	0	23	135	100.00%
6	0	15	107	100.00%
7	0	10	102	100.00%
8	0	12	64	100.00%
9	0	15	117	100.00%
10	0	22	128	100.00%
11	0	13	110	100.00%
12	0	12	76	100.00%
13	0	19	110	99.09%
14	0	19	100	100.00%
15	0	2	72	98.61%
16	0	18	99	97.98%
17	0	15	83	96.39%
18	0	18	101	99.01%
19	0	20	129	97.67%
20	0	13	115	100.00%
21	0	2	81	100.00%
22	0	4	65	100.00%
23	0	12	66	93.94%
24	0	30	108	98.15%
25	0	13	121	100.00%
26	0	24	108	99.07%
27	0	14	120	100.00%
28	0	19	98	97.96%
29	0	17	100	97.00%
30	0	16	105	96.19%
31	0	10	72	86.11%
32	0	9	48	81.25%

33	0	10	51	80.39%
34	0	32	75	57.33%
35	0	2	96	97.92%
36	0	3	78	96.15%
37	0	1	58	98.28%
38	0	12	84	86.00%
39	0	6	45	86.00%
40	0	17	118	86.00%
41	0	13	91	86.00%
42	0	7	50	86.00%
43	0	12	83	86.00%
44	0	11	112	90.46%
45	0	8	58	86.00%
46	0	6	114	94.77%
47	0	14	101	86.00%
48	0	6	78	92.41%
49	0	6	76	92.58%
50	0	8	59	86.00%
51	0	12	89	97.75%
52	0	7	121	94.21%
53	0	17	120	100.00%
54	0	14	132	100.00%
55	0	16	116	100.00%
Total	0	713	5099	86.03%

表 5.2 中文字合併效能評估(b)無反白字區域偵測

名片編號	局部區塊未合併	中文區塊合併錯誤	無反白字錯誤	中英文區塊總數	正確率
51	2	0	14	89	82.02%
52	3	4	15	121	81.82%
53	0	0	26	120	78.33%
54	0	0	10	132	92.42%
55	0	0	12	116	89.66%
Total	0	4	77	578	85.99%

表 5.3 中文字合併效能評估表(c)完整流程

名片編號	局部區塊未合併	中文區塊合併錯誤	中英文區塊總數	正確率
1	0	0	84	100.00%
2	0	0	102	100.00%
3	0	0	88	100.00%
4	0	2	80	97.50%
5	0	0	135	100.00%
6	0	0	107	100.00%
7	0	0	102	100.00%
8	0	0	64	100.00%
9	0	0	117	100.00%
10	0	0	128	100.00%
11	0	0	110	100.00%
12	0	0	76	100.00%
13	0	1	110	99.09%
14	0	0	100	100.00%
15	0	1	72	98.61%
16	0	2	99	97.98%
17	2	1	83	96.39%
18	0	1	101	99.01%
19	0	3	129	97.67%
20	0	0	115	100.00%
21	0	0	81	100.00%
22	0	0	65	100.00%
23	2	2	66	93.94%
24	0	2	108	98.15%
25	0	0	121	100.00%
26	0	1	108	99.07%
27	0	0	120	100.00%
28	0	2	98	97.96%
29	0	3	100	97.00%
30	0	4	105	96.19%
31	0	0	72	100.00%
32	0	0	48	100.00%
33	0	0	51	100.00%

34	0	6	75	92.00%
35	0	1	96	98.96%
36	0	3	78	96.15%
37	0	1	58	98.28%
38	0	0	84	100.00%
39	0	0	45	100.00%
40	1	1	118	98.31%
41	0	8	91	91.21%
42	0	0	50	100.00%
43	0	6	83	92.77%
44	1	5	112	94.64%
45	1	5	58	89.66%
46	0	0	114	100.00%
47	1	3	101	96.04%
48	0	0	78	100.00%
49	0	0	76	100.00%
50	1	0	59	98.31%
51	2	0	89	97.75%
52	3	4	121	94.21%
53	0	0	120	100.00%
54	0	0	132	100.00%
55	0	0	116	100.00%
Total	14	68	5099	98.39%

5.2 方向判別關鍵字效能評估

為檢測本研究中的文件方向判別效能，並為使測試樣本具通用性，本測試的中文字集採用 Big5 中文字庫共 5401 個中文字，針對標楷體與新細明體兩種常用字體，進行篩選關鍵字的效能評估。本實驗分成兩個階段，第一個階段，評估 Big5 中文字庫擷取出的關鍵字，由左、右側偏旁組成的比例。第二個階段，考慮中文字個別出現於一般文件

的機率不同，評估使用文字頻率加權後，正、逆向關鍵字出現的機率。

實驗評估一：

標楷體擷選出來 958 個正向關鍵字，108 個逆向關鍵字；新細明體擷選出來 700 個正向關鍵字，80 個逆向關鍵字。其中部分由左偏旁、或右偏旁構成的中文字，垂直投影完的波形，可能因為左右相連元件部份重疊過多，導致偏旁附近的波谷消失(例如：幼、致、旗)，重疊發生的情形可能因字體的影響而變異，而使得原本以部首概念，可由左右兩個區塊組成的中文字，投影後不符合「明顯單一波谷類型」，無法納入關鍵字集，輔助方向判別的中文字。

測試結果可知，標楷體中可擷取的關鍵字佔總字集的 20%，新細明體中可擷取的關鍵字佔總字集的 14%，但探討總體的結果，定義 positiveRotate 為正向關鍵字個數， negativeRotate 為逆向關鍵字個數為，文件旋轉與否的信心指數(RotateConfidence)定義於 5.2-1 式，無論標楷體或新細明體皆有接近 9 倍的數值，故以本文方法擷取關鍵字判斷文件方向仍有相當的可信度。

$$\text{RotateConfidence} = \frac{\max(\text{negativeRotate}, \text{positiveRotate})}{\min(\text{negativeRotate}, \text{positiveRotate})} \quad (5.2-1)$$

表 5.4 文件方向判斷關鍵字效能評估表

	正向關鍵字數	逆向關鍵字數	關鍵字佔總字集比重	旋轉信心指數
標楷體	958	108	20%	8.87 倍
新細明體	700	80	14%	8.75 倍

實驗評估二：

字頻資料使用國語推行委員會提供，八十七年口語問卷調查報告書所提供的字頻分析結果[20]，共 2893 個常用字。以新細明體測試，擷取出 435 個正向關鍵字與 51 個逆向關鍵字；以標楷體測試，擷取出 784 個正向關鍵字與 97 個逆向關鍵字。加上字頻做為權重統計的結果，標楷體文件預估每 100 個字可出現 25 個正向關鍵字，4 個逆向關鍵字；新細明體文件預估每 100 個字可出現 16 個正向關鍵字，3 個逆向關鍵字。

利用字頻概念去模擬一般文件中文字出現的頻率，而判斷出的正向、逆向的關鍵字出現頻率仍然有 5 倍以上的差距，由此可見，本文所提出的中文文件方向判斷方法確實有其應用價值。

表 5.5 加入字頻權重的關鍵字測試表

	正向關鍵字 出現頻率	逆向關鍵字 出現頻率	比例值
標楷體	25/100 字	4/100 字	6.25 倍
新細明體	16/100 字	3/100 字	5.30 倍

第六章

結論與未來工作

6.1 結論

本研究提出了一個實用的中文文件前處理系統。利用矩形邊緣追蹤，可判斷出常見的反白文字區域。本文提出的中文字行串連演算法，結合文件區域的文字大小屬性分析，可以克服字體大小變化對中文字合併造成的問題，在合併過程中，以 SVM 訓練模組判斷局部文字區塊的語言類別，使得中文字局部區塊不會誤與英數文字區塊合併，在中文名片的實驗測試，正確率高達 98.39%，可見中文字行抽取演算法有相當的穩定性。

本研究第二個貢獻，是提出了一個中文文件方向確認的方法，經過步驟一：垂直邊緣像素點與水平邊緣像素點的統計，與步驟二：中文字特有的垂直投影分布特性分析，可以使用中文文件內容，判斷中文字呈現的方向。由於本法是基於統計原理基礎，我們以常用字頻做關鍵字機率分析，每一百個字才會出現 29 個正、逆向關鍵字，故如果設定關鍵字判斷字數下限為 15 個字，則套用本統計方法的文件中文字數必須超過 50 個字。在中文字數樣本充足的情形下，無論是以整個中文字集討論，亦或以常用字頻的角度分析，判斷正、逆方向的關鍵字個數都達到 5 至 6 倍以上，算是相當顯著倍差，故進行文件方向判斷時，可視為一個穩定的參考指標。

6.2 未來工作

本研究提出的以相機取像的中文文件前處理系統，仍需要在良好採光、且沒有手震的情形下拍攝文件及名片影像，期望在未來研究中可以使用對於光影分布不均與模糊情形皆能處理良好的二值化演算法則，使得使用者拍攝影像時有更高的便利性；並且希望可以對非曼哈頓式的複雜排版文件進行文字內容及讀序擷取的討論，使得有更多以相機取像的文件類型可以使用本自動化的辨識前處理流程。中文文件方向判斷及校正方面，本研究提出的校正演算法是建立於統計的基礎上，故使用校正判斷的中文文件影像需要足夠數量的中文字數，雖然在中文字數較少的格式化文件影像(例如:名片)，可以用技巧性的關鍵字比對來判斷其文件的方向，但期望後續研究可以找到適用於較少中文字數文件，可穩定判斷文件走向的通用法則，使得可以使用自動化中文文件方向判斷及校正的文件類型更加多元。

參考文獻

- [1] F. Chang, C. H. Chou, C. C. Lin, and C. J. Chen “A Prototype Classification Method and its Application to Handwritten Character Recognition”. International Conference on Systems, Man and Cybernetics, 2004
- [2] Y. Cao, H. Li, "Skew Detection and Correction in Document Images Based on Straight-Line Fitting," Pattern Recognition Letters, , Vol. 24, No. 12, pp. 1871-1879, August. 2003.
- [3] R. G. Casey and E. Lecolinet, “A Survey of Methods and Strategies in Character Segmentation”, IEEE Transaction on Pattern Recognition Analysis and Machine Intelligence. Vol 18, No.7, pp 690 - 706, 1996.
- [4] J. Ha, R. M. Haralick and I. T. Phillips“Recursive X-Y Cut using Bounding boxes of Connected Components”, International Conference on Document Analysis and Recognition, Vol 2, 1995
- [5] L. Jagannathan and C.V. Jawahar, “perspective correction methods for camera-based document analysis” . International Workshop on Camera-based Document Analysis and Recognition, 2005
- [6] Y. Liu, S. Goto, T. Ikenaga. “A Robust Algorithm for Text Detection in Color Images”. International Conference on Document Analysis and Recognition, 2005
- [7] P. S. Liao, T. S. Chen, and P. C. Chung. “A Fast Algorithm for Multilevel Thresholding.” Journal of Information Science and Engineering, Vol 17, pp 713-737, 2001.
- [8] M. A. Fischler, R. C. Holles. “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography.” Comm. of the ACM, Vol 24, pp 381-395, 1981.
- [9] P. Clark, and M. Mirmehdi, “Rectifying perspective views of text in 3D scenes using vanishing points,” *Pattern Recognition*, Vol. 36, 2673-2683, 2003.

- [10] H. M. Sun , “Page Segmentation for Manhattan and Non-Manhattan Layout Documents via Selective CRLA”. International Conference on Document Analysis and Recognition, 2005
- [11] Y. Zhong, K. Karu, and A.K. Jain, "Locating Text in Complex Color Images," Pattern Recognition, Vol. 28, No. 10, pp. 1,523-1,536, Oct. 1995.
- [12] S. Zhao, Z. Chi, P. Shi and Q. Wang, “Handwritten Chinese Character Segmentation Using a Two-Stage Approach”, International Conference on Document Analysis and Recognition, 2001
- [13] 王亮聖, ”利用文件分析作文件之無失真重現”, 國立中央大學資訊工程研究所博士論文, 中華民國 86 年六月
- [14] 林宗勳, Support Vector Machines 簡介 , 台灣大學資訊工程研究所, 2000.
- [15] 林家禎, ”中英文名片商標的擷取及辨識”, 國立中央大學資訊工程研究所碩士論文, 中華民國 90 年六月
- [16] 李祐昇, “利用小波轉換自動偵測影像中的文字”, 國立台灣大學資訊管理研究所碩士論文, 中華民國 89 年六月
- [17] 孫宇、江崇禮、董明, “GIS 電子地圖中文字標注方向矯正演算法” , 大連理工大學學報, Vol. 42, 2002.
- [18] 賴逸嶺, ”中文名片處理系統”, 國立中央大學電機工程研究所碩士論文, 中華民國 87 年六月
- [19] 廖紹鋼(編譯), Gonzalez Woods(原著), ”數位影像處理”, 普林斯頓國際有限公司, 第二版
- [20] 教育部全球資訊網
- [21] <http://www.cs.mcgill.ca/~aghnei/square.html>
- [22] 維基百科 : <http://zh.wikipedia.org/w/index.php?title=%E9%83%A8%E9%A6%96&variant=zh-tw>

附錄 A

一、字頻總表

(一)本表字序依頻次高低安排，頻次表示每一百中文字，該字出現的次數。

(二)本表收字 2893 字，總頻次為 111940 次。

序號	字	頻次	序號	字	頻次	序號	字	頻次
1	人	1.332857	966	借	0.019653	1931	弦	0.003573
2	的	1.23191	967	滾	0.019653	1932	幼	0.003573
3	電	0.979989	968	園	0.019653	1933	瓶	0.003573
4	大	0.766482	969	親	0.019653	1934	飢	0.003573
5	好	0.756655	970	屬	0.01876	1935	另	0.003573
6	有	0.745042	971	盜	0.01876	1936	簽	0.003573
7	機	0.703055	972	謝	0.01876	1937	櫥	0.003573
8	車	0.656602	973	柯	0.01876	1938	駕	0.003573
9	業	0.632482	974	再	0.01876	1939	擾	0.003573
10	我	0.619975	975	摺	0.01876	1940	頤	0.003573
11	生	0.586028	976	零	0.01876	1941	頁	0.003573
12	子	0.566375	977	季	0.01876	1942	某	0.003573
13	看	0.526175	978	江	0.01876	1943	屠	0.003573
14	錢	0.511881	979	抓	0.01876	1944	辜	0.003573
15	不	0.507415	980	扮	0.01876	1945	授	0.003573
16	學	0.496695	981	碗	0.01876	1946	尺	0.003573
17	會	0.485975	982	炫	0.01876	1947	窩	0.003573
18	舞	0.485975	983	螺	0.01876	1948	待	0.003573
19	畫	0.485975	984	艦	0.01876	1949	稱	0.003573
20	保	0.482401	985	靜	0.01876	1950	婉	0.003573
21	國	0.463641	986	燒	0.01876	1951	供	0.003573
22	球	0.459175	987	查	0.01876	1952	韓	0.003573
23	打	0.451135	988	罰	0.01876	1953	葛	0.003573
24	家	0.451135	989	娃	0.01876	1954	碧	0.003573
25	很	0.451135	990	療	0.01876	1955	恥	0.003573
26	法	0.447561	991	櫃	0.01876	1956	減	0.003573
27	樂	0.434161	992	深	0.01876	1957	蛇	0.003573
28	影	0.424334	993	炒	0.01876	1958	貌	0.003573
29	多	0.421654	994	葉	0.01876	1959	禹	0.00268
30	工	0.406468	995	城	0.01876	1960	尹	0.00268
31	水	0.395748	996	莎	0.01876	1961	迂	0.00268
32	新	0.380561	997	幣	0.01876	1962	艱	0.00268
33	美	0.375201	998	裁	0.01876	1963	摘	0.00268
34	金	0.374308	999	版	0.01876	1964	肖	0.00268
35	魚	0.369841	1000	拿	0.017867	1965	姝	0.00268
36	行	0.346614	1001	腐	0.017867	1966	駭	0.00268
37	動	0.344828	1002	累	0.017867	1967	蘆	0.00268
38	音	0.341254	1003	覽	0.017867	1968	緣	0.00268

序號	字	頻次	序號	字	頻次	序號	字	頻次
39	吃	0.339468	1004	躲	0.017867	1969	與	0.00268
40	民	0.320708	1005	檔	0.017867	1970	礎	0.00268
41	政	0.318028	1006	墅	0.017867	1971	辰	0.00268
42	農	0.304627	1007	偶	0.017867	1972	佃	0.00268
43	險	0.303734	1008	吹	0.017867	1973	嚼	0.00268
44	喜	0.302841	1009	背	0.017867	1974	嶺	0.00268
45	卡	0.302841	1010	煩	0.017867	1975	穎	0.00268
46	網	0.296587	1011	擠	0.017867	1976	屈	0.00268
47	公	0.293014	1012	研	0.017867	1977	犁	0.00268
48	小	0.291227	1013	額	0.017867	1978	槓	0.00268
49	歡	0.288547	1014	修	0.017867	1979	鷺	0.00268
50	聞	0.280507	1015	債	0.017867	1980	豚	0.00268
51	事	0.278721	1016	亞	0.017867	1981	浹	0.00268
52	視	0.276041	1017	必	0.017867	1982	鮎	0.00268
53	可	0.266214	1018	細	0.017867	1983	喘	0.00268
54	地	0.266214	1019	探	0.017867	1984	儉	0.00268
55	牛	0.266214	1020	植	0.017867	1985	劈	0.00268
56	栗	0.262641	1021	婚	0.017867	1986	鏤	0.00268
57	產	0.261747	1022	蘭	0.017867	1987	傢	0.00268
58	風	0.261747	1023	求	0.017867	1988	虹	0.00268
59	股	0.258174	1024	秋	0.017867	1989		0.00268
60	歌	0.258174	1025	責	0.017867	1990	姨	0.00268
61	高	0.257281	1026	垃	0.016973	1991	迦	0.00268
62	了	0.256387	1027	檢	0.016973	1992	傭	0.00268
63	天	0.255494	1028	呆	0.016973	1993	卉	0.00268
64	化	0.250134	1029	只	0.016973	1994	忠	0.00268
65	交	0.249241	1030	抗	0.016973	1995	鶯	0.00268
66	台	0.246561	1031	宙	0.016973	1996	濃	0.00268
67	物	0.243881	1032	姦	0.016973	1997	扇	0.00268
68	術	0.242987	1033	令	0.016973	1998	朝	0.00268
69	漁	0.242094	1034	蓄	0.016973	1999	葫	0.00268
70	心	0.234947	1035	圾	0.016973	2000	迫	0.00268
71	藥	0.234054	1036	史	0.016973	2001	皆	0.00268
72	去	0.232267	1037	紀	0.016973	2002	汪	0.00268
73	員	0.231374	1038	柔	0.016973	2003	鷺	0.00268
74	上	0.231374	1039	萊	0.016973	2004	頌	0.00268
75	路	0.230481	1040	宣	0.016973	2005	敞	0.00268
76	演	0.229587	1041	蝦	0.016973	2006	垂	0.00268
77	腦	0.227801	1042	啟	0.016973	2007	嫖	0.00268
78	經	0.226907	1043	短	0.016973	2008	隔	0.00268
79	服	0.224227	1044	伯	0.016973	2009	沐	0.00268
80	太	0.223334	1045	庫	0.016973	2010	挫	0.00268
81	報	0.220654	1046	波	0.016973	2011	毫	0.00268
82	體	0.220654	1047	仿	0.016973	2012	妻	0.00268

序號	字	頻次	序號	字	頻次	序號	字	頻次
83	品	0.219761	1048	劑	0.016973	2013	襄	0.00268
84	長	0.219761	1049	輸	0.016973	2014	噹	0.00268
85	玩	0.219761	1050	暖	0.016973	2015	燙	0.00268
86	外	0.218867	1051	控	0.016973	2016	懸	0.00268
87	師	0.217974	1052	型	0.016973	2017	佼	0.00268
88	跳	0.211721	1053	座	0.016973	2018	玫	0.00268
89	文	0.210827	1054	越	0.016973	2019	剖	0.00268
90	一	0.210827	1055	廢	0.016973	2020	邏	0.00268
91	星	0.210827	1056	兩	0.016973	2021	塢	0.00268
92	教	0.210827	1057	洗	0.016973	2022	叮	0.00268
93	中	0.210827	1058	厭	0.01608	2023	墓	0.00268
94	科	0.209934	1059	紛	0.01608	2024	般	0.00268
95	馬	0.205467	1060	夾	0.01608	2025	坪	0.00268
96	劇	0.204574	1061	剪	0.01608	2026	騰	0.00268
97	力	0.203681	1062	次	0.01608	2027	乃	0.00268
98	自	0.202787	1063	誰	0.01608	2028	且	0.00268
99	要	0.201894	1064	促	0.01608	2029	侶	0.00268
100	戲	0.201894	1065	雅	0.01608	2030	擔	0.00268
101	資	0.201894	1066	妝	0.01608	2031	恩	0.00268
102	院	0.201894	1067	噴	0.01608	2032	嘿	0.00268
103	款	0.201001	1068	島	0.01608	2033	依	0.00268
104	氣	0.200107	1069	粉	0.01608	2034	哭	0.00268
105	藝	0.199214	1070	宋	0.01608	2035	杜	0.00268
106	空	0.199214	1071	緊	0.01608	2036	闊	0.00268
107	理	0.198321	1072	掛	0.01608	2037	帆	0.00268
108	作	0.198321	1073	位	0.01608	2038	桑	0.00268
109	治	0.197427	1074	肚	0.01608	2039	跨	0.00268
110	廣	0.195641	1075	獎	0.01608	2040	稀	0.00268
111	牧	0.195641	1076	購	0.01608	2041	邁	0.00268
112	存	0.191174	1077	娜	0.01608	2042	邪	0.00268
113	發	0.191174	1078	雪	0.01608	2043	亦	0.00268
114	片	0.190281	1079	接	0.01608	2044	撿	0.00268
115	出	0.189387	1080	晚	0.01608	2045	瑪	0.00268
116	羊	0.187601	1081	噪	0.015187	2046	陷	0.00268
117	在	0.186707	1082	策	0.015187	2047	峻	0.00268
118	財	0.186707	1083	毬	0.015187	2048	蠟	0.00268
119	情	0.185814	1084	塔	0.015187	2049	玻	0.00268
120	海	0.18492	1085	鄰	0.015187	2050	誼	0.00268
121	社	0.184027	1086	壯	0.015187	2051	蛤	0.00268
122	全	0.18224	1087	夠	0.015187	2052	佛	0.00268
123	活	0.17956	1088	怎	0.015187	2053	壇	0.00268
124	老	0.178667	1089	究	0.015187	2054	窄	0.00268
125	無	0.178667	1090	劉	0.015187	2055	即	0.00268
126	飛	0.17688	1091	穫	0.015187	2056	彭	0.00268

序號	字	頻次	序號	字	頻次	序號	字	頻次
127	真	0.17688	1092	訴	0.015187	2057	屬	0.00268
128	種	0.175987	1093	叫	0.015187	2058	奴	0.00268
129	主	0.1742	1094	留	0.015187	2059	蚶	0.00268
130	器	0.173307	1095	殖	0.015187	2060	瑩	0.00268
131	山	0.173307	1096	辣	0.015187	2061	突	0.00268
132	火	0.17152	1097	河	0.015187	2062	遇	0.00268
133	醫	0.170627	1098	潛	0.015187	2063	愈	0.00268
134	安	0.170627	1099	攻	0.015187	2064	玄	0.00268
135	播	0.170627	1100	藍	0.015187	2065	蘋	0.00268
136	夫	0.169734	1101	絲	0.015187	2066	迴	0.00268
137	通	0.16884	1102	漏	0.015187	2067	贈	0.00268
138	軍	0.167054	1103	半	0.015187	2068	梁	0.00268
139	流	0.167054	1104	脫	0.015187	2069	巨	0.00268
140	遊	0.16616	1105	閃	0.015187	2070	娥	0.00268
141	用	0.16616	1106	認	0.015187	2071	莉	0.00268
142	色	0.165267	1107	殼	0.015187	2072	妖	0.00268
143	下	0.165267	1108	盡	0.015187	2073	謎	0.00268
144	衣	0.165267	1109	烏	0.015187	2074	冠	0.00268
145	收	0.164374	1110	送	0.014293	2075	縫	0.00268
146	戰	0.16348	1111	飽	0.014293	2076	荒	0.00268
147	船	0.161694	1112	硬	0.014293	2077	漆	0.00268
148	健	0.159907	1113	哪	0.014293	2078	倦	0.00268
149	房	0.159907	1114	戴	0.014293	2079	攜	0.00268
150	手	0.159907	1115	幾	0.014293	2080	陪	0.00268
151	利	0.154547	1116	朗	0.014293	2081	珊	0.00268
152	官	0.154547	1117	捏	0.014293	2082	獵	0.00268
153	西	0.154547	1118	繳	0.014293	2083	鉅	0.00268
154	節	0.151867	1119	鋤	0.014293	2084	衫	0.00268
155	花	0.151867	1120	糟	0.014293	2085	蝕	0.00268
156	光	0.149187	1121	努	0.014293	2086	衰	0.00268
157	告	0.148294	1122	腸	0.014293	2087	魅	0.00268
158	選	0.148294	1123	札	0.014293	2088	辭	0.00268
159	育	0.148294	1124	淨	0.014293	2089	脈	0.00268
160	濟	0.148294	1125	何	0.014293	2090	穗	0.00268
161	以	0.146507	1126	汞	0.014293	2091	縱	0.00268
162	統	0.146507	1127	徒	0.014293	2092	坡	0.00268
163	市	0.146507	1128	取	0.014293	2093	敷	0.00268
164	書	0.146507	1129	旋	0.014293	2094	觸	0.00268
165	身	0.145614	1130	煙	0.014293	2095	詳	0.00268
166	草	0.142934	1131	呼	0.014293	2096	紫	0.00268
167	放	0.14204	1132	揮	0.014293	2097	凡	0.00268
168	場	0.14204	1133	周	0.014293	2098	君	0.00268
169	彩	0.141147	1134	虛	0.014293	2099	絡	0.00268
170	日	0.140254	1135	執	0.014293	2100	串	0.00268

序號	字	頻次	序號	字	頻次	序號	字	頻次
171	司	0.140254	1136	梯	0.014293	2101	慌	0.00268
172	開	0.138467	1137	涼	0.014293	2102	倍	0.00268
173	明	0.138467	1138	壁	0.014293	2103	鼎	0.00268
174	道	0.137574	1139	惠	0.014293	2104	豎	0.00268
175	時	0.137574	1140	恰	0.014293	2105	鮪	0.00268
176	提	0.134	1141	喇	0.014293	2106	熊	0.00268
177	燈	0.134	1142	艇	0.014293	2107	仕	0.00268
178	導	0.133107	1143	譜	0.014293	2108	琅	0.00268
179	聽	0.133107	1144	郭	0.014293	2109	仲	0.00268
180	律	0.13132	1145	則	0.014293	2110	右	0.00268
181	銀	0.129534	1146	航	0.014293	2111	沸	0.00268
182	步	0.12864	1147	沫	0.014293	2112	蔣	0.00268
183	跑	0.12864	1148	緻	0.014293	2113	蔔	0.00268
184	暴	0.127747	1149	叭	0.014293	2114	瘟	0.00268
185	進	0.127747	1150	勾	0.014293	2115	蘿	0.00268
186	紅	0.12596	1151	鉛	0.014293	2116	軌	0.00268
187	龍	0.12596	1152	普	0.014293	2117	縮	0.00268
188	目	0.125067	1153	臭	0.014293	2118	瞎	0.00268
189	是	0.125067	1154	劃	0.014293	2119	託	0.00268
190	稻	0.125067	1155	俠	0.014293	2120	慰	0.00268
191	到	0.125067	1156	浮	0.014293	2121	猴	0.00268
192	唱	0.124174	1157	溉	0.0134	2122	袖	0.00268
193	達	0.124174	1158	雄	0.0134	2123	誨	0.00268
194	米	0.124174	1159	醜	0.0134	2124	俊	0.00268
195	和	0.12328	1160	符	0.0134	2125	擬	0.00268
196	技	0.12328	1161	未	0.0134	2126	姍	0.00268
197	象	0.122387	1162	序	0.0134	2127	矩	0.00268
198	運	0.1206	1163	談	0.0134	2128	鷗	0.001787
199	們	0.1206	1164	噁	0.0134	2129	巷	0.001787
200	黑	0.119707	1165	梅	0.0134	2130	敘	0.001787
201	融	0.118814	1166	陣	0.0134	2131	茄	0.001787
202	土	0.118814	1167	婆	0.0134	2132	腎	0.001787
203	愛	0.118814	1168	標	0.0134	2133	液	0.001787
204	來	0.11792	1169	犬	0.0134	2134	述	0.001787
205	許	0.11792	1170	趕	0.0134	2135	寮	0.001787
206	者	0.117027	1171	尋	0.0134	2136	淹	0.001787
207	料	0.117027	1172	捉	0.0134	2137	瞋	0.001787
208	總	0.117027	1173	今	0.0134	2138	飧	0.001787
209	東	0.117027	1174	焦	0.0134	2139	稍	0.001787
210	買	0.11524	1175	箏	0.0134	2140	弗	0.001787
211	食	0.11524	1176	列	0.0134	2141	若	0.001787
212	感	0.11524	1177	喬	0.0134	2142	龐	0.001787
213	意	0.11524	1178	蟲	0.0134	2143	珮	0.001787
214	企	0.11524	1179	楊	0.0134	2144	攬	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
215	務	0.114347	1180	貓	0.0134	2145	渲	0.001787
216	死	0.113454	1181	賭	0.0134	2146	鹹	0.001787
217	寫	0.11256	1182	迷	0.0134	2147	契	0.001787
218	頭	0.11256	1183	狂	0.012507	2148	粹	0.001787
219	排	0.110774	1184	宜	0.012507	2149	宰	0.001787
220	警	0.110774	1185	奈	0.012507	2150	硫	0.001787
221	病	0.10988	1186	終	0.012507	2151	跡	0.001787
222	成	0.10988	1187	廷	0.012507	2152	慈	0.001787
223	景	0.108987	1188	週	0.012507	2153	饋	0.001787
224	肉	0.108987	1189	引	0.012507	2154	祈	0.001787
225	傳	0.108094	1190	講	0.012507	2155	諷	0.001787
226	察	0.108094	1191	鑽	0.012507	2156	橄	0.001787
227	沒	0.106307	1192	完	0.012507	2157	哺	0.001787
228	攝	0.106307	1193	楚	0.012507	2158	莓	0.001787
229	本	0.105414	1194	及	0.012507	2159	屆	0.001787
230	犯	0.10452	1195	撒	0.012507	2160	基	0.001787
231	常	0.10452	1196	繁	0.012507	2161	弄	0.001787
232	鋼	0.103627	1197	於	0.012507	2162	侖	0.001787
233	殺	0.103627	1198	復	0.012507	2163	弋	0.001787
234	合	0.103627	1199	曉	0.012507	2164	嶼	0.001787
235	照	0.103627	1200	摸	0.012507	2165	曠	0.001787
236	田	0.102734	1201	換	0.012507	2166	庶	0.001787
237	計	0.10184	1202	椅	0.012507	2167	矣	0.001787
238	口	0.10184	1203	詐	0.012507	2168	磚	0.001787
239	能	0.10184	1204	橋	0.012507	2169	蠅	0.001787
240	板	0.100947	1205	癌	0.012507	2170	始	0.001787
241	方	0.100054	1206	秩	0.012507	2171	碌	0.001787
242	快	0.100054	1207	姊	0.012507	2172	柱	0.001787
243	年	0.100054	1208	鯊	0.012507	2173	歉	0.001787
244	媽	0.100054	1209	池	0.012507	2174	蓬	0.001787
245	做	0.100054	1210	虎	0.012507	2175	囚	0.001787
246	點	0.100054	1211	副	0.012507	2176	顫	0.001787
247	籃	0.100054	1212	智	0.012507	2177	剩	0.001787
248	林	0.09916	1213	蓋	0.012507	2178	嫩	0.001787
249	八	0.09916	1214	檳	0.012507	2179	饒	0.001787
250	裝	0.09916	1215	篋	0.012507	2180	涎	0.001787
251	表	0.098267	1216	凱	0.012507	2181	盃	0.001787
252	跌	0.098267	1217	黏	0.012507	2182	跌	0.001787
253	彈	0.098267	1218	欺	0.012507	2183	捍	0.001787
254	漲	0.098267	1219	每	0.011613	2184	扶	0.001787
255	展	0.097374	1220	詞	0.011613	2185	痺	0.001787
256	衛	0.09648	1221	末	0.011613	2186	誹	0.001787
257	投	0.09648	1222	隨	0.011613	2187	屛	0.001787
258	易	0.09648	1223	燕	0.011613	2188	俏	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
259	搶	0.095587	1224	淋	0.011613	2189	醺	0.001787
260	部	0.095587	1225	浴	0.011613	2190	驅	0.001787
261	樓	0.094694	1226	諾	0.011613	2191	絮	0.001787
262	內	0.0938	1227	臉	0.011613	2192	囟	0.001787
263	爭	0.092907	1228	介	0.011613	2193	逼	0.001787
264	平	0.092907	1229	刷	0.011613	2194	憫	0.001787
265	特	0.092907	1230	斷	0.011613	2195	抬	0.001787
266	勞	0.092907	1231	潑	0.011613	2196	堪	0.001787
267	級	0.092907	1232	組	0.011613	2197	攏	0.001787
268	鐵	0.092014	1233	舍	0.011613	2198	鍍	0.001787
269	果	0.092014	1234	懷	0.011613	2199	犧	0.001787
270	三	0.092014	1235	蹂	0.011613	2200	輩	0.001787
271	游	0.09112	1236	榔	0.011613	2201	謹	0.001787
272	養	0.09112	1237	靠	0.011613	2202	括	0.001787
273	舉	0.09112	1238	蕭	0.011613	2203	潢	0.001787
274	個	0.09112	1239	骨	0.011613	2204	壺	0.001787
275	李	0.09112	1240	儀	0.011613	2205	穴	0.001787
276	護	0.090227	1241	恢	0.011613	2206	筍	0.001787
277	造	0.090227	1242	預	0.011613	2207	拾	0.001787
278	雕	0.090227	1243	薪	0.011613	2208	彫	0.001787
279	集	0.089334	1244	瑜	0.011613	2209	鴿	0.001787
280	園	0.089334	1245	哲	0.011613	2210	宴	0.001787
281	亂	0.089334	1246	歲	0.011613	2211	鎂	0.001787
282	士	0.089334	1247	競	0.011613	2212	哨	0.001787
283	亮	0.087547	1248	禮	0.011613	2213	卵	0.001787
284	實	0.087547	1249	丹	0.011613	2214	羞	0.001787
285	強	0.086654	1250	胡	0.011613	2215	舅	0.001787
286	商	0.086654	1251	示	0.011613	2216	桶	0.001787
287	盤	0.086654	1252	離	0.011613	2217	亨	0.001787
288	分	0.086654	1253	餓	0.011613	2218	塘	0.001787
289	菜	0.08576	1254	殘	0.011613	2219	祝	0.001787
290	環	0.08576	1255	松	0.011613	2220	澳	0.001787
291	線	0.08576	1256	診	0.011613	2221	蒼	0.001787
292	記	0.08576	1257	罐	0.011613	2222	斐	0.001787
293	立	0.084867	1258	耐	0.011613	2223	珈	0.001787
294	賺	0.084867	1259	髒	0.011613	2224	挖	0.001787
295	想	0.084867	1260	薯	0.011613	2225	槽	0.001787
296	飯	0.083974	1261	娘	0.011613	2226	惱	0.001787
297	都	0.083974	1262	祖	0.011613	2227	菠	0.001787
298	這	0.083974	1263	審	0.011613	2228	赫	0.001787
299	蹈	0.083974	1264	震	0.011613	2229	既	0.001787
300	像	0.08308	1265	差	0.011613	2230	瑚	0.001787
301	汽	0.08308	1266	竿	0.011613	2231	覓	0.001787
302	紙	0.08308	1267	略	0.011613	2232	鋁	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
303	件	0.08308	1268	係	0.011613	2233	曆	0.001787
304	月	0.08308	1269	伍	0.011613	2234	湧	0.001787
305	白	0.08308	1270	姆	0.011613	2235	啥	0.001787
306	精	0.082187	1271	念	0.011613	2236	鮭	0.001787
307	麥	0.082187	1272	隻	0.011613	2237	亢	0.001787
308	索	0.082187	1273	搞	0.011613	2238	薙	0.001787
309	琴	0.082187	1274	久	0.01072	2239	謊	0.001787
310	釣	0.082187	1275	驚	0.01072	2240	俑	0.001787
311	雞	0.081294	1276	獲	0.01072	2241	吞	0.001787
312	仔	0.081294	1277	堆	0.01072	2242	襯	0.001787
313	賣	0.081294	1278	粗	0.01072	2243	妨	0.001787
314	訊	0.0804	1279	歷	0.01072	2244	煞	0.001787
315	府	0.0804	1280	歸	0.01072	2245	肝	0.001787
316	相	0.0804	1281	症	0.01072	2246	劍	0.001787
317	知	0.079507	1282	瓦	0.01072	2247	啪	0.001787
318	息	0.079507	1283	積	0.01072	2248	嗦	0.001787
319	豬	0.078614	1284	娼	0.01072	2249	洵	0.001787
320	張	0.078614	1285	鍛	0.01072	2250	玫	0.001787
321	命	0.078614	1286	筋	0.01072	2251	瑰	0.001787
322	哥	0.07772	1287	段	0.01072	2252	墳	0.001787
323	當	0.07772	1288	滋	0.01072	2253	狹	0.001787
324	克	0.07772	1289	賂	0.01072	2254	燃	0.001787
325	碟	0.076827	1290	負	0.01072	2255	盧	0.001787
326	聊	0.076827	1291	尚	0.01072	2256	磅	0.001787
327	黨	0.076827	1292	希	0.01072	2257	答	0.001787
328	染	0.075934	1293	宅	0.01072	2258	崙	0.001787
329	兒	0.075934	1294	哇	0.01072	2259	咪	0.001787
330	起	0.075934	1295	狼	0.01072	2260	饅	0.001787
331	關	0.075934	1296	蝸	0.01072	2261	孜	0.001787
332	停	0.075934	1297	訪	0.01072	2262	腑	0.001787
333	禍	0.075934	1298	爐	0.01072	2263	弛	0.001787
334	斯	0.07504	1299	韻	0.01072	2264	萱	0.001787
335	管	0.07504	1300	里	0.01072	2265	霄	0.001787
336	畢	0.074147	1301	臺	0.01072	2266	凹	0.001787
337	代	0.074147	1302	幻	0.01072	2267	傲	0.001787
338	性	0.074147	1303	斃	0.01072	2268	譯	0.001787
339	王	0.073254	1304	向	0.01072	2269	狄	0.001787
340	屋	0.073254	1305	襪	0.01072	2270	陌	0.001787
341	門	0.073254	1306	佳	0.01072	2271	吾	0.001787
342	基	0.073254	1307	夢	0.01072	2272	淒	0.001787
343	局	0.073254	1308	碼	0.01072	2273	痞	0.001787
344	帶	0.073254	1309	佈	0.01072	2274	敬	0.001787
345	針	0.073254	1310	穀	0.01072	2275	疲	0.001787
346	登	0.073254	1311	午	0.01072	2276	敵	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
347	具	0.07236	1312	幹	0.01072	2277	澎	0.001787
348	話	0.07236	1313	采	0.01072	2278	朽	0.001787
349	綠	0.07236	1314	糾	0.009827	2279	刮	0.001787
350	字	0.07236	1315	爽	0.009827	2280	併	0.001787
351	費	0.07236	1316	該	0.009827	2281	瓊	0.001787
352	課	0.07236	1317	伏	0.009827	2282	儂	0.001787
353	建	0.07236	1318	誇	0.009827	2283	晴	0.001787
354	泳	0.071467	1319	肢	0.009827	2284	賈	0.001787
355	現	0.071467	1320	享	0.009827	2285	捰	0.001787
356	拉	0.071467	1321	崩	0.009827	2286	歪	0.001787
357	原	0.071467	1322	追	0.009827	2287	莘	0.001787
358	污	0.071467	1323	榮	0.009827	2288	稚	0.001787
359	制	0.071467	1324	鎖	0.009827	2289	鬱	0.001787
360	毛	0.070574	1325	擴	0.009827	2290	齒	0.001787
361	油	0.070574	1326	侵	0.009827	2291	霸	0.001787
362	校	0.070574	1327	糕	0.009827	2292	惑	0.001787
363	遠	0.06968	1328	豪	0.009827	2293	搜	0.001787
364	你	0.06968	1329	渡	0.009827	2294	慾	0.001787
365	圖	0.06968	1330	鍋	0.009827	2295	扭	0.001787
366	而	0.06968	1331	濫	0.009827	2296	撐	0.001787
367	女	0.06968	1332	瓷	0.009827	2297	寸	0.001787
368	石	0.068787	1333	憶	0.009827	2298	拮	0.001787
369	得	0.068787	1334	圍	0.009827	2299	溪	0.001787
370	廠	0.068787	1335	唯	0.009827	2300	返	0.001787
371	友	0.068787	1336	置	0.009827	2301	汲	0.001787
372	便	0.067894	1337	匹	0.009827	2302	諄	0.001787
373	苦	0.067894	1338	席	0.009827	2303	啞	0.001787
374	陳	0.067894	1339	遺	0.009827	2304	旭	0.001787
375	信	0.067894	1340	籠	0.009827	2305	牙	0.001787
376	洋	0.067894	1341	珍	0.009827	2306	賤	0.001787
377	華	0.067894	1342	寧	0.009827	2307	湊	0.001787
378	包	0.067894	1343	翁	0.009827	2308	振	0.001787
379	筆	0.067	1344	擁	0.009827	2309	榜	0.001787
380	住	0.067	1345	腥	0.009827	2310	訟	0.001787
381	熱	0.067	1346	擊	0.009827	2311	鋪	0.001787
382	康	0.067	1347	廉	0.009827	2312	彰	0.001787
383	設	0.066107	1348	況	0.009827	2313	跽	0.001787
384	之	0.066107	1349	藏	0.009827	2314	朱	0.001787
385	為	0.065214	1350	缺	0.009827	2315	擎	0.001787
386	漂	0.065214	1351	邦	0.009827	2316	盪	0.001787
387	肥	0.065214	1352	顯	0.009827	2317	抄	0.001787
388	間	0.065214	1353	雙	0.009827	2318	趨	0.001787
389	館	0.065214	1354	膠	0.009827	2319	慕	0.001787
390	聲	0.065214	1355	貧	0.009827	2320	丈	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
391	稅	0.06432	1356	揚	0.009827	2321	脹	0.001787
392	踏	0.06432	1357	弊	0.009827	2322	匆	0.001787
393	然	0.06432	1358	俱	0.009827	2323	嶽	0.001787
394	支	0.06432	1359	莊	0.009827	2324	濤	0.001787
395	麼	0.063427	1360	茲	0.009827	2325	睦	0.001787
396	休	0.063427	1361	聚	0.009827	2326	茱	0.001787
397	世	0.063427	1362	洲	0.009827	2327	倉	0.001787
398	劫	0.063427	1363	晶	0.009827	2328	琳	0.001787
399	青	0.062534	1364	急	0.009827	2329	岡	0.001787
400	製	0.062534	1365	餘	0.009827	2330	墮	0.001787
401	走	0.062534	1366	隆	0.009827	2331	甫	0.001787
402	英	0.062534	1367	它	0.008933	2332	桂	0.001787
403	古	0.062534	1368	哀	0.008933	2333	纜	0.001787
404	所	0.062534	1369	淇	0.008933	2334	占	0.001787
405	娛	0.06164	1370	擦	0.008933	2335	煎	0.001787
406	數	0.06164	1371	遙	0.008933	2336	銘	0.001787
407	重	0.06164	1372	窮	0.008933	2337	咻	0.001787
408	加	0.06164	1373	灸	0.008933	2338	嫂	0.001787
409	木	0.06164	1374	擺	0.008933	2339	渝	0.001787
410	案	0.060747	1375	忽	0.008933	2340	愚	0.001787
411	味	0.060747	1376	閭	0.008933	2341	菱	0.001787
412	滑	0.060747	1377	疫	0.008933	2342	拚	0.001787
413	難	0.059853	1378	奔	0.008933	2343	乚	0.001787
414	漫	0.059853	1379	尖	0.008933	2344	凸	0.001787
415	團	0.059853	1380	嘴	0.008933	2345	蟻	0.001787
416	素	0.059853	1381	痴	0.008933	2346	餌	0.001787
417	防	0.059853	1382	錶	0.008933	2347	鯉	0.001787
418	神	0.059853	1383	慘	0.008933	2348	閩	0.001787
419	危	0.059853	1384	憐	0.008933	2349	昌	0.001787
420	爸	0.05896	1385	勿	0.008933	2350	彬	0.001787
421	災	0.05896	1386	廚	0.008933	2351	澡	0.001787
422	店	0.05896	1387	棚	0.008933	2352	祟	0.001787
423	布	0.05896	1388	壘	0.008933	2353	痕	0.001787
424	奶	0.05896	1389	堂	0.008933	2354	娑	0.001787
425	單	0.05896	1390	綿	0.008933	2355	氏	0.001787
426	領	0.058067	1391	蹺	0.008933	2356	憤	0.001787
427	足	0.058067	1392	懶	0.008933	2357	膀	0.001787
428	餐	0.058067	1393	腿	0.008933	2358	媚	0.001787
429	類	0.058067	1394	售	0.008933	2359	掏	0.001787
430	習	0.058067	1395	甜	0.008933	2360	臨	0.001787
431	錄	0.058067	1396	惶	0.008933	2361	淵	0.001787
432	別	0.058067	1397	宵	0.008933	2362	勵	0.001787
433	觀	0.058067	1398	掉	0.008933	2363	殆	0.001787
434	耕	0.057173	1399	孟	0.008933	2364	悔	0.001787

序號	字	頻次	序號	字	頻次	序號	字	頻次
435	德	0.057173	1400	坯	0.008933	2365	黎	0.001787
436	九	0.057173	1401	凍	0.008933	2366	穎	0.001787
437	捕	0.057173	1402	甘	0.008933	2367	頤	0.001787
438	受	0.057173	1403	蹟	0.008933	2368	彎	0.001787
439	印	0.057173	1404	仙	0.008933	2369	恨	0.001787
440	連	0.057173	1405	閣	0.008933	2370	尤	0.001787
441	界	0.057173	1406	頻	0.008933	2371	喧	0.001787
442	繪	0.05628	1407	倫	0.008933	2372	淪	0.001787
443	說	0.05628	1408	伊	0.008933	2373	趙	0.001787
444	壽	0.05628	1409	沈	0.008933	2374	驕	0.001787
445	入	0.05628	1410	否	0.008933	2375	据	0.001787
446	架	0.05628	1411	職	0.008933	2376	灘	0.001787
447	貝	0.05628	1412	爵	0.008933	2377	億	0.001787
448	戶	0.05628	1413	舌	0.008933	2378	唇	0.001787
449	武	0.05628	1414	曼	0.008933	2379	偏	0.001787
450	禪	0.05628	1415	往	0.008933	2380	州	0.001787
451	冰	0.05628	1416	聖	0.008933	2381	奕	0.001787
452	處	0.055387	1417	析	0.008933	2382	貫	0.001787
453	街	0.055387	1418	純	0.008933	2383	沉	0.001787
454	讓	0.055387	1419	徑	0.008933	2384	讚	0.001787
455	錯	0.055387	1420	距	0.008933	2385	錦	0.001787
456	賽	0.055387	1421	折	0.008933	2386	繚	0.001787
457	曲	0.055387	1422	賓	0.008933	2387	覆	0.001787
458	少	0.055387	1423	切	0.008933	2388	溺	0.001787
459	鞋	0.054493	1424	首	0.008933	2389	荷	0.001787
460	棒	0.054493	1425	墜	0.00804	2390	愷	0.001787
461	麗	0.054493	1426	憂	0.00804	2391	佔	0.001787
462	質	0.054493	1427	戀	0.00804	2392	榻	0.001787
463	營	0.054493	1428	寬	0.00804	2393	矚	0.001787
464	考	0.054493	1429	烈	0.00804	2394	遲	0.001787
465	輕	0.054493	1430	誠	0.00804	2395	劣	0.001787
466	室	0.054493	1431	泉	0.00804	2396	寡	0.001787
467	角	0.054493	1432	志	0.00804	2397	繽	0.001787
468	慶	0.0536	1433	籍	0.00804	2398	伽	0.000893
469	拍	0.0536	1434	胖	0.00804	2399	鉗	0.000893
470	被	0.0536	1435	嘛	0.00804	2400	塾	0.000893
471	使	0.0536	1436	輻	0.00804	2401	筴	0.000893
472	威	0.0536	1437	悅	0.00804	2402	脂	0.000893
473	義	0.0536	1438	拼	0.00804	2403	菊	0.000893
474	比	0.052707	1439	迅	0.00804	2404	鷓	0.000893
475	也	0.052707	1440	勒	0.00804	2405	淳	0.000893
476	髮	0.052707	1441	旗	0.00804	2406	棍	0.000893
477	永	0.052707	1442	紗	0.00804	2407	埔	0.000893
478	尼	0.052707	1443	困	0.00804	2408	賬	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
479	號	0.052707	1444	蹄	0.00804	2409	喂	0.000893
480	閒	0.052707	1445	匪	0.00804	2410	駟	0.000893
481	率	0.052707	1446	輯	0.00804	2411	損	0.000893
482	雜	0.052707	1447	達	0.00804	2412	昭	0.000893
483	嗎	0.051813	1448	範	0.00804	2413	飪	0.000893
484	腳	0.051813	1449	甲	0.00804	2414	鮑	0.000893
485	福	0.051813	1450	獸	0.00804	2415	埂	0.000893
486	炮	0.051813	1451	膏	0.00804	2416	嫦	0.000893
487	速	0.051813	1452	互	0.00804	2417	迎	0.000893
488	狗	0.051813	1453	挺	0.00804	2418	枕	0.000893
489	穿	0.051813	1454	蔡	0.00804	2419	瞬	0.000893
490	鍊	0.051813	1455	蹤	0.00804	2420	離	0.000893
491	塑	0.051813	1456	喉	0.00804	2421	扎	0.000893
492	貨	0.05092	1457	芋	0.00804	2422	敖	0.000893
493	套	0.05092	1458	冬	0.00804	2423	禱	0.000893
494	五	0.05092	1459	棉	0.00804	2424	狸	0.000893
495	逛	0.05092	1460	償	0.00804	2425	脖	0.000893
496	百	0.05092	1461	欠	0.00804	2426	盪	0.000893
497	定	0.05092	1462	慧	0.00804	2427	艾	0.000893
498	麵	0.050027	1463	從	0.00804	2428	蜈	0.000893
499	貿	0.050027	1464	眠	0.00804	2429	膩	0.000893
500	野	0.050027	1465	朵	0.00804	2430	鈕	0.000893
501	名	0.050027	1466	轎	0.00804	2431	呂	0.000893
502	耳	0.050027	1467	爺	0.00804	2432	譽	0.000893
503	汙	0.050027	1468	顧	0.00804	2433	燭	0.000893
504	害	0.050027	1469	逃	0.00804	2434	儒	0.000893
505	元	0.049133	1470	窗	0.00804	2435	唆	0.000893
506	俗	0.049133	1471	退	0.00804	2436	偶	0.000893
507	過	0.049133	1472	孫	0.00804	2437	昨	0.000893
508	陸	0.049133	1473	宏	0.00804	2438	烹	0.000893
509	刻	0.049133	1474	桃	0.00804	2439	肴	0.000893
510	爾	0.049133	1475	帝	0.00804	2440	隧	0.000893
511	貸	0.049133	1476	蓮	0.00804	2441	凰	0.000893
512	童	0.049133	1477	啡	0.00804	2442	騫	0.000893
513	清	0.049133	1478	翻	0.00804	2443	蟬	0.000893
514	鼠	0.049133	1479	庸	0.007147	2444	佐	0.000893
515	同	0.049133	1480	糧	0.007147	2445	樊	0.000893
516	騎	0.04824	1481	賢	0.007147	2446	贏	0.000893
517	識	0.04824	1482	勁	0.007147	2447	藉	0.000893
518	轉	0.04824	1483	偽	0.007147	2448	藩	0.000893
519	著	0.04824	1484	奮	0.007147	2449	匈	0.000893
520	爬	0.04824	1485	屁	0.007147	2450	揆	0.000893
521	銷	0.047347	1486	攤	0.007147	2451	棧	0.000893
522	笑	0.047347	1487	潔	0.007147	2452	蚊	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
523	證	0.047347	1488	較	0.007147	2453	后	0.000893
524	貪	0.047347	1489	恆	0.007147	2454	潭	0.000893
525	茶	0.047347	1490	砂	0.007147	2455	椰	0.000893
526	失	0.047347	1491	申	0.007147	2456	蟠	0.000893
527	描	0.046453	1492	敦	0.007147	2457	昏	0.000893
528	指	0.046453	1493	氣	0.007147	2458	撕	0.000893
529	第	0.046453	1494	赤	0.007147	2459	轡	0.000893
530	權	0.046453	1495	裕	0.007147	2460	湄	0.000893
531	期	0.046453	1496	鹿	0.007147	2461	啣	0.000893
532	居	0.04556	1497	根	0.007147	2462	駛	0.000893
533	溜	0.04556	1498	順	0.007147	2463	軸	0.000893
534	香	0.04556	1499	乏	0.007147	2464	蠹	0.000893
535	升	0.04556	1500	魯	0.007147	2465	伙	0.000893
536	巴	0.04556	1501	閔	0.007147	2466	僑	0.000893
537	飾	0.04556	1502	鹽	0.007147	2467	吟	0.000893
538	箱	0.044667	1503	薄	0.007147	2468	隴	0.000893
539	扁	0.044667	1504	崇	0.007147	2469	狐	0.000893
540	際	0.044667	1505	泛	0.007147	2470	御	0.000893
541	如	0.044667	1506	拳	0.007147	2471	叛	0.000893
542	袋	0.044667	1507	並	0.007147	2472	羿	0.000893
543	黃	0.044667	1508	盈	0.007147	2473	佑	0.000893
544	四	0.044667	1509	津	0.007147	2474	胙	0.000893
545	倒	0.044667	1510	役	0.007147	2475	臂	0.000893
546	條	0.044667	1511	弱	0.007147	2476	恭	0.000893
547	超	0.044667	1512	臥	0.007147	2477	藹	0.000893
548	烤	0.043773	1513	尾	0.007147	2478	徹	0.000893
549	毒	0.043773	1514	蘇	0.007147	2479	穀	0.000893
550	又	0.043773	1515	柵	0.007147	2480	粟	0.000893
551	郵	0.043773	1516	狀	0.007147	2481	蚯	0.000893
552	派	0.043773	1517	塗	0.007147	2482	鴻	0.000893
553	喝	0.043773	1518	匙	0.007147	2483	伺	0.000893
554	槍	0.043773	1519	蛙	0.007147	2484	幢	0.000893
555	裡	0.043773	1520	默	0.007147	2485	甚	0.000893
556	塞	0.043773	1521	杯	0.007147	2486	躬	0.000893
557	消	0.043773	1522	紹	0.007147	2487	蚓	0.000893
558	皮	0.043773	1523	鼻	0.007147	2488	嚮	0.000893
559	芬	0.043773	1524	昂	0.007147	2489	濕	0.000893
560	鈴	0.04288	1525	封	0.007147	2490	榴	0.000893
561	減	0.04288	1526	苗	0.007147	2491	螞	0.000893
562	匯	0.04288	1527	暢	0.007147	2492	禪	0.000893
563	玉	0.041987	1528	魔	0.007147	2493	姻	0.000893
564	母	0.041987	1529	絕	0.007147	2494	蜀	0.000893
565	刑	0.041987	1530	衝	0.007147	2495	囊	0.000893
566	廈	0.041987	1531	夏	0.007147	2496	礁	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
567	興	0.041987	1532	止	0.007147	2497	悚	0.000893
568	罪	0.041987	1533	逸	0.007147	2498	鈎	0.000893
569	最	0.041987	1534	曝	0.007147	2499	碾	0.000893
570	羅	0.041987	1535	廊	0.007147	2500	瞻	0.000893
571	桌	0.041987	1536	仗	0.007147	2501	哼	0.000893
572	面	0.041987	1537	衡	0.007147	2502	蚣	0.000893
573	價	0.041987	1538	酸	0.007147	2503	徙	0.000893
574	族	0.041987	1539	淡	0.007147	2504	履	0.000893
575	兵	0.041987	1540	祿	0.006253	2505	鱒	0.000893
576	度	0.041987	1541	霹	0.006253	2506	穢	0.000893
577	芭	0.041093	1542	灰	0.006253	2507	皆	0.000893
578	章	0.041093	1543	確	0.006253	2508	儘	0.000893
579	哦	0.041093	1544	汁	0.006253	2509	閤	0.000893
580	幫	0.041093	1545	紐	0.006253	2510	籟	0.000893
581	飲	0.041093	1546	姓	0.006253	2511	簪	0.000893
582	博	0.041093	1547	蔗	0.006253	2512	濁	0.000893
583	核	0.041093	1548	免	0.006253	2513	濱	0.000893
584	朋	0.041093	1549	搭	0.006253	2514	壑	0.000893
585	程	0.041093	1550	豔	0.006253	2515	乘	0.000893
586	萬	0.041093	1551	禦	0.006253	2516	拋	0.000893
587	扯	0.041093	1552	仰	0.006253	2517	裸	0.000893
588	血	0.041093	1553	映	0.006253	2518	棗	0.000893
589	沙	0.041093	1554	佰	0.006253	2519	饒	0.000893
590	雷	0.041093	1555	霍	0.006253	2520	倩	0.000893
591	功	0.0402	1556	牲	0.006253	2521	駢	0.000893
592	直	0.0402	1557	虱	0.006253	2522	賦	0.000893
593	鈔	0.0402	1558	幸	0.006253	2523	仇	0.000893
594	賞	0.0402	1559	勇	0.006253	2524	敞	0.000893
595	誌	0.0402	1560	猛	0.006253	2525	蒜	0.000893
596	飆	0.0402	1561	臟	0.006253	2526	鶴	0.000893
597	材	0.0402	1562	蓄	0.006253	2527	紉	0.000893
598	割	0.0402	1563	囉	0.006253	2528	簇	0.000893
599	模	0.0402	1564	敢	0.006253	2529	曄	0.000893
600	拜	0.0402	1565	蜓	0.006253	2530	嶇	0.000893
601	什	0.0402	1566	承	0.006253	2531	櫻	0.000893
602	射	0.0402	1567	砲	0.006253	2532	鈣	0.000893
603	式	0.0402	1568	蟹	0.006253	2533	籽	0.000893
604	富	0.0402	1569	降	0.006253	2534	臣	0.000893
605	趣	0.039307	1570	蜻	0.006253	2535	杰	0.000893
606	把	0.039307	1571	棟	0.006253	2536	愁	0.000893
607	非	0.039307	1572	猜	0.006253	2537	襟	0.000893
608	帥	0.039307	1573	誕	0.006253	2538	謗	0.000893
609	給	0.039307	1574	謠	0.006253	2539	稼	0.000893
610	壞	0.039307	1575	瘦	0.006253	2540	眩	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
611	庭	0.039307	1576	署	0.006253	2541	佩	0.000893
612	讀	0.039307	1577	簡	0.006253	2542	葡	0.000893
613	浪	0.039307	1578	扣	0.006253	2543	憔	0.000893
614	聯	0.039307	1579	穩	0.006253	2544	歹	0.000893
615	舒	0.039307	1580	婦	0.006253	2545	悴	0.000893
616	低	0.038413	1581	丁	0.006253	2546	藻	0.000893
617	境	0.038413	1582	徵	0.006253	2547	殯	0.000893
618	益	0.038413	1583	斗	0.006253	2548	哎	0.000893
619	蕾	0.038413	1584	隱	0.006253	2549	猥	0.000893
620	覺	0.038413	1585	秘	0.006253	2550	寺	0.000893
621	因	0.038413	1586	層	0.006253	2551	樟	0.000893
622	繩	0.038413	1587	擇	0.006253	2552	擄	0.000893
623	棋	0.038413	1588	忘	0.006253	2553	唸	0.000893
624	刺	0.038413	1589	滴	0.006253	2554	蟒	0.000893
625	睡	0.038413	1590	醉	0.006253	2555	煌	0.000893
626	糖	0.038413	1591	績	0.006253	2556	遏	0.000893
627	千	0.038413	1592	郎	0.006253	2557	醒	0.000893
628	羽	0.038413	1593	碩	0.006253	2558	烘	0.000893
629	規	0.038413	1594	伐	0.006253	2559	檜	0.000893
630	鏡	0.038413	1595	撥	0.006253	2560	蚌	0.000893
631	亡	0.03752	1596	決	0.006253	2561	暮	0.000893
632	寶	0.03752	1597	循	0.006253	2562	爹	0.000893
633	餅	0.03752	1598	招	0.006253	2563	征	0.000893
634	等	0.03752	1599	愉	0.006253	2564	紆	0.000893
635	旅	0.03752	1600	付	0.006253	2565	美	0.000893
636	輝	0.03752	1601	丟	0.006253	2566	尉	0.000893
637	郊	0.03752	1602	咖	0.006253	2567	屐	0.000893
638	陽	0.03752	1603	遣	0.006253	2568	攢	0.000893
639	父	0.03752	1604	菲	0.006253	2569	乙	0.000893
640	項	0.03752	1605	曳	0.006253	2570	剔	0.000893
641	七	0.03752	1606	繞	0.006253	2571	醅	0.000893
642	券	0.03752	1607	奐	0.006253	2572	橈	0.000893
643	十	0.03752	1608	銅	0.006253	2573	昧	0.000893
644	村	0.03752	1609	巾	0.006253	2574	姜	0.000893
645	班	0.03752	1610	胎	0.006253	2575	墊	0.000893
646	妹	0.03752	1611	劾	0.006253	2576	眉	0.000893
647	憲	0.036627	1612	階	0.006253	2577	傻	0.000893
648	春	0.036627	1613	旦	0.006253	2578	丘	0.000893
649	試	0.036627	1614	昇	0.006253	2579	昊	0.000893
650	態	0.036627	1615	耍	0.006253	2580	皺	0.000893
651	粽	0.036627	1616	兌	0.006253	2581	殃	0.000893
652	議	0.036627	1617	販	0.006253	2582	怨	0.000893
653	泡	0.036627	1618	粒	0.006253	2583	鎚	0.000893
654	幕	0.036627	1619	嚇	0.006253	2584	芹	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
655	施	0.036627	1620	膽	0.006253	2585	茱	0.000893
656	陶	0.036627	1621	謀	0.006253	2586	釜	0.000893
657	酒	0.035733	1622	訓	0.006253	2587	蓉	0.000893
658	卦	0.035733	1623	抹	0.006253	2588	泗	0.000893
659	帳	0.035733	1624	溝	0.006253	2589	妍	0.000893
660	滿	0.035733	1625	冶	0.006253	2590	弑	0.000893
661	軟	0.035733	1626	兇	0.006253	2591	谿	0.000893
662	阿	0.035733	1627	免	0.00536	2592	陰	0.000893
663	六	0.035733	1628	雀	0.00536	2593	黍	0.000893
664	鍵	0.035733	1629	至	0.00536	2594	鞠	0.000893
665	鬆	0.035733	1630	替	0.00536	2595	孃	0.000893
666	革	0.035733	1631	貢	0.00536	2596	柚	0.000893
667	乳	0.035733	1632	鯨	0.00536	2597	呈	0.000893
668	對	0.035733	1633	箋	0.00536	2598	烽	0.000893
669	才	0.035733	1634	勤	0.00536	2599	炳	0.000893
670	散	0.03484	1635	勝	0.00536	2600	閼	0.000893
671	靈	0.03484	1636	拔	0.00536	2601	鐺	0.000893
672	那	0.03484	1637	捐	0.00536	2602	蹶	0.000893
673	抽	0.03484	1638	桿	0.00536	2603	臻	0.000893
674	與	0.03484	1639	煮	0.00536	2604	籌	0.000893
675	變	0.03484	1640	束	0.00536	2605	罔	0.000893
676	喔	0.03484	1641	兄	0.00536	2606	荀	0.000893
677	搖	0.03484	1642	遵	0.00536	2607	痺	0.000893
678	廳	0.03484	1643	址	0.00536	2608	赴	0.000893
679	典	0.03484	1644	滷	0.00536	2609	迭	0.000893
680	優	0.03484	1645	宿	0.00536	2610	夷	0.000893
681	請	0.03484	1646	胞	0.00536	2611	邸	0.000893
682	破	0.03484	1647	躍	0.00536	2612	踐	0.000893
683	帽	0.03484	1648	慎	0.00536	2613	粧	0.000893
684	牢	0.03484	1649	燥	0.00536	2614	賴	0.000893
685	奧	0.03484	1650	餵	0.00536	2615	盾	0.000893
686	假	0.03484	1651	孕	0.00536	2616	崎	0.000893
687	監	0.033947	1652	蒂	0.00536	2617	懿	0.000893
688	獅	0.033947	1653	緒	0.00536	2618	踞	0.000893
689	畜	0.033947	1654	潤	0.00536	2619	陋	0.000893
690	底	0.033947	1655	曬	0.00536	2620	閼	0.000893
691	創	0.033947	1656	殊	0.00536	2621	剝	0.000893
692	辛	0.033947	1657	郡	0.00536	2622	琢	0.000893
693	偷	0.033947	1658	瑞	0.00536	2623	萎	0.000893
694	酷	0.033947	1659	悶	0.00536	2624	添	0.000893
695	顏	0.033947	1660	晴	0.00536	2625	諸	0.000893
696	箭	0.033947	1661	螃	0.00536	2626	禿	0.000893
697	敗	0.033947	1662	吏	0.00536	2627	兼	0.000893
698	更	0.033947	1663	渴	0.00536	2628	椒	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
699	己	0.033947	1664	煤	0.00536	2629	匠	0.000893
700	操	0.033947	1665	屎	0.00536	2630	膝	0.000893
701	夜	0.033947	1666	塢	0.00536	2631	欲	0.000893
702	怪	0.033947	1667	妓	0.00536	2632	凶	0.000893
703	量	0.033947	1668	笠	0.00536	2633	勘	0.000893
704	恐	0.033053	1669	笨	0.00536	2634	餒	0.000893
705	結	0.033053	1670	馨	0.00536	2635	鋤	0.000893
706	漢	0.033053	1671	俄	0.00536	2636	嚐	0.000893
707	眾	0.033053	1672	撲	0.00536	2637	僅	0.000893
708	樹	0.033053	1673	固	0.00536	2638	憾	0.000893
709	奇	0.033053	1674	霖	0.00536	2639	遭	0.000893
710	貴	0.033053	1675	牽	0.00536	2640	鯖	0.000893
711	冷	0.033053	1676	鄭	0.00536	2641	痔	0.000893
712	築	0.033053	1677	胸	0.00536	2642	鷹	0.000893
713	問	0.033053	1678	祕	0.00536	2643	嗲	0.000893
714	源	0.033053	1679	茅	0.00536	2644	渣	0.000893
715	暗	0.033053	1680	洩	0.00536	2645	乞	0.000893
716	改	0.033053	1681	戈	0.00536	2646	涉	0.000893
717	隊	0.033053	1682	鳳	0.00536	2647	註	0.000893
718	由	0.033053	1683	輻	0.00536	2648	嘗	0.000893
719	湯	0.033053	1684	寵	0.00536	2649	腔	0.000893
720	閉	0.033053	1685	頂	0.00536	2650	簾	0.000893
721	響	0.03216	1686	傳	0.00536	2651	窯	0.000893
722	持	0.03216	1687	綱	0.00536	2652	鳶	0.000893
723	瓜	0.03216	1688	晒	0.00536	2653	昔	0.000893
724	森	0.03216	1689	煉	0.00536	2654	鸚	0.000893
725	混	0.03216	1690	抵	0.00536	2655	舊	0.000893
726	故	0.03216	1691	酬	0.00536	2656	杵	0.000893
727	補	0.03216	1692	彼	0.00536	2657	帑	0.000893
728	男	0.03216	1693	遷	0.00536	2658	凋	0.000893
729	鮮	0.03216	1694	怒	0.00536	2659	延	0.000893
730	客	0.03216	1695	魄	0.00536	2660	糲	0.000893
731	豐	0.031267	1696	撲	0.00536	2661	屍	0.000893
732	牌	0.031267	1697	己	0.00536	2662	傀	0.000893
733	輪	0.031267	1698	悠	0.00536	2663	估	0.000893
734	捷	0.031267	1699	寒	0.00536	2664	錠	0.000893
735	梭	0.031267	1700	孤	0.00536	2665	燭	0.000893
736	鬼	0.031267	1701	糞	0.00536	2666	遷	0.000893
737	調	0.031267	1702	輔	0.00536	2667		0.000893
738	港	0.031267	1703	姑	0.00536	2668	唉	0.000893
739	賄	0.031267	1704	偉	0.00536	2669	韁	0.000893
740	齡	0.030373	1705	虧	0.00536	2670	嘩	0.000893
741	端	0.030373	1706	載	0.00536	2671	啤	0.000893
742	就	0.030373	1707	磨	0.00536	2672	貉	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
743	孔	0.030373	1708	評	0.00536	2673	猷	0.000893
744	縣	0.030373	1709	例	0.00536	2674	咧	0.000893
745	他	0.030373	1710	臘	0.00536	2675	蛀	0.000893
746	舟	0.030373	1711	丑	0.00536	2676	菸	0.000893
747	鬥	0.030373	1712	璃	0.00536	2677	蜂	0.000893
748	注	0.030373	1713	獻	0.00536	2678	泣	0.000893
749	泥	0.030373	1714	測	0.00536	2679	竊	0.000893
750	障	0.030373	1715	井	0.00536	2680	抑	0.000893
751	插	0.030373	1716	肺	0.00536	2681	幄	0.000893
752	灣	0.030373	1717	龜	0.00536	2682	孽	0.000893
753	撞	0.030373	1718	翹	0.00536	2683	蔥	0.000893
754	區	0.030373	1719	鎮	0.00536	2684	噶	0.000893
755	傷	0.030373	1720	禁	0.00536	2685	渺	0.000893
756	刀	0.030373	1721	鴉	0.00536	2686	吠	0.000893
757	吳	0.02948	1722	旺	0.00536	2687	蜊	0.000893
758	納	0.02948	1723	舊	0.00536	2688	簣	0.000893
759	礦	0.02948	1724	盲	0.00536	2689	滯	0.000893
760	丸	0.02948	1725	塊	0.00536	2690	槁	0.000893
761	複	0.02948	1726	幽	0.00536	2691	輟	0.000893
762	鵝	0.02948	1727	鈞	0.00536	2692	余	0.000893
763	洞	0.02948	1728	詩	0.00536	2693	白	0.000893
764	推	0.02948	1729	賊	0.00536	2694	鷓	0.000893
765	系	0.02948	1730	均	0.00536	2695	杏	0.000893
766	共	0.02948	1731	漠	0.00536	2696	壅	0.000893
767	各	0.02948	1732	岸	0.00536	2697	誘	0.000893
768	宮	0.02948	1733	梨	0.00536	2698	育	0.000893
769	秀	0.02948	1734	鋒	0.00536	2699	澱	0.000893
770	眼	0.02948	1735	洛	0.00536	2700	瀋	0.000893
771	跟	0.028587	1736	域	0.00536	2701	愣	0.000893
772	怖	0.028587	1737	巡	0.00536	2702	閏	0.000893
773	緋	0.028587	1738	蒸	0.004467	2703	蒲	0.000893
774	驗	0.028587	1739	孝	0.004467	2704	鉀	0.000893
775	續	0.028587	1740	拖	0.004467	2705	眷	0.000893
776	題	0.028587	1741	梳	0.004467	2706	揍	0.000893
777	維	0.028587	1742	皇	0.004467	2707	愿	0.000893
778	語	0.028587	1743	損	0.004467	2708	膳	0.000893
779	落	0.028587	1744	熟	0.004467	2709	昆	0.000893
780	群	0.028587	1745	柴	0.004467	2710	顛	0.000893
781	容	0.027693	1746	宗	0.004467	2711	涂	0.000893
782	炸	0.027693	1747	灑	0.004467	2712	緩	0.000893
783	站	0.027693	1748	嚥	0.004467	2713	妮	0.000893
784	磁	0.027693	1749	逐	0.004467	2714	桐	0.000893
785	移	0.027693	1750	寄	0.004467	2715	泄	0.000893
786	委	0.027693	1751	此	0.004467	2716	肫	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
787	怕	0.027693	1752	裁	0.004467	2717	喊	0.000893
788	莫	0.027693	1753	貶	0.004467	2718	漿	0.000893
789	壓	0.027693	1754	似	0.004467	2719	閨	0.000893
790	論	0.027693	1755	慣	0.004467	2720	漪	0.000893
791	堡	0.027693	1756	氓	0.004467	2721	頸	0.000893
792	除	0.027693	1757	罷	0.004467	2722	于	0.000893
793	救	0.027693	1758	阻	0.004467	2723	葯	0.000893
794	微	0.027693	1759	姐	0.004467	2724	釀	0.000893
795	獨	0.027693	1760	協	0.004467	2725	坏	0.000893
796	啦	0.027693	1761	呷	0.004467	2726	苛	0.000893
797	竹	0.027693	1762	淚	0.004467	2727	腰	0.000893
798	整	0.0268	1763	蕩	0.004467	2728	裸	0.000893
799	些	0.0268	1764	但	0.004467	2729	鏹	0.000893
800	闊	0.0268	1765	漸	0.004467	2730	餃	0.000893
801	近	0.0268	1766	繡	0.004467	2731	丙	0.000893
802	踢	0.0268	1767	矮	0.004467	2732	甕	0.000893
803	織	0.0268	1768	皂	0.004467	2733	聆	0.000893
804	奏	0.0268	1769	忍	0.004467	2734	蕘	0.000893
805	妙	0.0268	1770	疏	0.004467	2735	紬	0.000893
806	正	0.0268	1771	肅	0.004467	2736	謁	0.000893
807	先	0.0268	1772	徐	0.004467	2737	諂	0.000893
808	限	0.0268	1773	壤	0.004467	2738	揭	0.000893
809	回	0.0268	1774	橫	0.004467	2739	骯	0.000893
810	划	0.0268	1775	督	0.004467	2740	濮	0.000893
811	賠	0.0268	1776	薩	0.004467	2741	勃	0.000893
812	二	0.0268	1777	填	0.004467	2742	瞞	0.000893
813	溫	0.0268	1778	裂	0.004467	2743	蹲	0.000893
814	汗	0.0268	1779	廁	0.004467	2744	猿	0.000893
815	良	0.0268	1780	盛	0.004467	2745	飄	0.000893
816	掃	0.0268	1781	蝶	0.004467	2746	嘲	0.000893
817	姿	0.0268	1782	翼	0.004467	2747	巫	0.000893
818	南	0.025907	1783	枯	0.004467	2748	釘	0.000893
819	孩	0.025907	1784	馳	0.004467	2749	閨	0.000893
820	守	0.025907	1785	乒	0.004467	2750	瀑	0.000893
821	綁	0.025907	1786	披	0.004467	2751	廓	0.000893
822	寓	0.025907	1787	晰	0.004467	2752	糊	0.000893
823	準	0.025907	1788	蜜	0.004467	2753	鳴	0.000893
824	頓	0.025907	1789	砍	0.004467	2754	睹	0.000893
825	湖	0.025907	1790	悟	0.004467	2755	坊	0.000893
826	爆	0.025907	1791	繼	0.004467	2756	儼	0.000893
827	懂	0.025013	1792	吻	0.004467	2757	趾	0.000893
828	啊	0.025013	1793	情	0.004467	2758	葩	0.000893
829	份	0.025013	1794	誤	0.004467	2759	塵	0.000893
830	慢	0.025013	1795	帖	0.004467	2760	靶	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
831	思	0.025013	1796	傾	0.004467	2761	跪	0.000893
832	坐	0.025013	1797	兵	0.004467	2762	霞	0.000893
833	迪	0.025013	1798	奸	0.004467	2763		0.000893
834	蔬	0.025013	1799	捲	0.004467	2764	娣	0.000893
835	摩	0.025013	1800	遍	0.004467	2765	瑟	0.000893
836	增	0.025013	1801	茫	0.004467	2766	粥	0.000893
837	找	0.025013	1802	卷	0.004467	2767	川	0.000893
838	冒	0.025013	1803	澤	0.004467	2768	汀	0.000893
839	戒	0.025013	1804	柏	0.004467	2769	刨	0.000893
840	乾	0.025013	1805	仁	0.004467	2770	坤	0.000893
841	激	0.025013	1806	翩	0.004467	2771	函	0.000893
842	麻	0.025013	1807	瘋	0.004467	2772	暇	0.000893
843	惡	0.025013	1808	胃	0.004467	2773	瑤	0.000893
844	裸	0.025013	1809	罩	0.004467	2774	纍	0.000893
845	蒙	0.025013	1810	致	0.004467	2775	猖	0.000893
846	值	0.025013	1811	須	0.004467	2776	琦	0.000893
847	善	0.025013	1812	轟	0.004467	2777	妃	0.000893
848	密	0.025013	1813	峰	0.004467	2778	惜	0.000893
849	北	0.025013	1814	嬰	0.004467	2779	婪	0.000893
850	專	0.025013	1815	嘆	0.004467	2780	鈹	0.000893
851	蛋	0.025013	1816	沿	0.004467	2781	篆	0.000893
852	省	0.02412	1817	傘	0.004467	2782	婷	0.000893
853	將	0.02412	1818	左	0.004467	2783	漓	0.000893
854	忙	0.02412	1819	虐	0.004467	2784	雁	0.000893
855	避	0.02412	1820	傑	0.004467	2785	糙	0.000893
856	媒	0.02412	1821	框	0.004467	2786	拷	0.000893
857	儲	0.02412	1822	抒	0.004467	2787	炎	0.000893
858	坦	0.02412	1823	搏	0.004467	2788	芒	0.000893
859	鴨	0.02412	1824	盒	0.004467	2789	卻	0.000893
860	爛	0.02412	1825	挑	0.004467	2790	批	0.000893
861	採	0.02412	1826	按	0.004467	2791	瘍	0.000893
862	弟	0.02412	1827	耗	0.004467	2792	嬉	0.000893
863	刊	0.02412	1828	摔	0.004467	2793	崑	0.000893
864	嚴	0.02412	1829	厝	0.004467	2794	惻	0.000893
865	應	0.02412	1830	僕	0.004467	2795	怡	0.000893
866	欄	0.023227	1831	鑑	0.004467	2796	鑰	0.000893
867	雲	0.023227	1832	聰	0.003573	2797	侯	0.000893
868	谷	0.023227	1833	遜	0.003573	2798	腔	0.000893
869	廟	0.023227	1834	汜	0.003573	2799	憑	0.000893
870	珠	0.023227	1835	忱	0.003573	2800	斂	0.000893
871	歐	0.023227	1836	斧	0.003573	2801	紮	0.000893
872	吧	0.023227	1837	釋	0.003573	2802	贊	0.000893
873	邊	0.023227	1838	鈺	0.003573	2803	偕	0.000893
874	雨	0.023227	1839	翠	0.003573	2804	辱	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
875	據	0.023227	1840	辯	0.003573	2805	舩	0.000893
876	任	0.023227	1841	褶	0.003573	2806	漣	0.000893
877	鄉	0.023227	1842	罵	0.003573	2807	潰	0.000893
878	效	0.023227	1843	靴	0.003573	2808	諦	0.000893
879	樣	0.023227	1844	疾	0.003573	2809	繹	0.000893
880	構	0.023227	1845	措	0.003573	2810	鑲	0.000893
881	祭	0.023227	1846	惟	0.003573	2811	綴	0.000893
882	格	0.023227	1847	肌	0.003573	2812	圃	0.000893
883	配	0.023227	1848	町	0.003573	2813	肩	0.000893
884	灌	0.022333	1849	附	0.003573	2814	峽	0.000893
885	斑	0.022333	1850	楓	0.003573	2815	僵	0.000893
886	或	0.022333	1851	紋	0.003573	2816	螞	0.000893
887	綜	0.022333	1852	詢	0.003573	2817	搬	0.000893
888	還	0.022333	1853	噁	0.003573	2818	逍	0.000893
889	螢	0.022333	1854	枝	0.003573	2819	溯	0.000893
890	助	0.022333	1855	押	0.003573	2820	嘔	0.000893
891	貼	0.022333	1856	尊	0.003573	2821	怠	0.000893
892	齊	0.022333	1857	呢	0.003573	2822	旱	0.000893
893	耘	0.022333	1858	疼	0.003573	2823	膨	0.000893
894	哈	0.022333	1859	晨	0.003573	2824	願	0.000893
895	颯	0.022333	1860	岩	0.003573	2825	耶	0.000893
896	適	0.022333	1861	夕	0.003573	2826	瞳	0.000893
897	判	0.022333	1862	攀	0.003573	2827	緯	0.000893
898	私	0.022333	1863	螂	0.003573	2828	賀	0.000893
899	吉	0.022333	1864	坎	0.003573	2829	擲	0.000893
900	裙	0.022333	1865	蟑	0.003573	2830	帕	0.000893
901	呀	0.02144	1866	澆	0.003573	2831	阪	0.000893
902	望	0.02144	1867	叔	0.003573	2832	挪	0.000893
903	梵	0.02144	1868	撩	0.003573	2833	攸	0.000893
904	吵	0.02144	1869	訂	0.003573	2834	啣	0.000893
905	笛	0.02144	1870	耀	0.003573	2835	痿	0.000893
906	後	0.02144	1871	遞	0.003573	2836	截	0.000893
907	租	0.02144	1872	禽	0.003573	2837	紊	0.000893
908	央	0.02144	1873	渾	0.003573	2838	禾	0.000893
909	陀	0.02144	1874	髦	0.003573	2839	隕	0.000893
910	算	0.02144	1875	厚	0.003573	2840	慮	0.000893
911	筒	0.02144	1876	您	0.003573	2841	苜	0.000893
912	泰	0.02144	1877	汐	0.003573	2842	碎	0.000893
913	獄	0.02144	1878	叉	0.003573	2843	蓆	0.000893
914	秧	0.02144	1879	掌	0.003573	2844	庇	0.000893
915	械	0.02144	1880	喪	0.003573	2845	斤	0.000893
916	前	0.02144	1881	諜	0.003573	2846	薑	0.000893
917	墨	0.02144	1882	奪	0.003573	2847	寢	0.000893
918	棚	0.02144	1883	冊	0.003573	2848	凳	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
919	需	0.02144	1884	催	0.003573	2849	垣	0.000893
920	董	0.020547	1885	滔	0.003573	2850	唬	0.000893
921	培	0.020547	1886	涵	0.003573	2851	鋪	0.000893
922	悲	0.020547	1887	夭	0.003573	2852	祥	0.000893
923	簿	0.020547	1888	兜	0.003573	2853	蚤	0.000893
924	討	0.020547	1889	乎	0.003573	2854	叩	0.000893
925	肯	0.020547	1890	乖	0.003573	2855	碘	0.000893
926	鬧	0.020547	1891	幅	0.003573	2856	璘	0.000893
927	異	0.020547	1892	斥	0.003573	2857	初	0.000893
928	巧	0.020547	1893	橡	0.003573	2858	蜃	0.000893
929	欣	0.020547	1894	敲	0.003573	2859	戎	0.000893
930	練	0.020547	1895	棄	0.003573	2860	簣	0.000893
931	剛	0.020547	1896	撼	0.003573	2861	干	0.000893
932	充	0.020547	1897	焚	0.003573	2862	睏	0.000893
933	極	0.020547	1898	僚	0.003573	2863	纏	0.000893
934	勢	0.020547	1899	拘	0.003573	2864	唐	0.000893
935	早	0.020547	1900	蚰	0.003573	2865	虞	0.000893
936	鼓	0.020547	1901	暈	0.003573	2866	黛	0.000893
937	吸	0.020547	1902	鬚	0.003573	2867	鞏	0.000893
938	見	0.020547	1903	撈	0.003573	2868	悱	0.000893
939	解	0.020547	1904	堵	0.003573	2869	邱	0.000893
940	透	0.020547	1905	禧	0.003573	2870	鴛	0.000893
941	言	0.020547	1906	鍾	0.003573	2871	莠	0.000893
942	痛	0.020547	1907	堅	0.003573	2872	洪	0.000893
943	編	0.020547	1908	藤	0.003573	2873		0.000893
944	紡	0.019653	1909	途	0.003573	2874	癢	0.000893
945	露	0.019653	1910	蠻	0.003573	2875	泮	0.000893
946	宇	0.019653	1911	鐘	0.003573	2876	稿	0.000893
947	騙	0.019653	1912	伴	0.003573	2877	渠	0.000893
948	飼	0.019653	1913	纖	0.003573	2878	塚	0.000893
949	京	0.019653	1914	沖	0.003573	2879	喚	0.000893
950	備	0.019653	1915	樑	0.003573	2880	熬	0.000893
951	托	0.019653	1916	碰	0.003573	2881	缸	0.000893
952	鳥	0.019653	1917	棺	0.003573	2882	暑	0.000893
953	床	0.019653	1918	琉	0.003573	2883	梗	0.000893
954	酪	0.019653	1919	伸	0.003573	2884	蹋	0.000893
955	其	0.019653	1920	番	0.003573	2885	昱	0.000893
956	鞭	0.019653	1921	鑼	0.003573	2886	坑	0.000893
957	圓	0.019653	1922	患	0.003573	2887	髀	0.000893
958	形	0.019653	1923	逆	0.003573	2888	逗	0.000893
959	反	0.019653	1924	偵	0.003573	2889	嗑	0.000893
960	參	0.019653	1925	牆	0.003573	2890	毀	0.000893
961	約	0.019653	1926	吐	0.003573	2891	敏	0.000893
962	候	0.019653	1927	杆	0.003573	2892	淆	0.000893

序號	字	頻次	序號	字	頻次	序號	字	頻次
963	辦	0.019653	1928	吊	0.003573	2893	惘	0.000893
964	豆	0.019653	1929	概	0.003573			
965	潮	0.019653	1930	襲	0.003573			