國立臺灣大學電機資訊學院資訊工程研究所

碩士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

古籍影像與文本之對應－以《古今圖書集成》為例

Mapping between Images and Texts of *Completed Collection of Graphs and Writings of Ancient and Modern Times*

陳冠仲

Kuan-Chung Chen

指導教授：項潔 教授

Advisor: Jieh Hsiang, Ph.D

中華民國 104 年 7 月

July, 2015

誌謝

四年的碩士生涯，一路走來跌跌撞撞有些漫長，中間經歷了休學、當兵；復學後加入了項潔老師的實驗室，非常高興也非常幸運能加入 303 實驗室，在 303 的兩年碩士結束後回頭看，更覺得當初做了正確的決定。303 就像個大家庭，彼此間互相幫助、成長，兩年來受到了許多人的幫助。

首先當然要感謝項潔老師的指導，除了在學業上給了我明確的研究方向，並且時常分享許多課本上學習不到的知識，教導我們要去「思考問題」，再進一步去「解決問題」；同時在我生活上遭遇困難時給了許多鼓勵，兩年下來實在獲益良多。

再來要感謝兩位口試委員謝育平教授以及蔡宗翰教授，不僅指正了我論文上的許多缺失，更給了明確的改正方向，在此再次感謝二位的用心。

這篇論文能完成還要感謝許多人的幫忙。感謝 303 實驗室的杜協昌博士，博士在實務上的經驗豐富，所以總是想的比別人深比別人遠，常常能看見我們沒發現的問題；在技術上也給了相當多的指導。熱心的農堯學長除了分享 GIS 的相關研究之外，也給了我許多與我論文相關的參考資料。稷安學長與宋浩學長是實驗室的大家長，幫忙處理了實驗室許多大小問題，也感謝稷安學長透過人文學者的身分給了許多不同角度的意見；感謝宋浩學長分享了許多相關技術與實作開發的經驗。感謝上一屆畢業的沛強，告訴我實驗室的大小事，讓我能快速融入這個環境。還要感謝開心果小黑、一起畢業的好夥伴秀萍。

最後要感謝我的家人，即使我的求學過程比別人漫長，他們仍舊支持我，讓我能夠專心在自己的學業上，在此致上最高的謝意。

陳冠仲　民國 104 年 7 月

中文摘要

《古今圖書集成》為現存最大類書，因此有不少數位人文學者將其與資料庫系統結合，做成《古今圖書集成》全文檢索系統，內容大多包含文字及影像的搜索功能，但在結果的呈現上皆重於文字，對影像的部分並無多加著墨，所以當使用者想從影像中獲取一些資訊，例如找某個關鍵字詞時，只能用肉眼觀察影像的內容，無法從系統提供幫助。

在本研究中，試圖避開 OCR 技術的輔助，直接對影像及文本處理，讓兩者間有高度的對應關係，再利用文本來尋找文字在影像中的位置。首先對所有影像做一些影像處理，包含了旋轉與切割，使每張影像有著相同的格式與排版，再分析影像特性，如：文字的排版方式、影像中圖像有固定大小與位置等等，利用這些特性以行為單位將影像的狀態完整對應到文本中，最後文本每一行對應到影像中文字、空行、圖像三種狀態其一。

最後再利用對應完成的文本及處理過的影像，先計算文字在文本中的位置，再透過對應座標的方式找出文字在影像中的位置。如此使得《古今圖書集成》影像將不再只是以插圖的形式點綴系統，而是能實際提供有用的資訊給使用者。

關鍵字：古今圖書集成、數位人文、影像處理

ABSTRACT

The *Complete Collection of Graphs and Writings of Ancient and Modern Times* (*Gujintushujicheng*, or *Jicheng* for short), completed in the early 18th century, is the largest book in the world in existence. Containing over one million Chinese characters, almost 100,000 pages, and cover over 6,000 subjects, *Jicheng* is also difficult to use. During the past decade, several digital systems have been developed so that people can use *Jicheng* through fulltext search. However, all of these system did not attempt to match images and texts, which would make using *Jicheng* even easier. This difficult arises partly because for old Chinese books, OCR is still not an effective technology.

In this thesis we develop a method that tries to find direct correspondence between an image of *Jicheng* and its associated text without resorting to OCR. We first calibrate the images so that all 100,000 pages in the book have the same size and format. We then analyze the characteristics such as the format, number of lines, position of graphs, etc, so that each line in the typed text maps to either a line of text, a blank line, of part of a graph in a page image. Once this is done, we then do a character-by-character mapping between each character in the typed text and a character in a page image.

Our method is quite effective. The accuracy in mapping the entire contain of *Jicheng* is 98,7%. The rest is mainly due to typographic errors occurred when typing the full text, which can be easily corrected by hand.

Keywords: *Gujintushujicheng*, Digital Humanities, Image Processing

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1　緒論

1.1　研究背景

《古今圖書集成》(以下皆簡稱《集成》)是現存最大的類書，因原中國最大類書《永樂大典》幾乎已燬，正本不知去向，而副本僅存原書的 4% [1]。《集成》不僅規模大且體例完整[2]，因此不少數位人文學者將其與資料庫系統結合，做成《集成》的全文檢索系統，例如廣西大學開發的「古今圖書集成索引&全書圖像[3]」，東吳大學與國立故宮博物院所共同開發的「數位古今圖書集成[4]」，以及台大資工數位典藏與自動推論實驗室的 THDL-based 古今圖書集成[5]，以上系統皆提供了文字或圖像的搜索功能，但在搜索結果的呈現上皆只著重於文字部分。

以 THDL-based 古今圖書集成來說明，(a.)輸入關鍵字「赤壁」搜索後，系統會列出出現關鍵字的(b.)條目內文列表，而每個條目右方會顯示(c.)與其對應的影像

Fig. 1-1　THDL-based 古今圖書集成

（條目在書本中出現的頁面）。在內文的部分可以清楚的看到關鍵字「赤壁」被標註成黃色底色，但在右方影像中就僅僅只是一張《集成》影像，若想要在影像中找尋「赤壁」的位置並沒有提供特別的方法，只能靠肉眼搜尋。

若是想要由系統來提示關鍵字在影像中的位置，最直覺的方法便是使用 OCR（**O**ptical **C**haracter **R**ecognition）光學字符識別，但由於中文字每個字的間距固定，不像歐美文字 Words 的長度不一，因此僅能從字體作判斷，且中文字的複雜度非常高，因此中文 OCR 的辨識率一直都不高 [6]；再考慮到古籍的影像保存不易，經過影印、掃描清晰度又大打折扣，因此中文 OCR 並不適合作為系統使用。

1.2 研究目的

隨著數位人文領域的發展，越來越多的古籍被建立成數位化的文本及書本頁面的影像（可能為照片檔或掃描檔），現今這些古籍的資料庫系統多著重在文本的呈現上，做了許多後分類、自動斷詞等等的處理，但影像的部分往往只是以圖片的形式貼在一旁，以《集成》影像為例，一張影像內文字動輒上千字，想從影像中找到自己想要的資訊需要花很大的力氣，本研究希望能改善使用者與影像的互動性，提升這些古籍影像在系統中的意義。

本研究最終目的在於：不透過 OCR 的輔助，而是利用與影像高度對應的文本來獲得影像中的文字資訊。本研究以《集成》為例，希望透過資訊技術自動化分析並處理影像與文本資訊，最後生成一組格式統一的影像（排版、影像傾斜度）與一組和影像高度對應的文本檔案，此檔案不僅具備文本資訊（文字、空行、圖像）也包含了影像資訊（內文對應的影像檔名），透過此文本檔案可清楚的還原文字在影像中的位置，包括第幾行、第幾列、中間有無空行、圖像等等。有了這些

資訊便可知道文字在影像中對應的位置，再加上處理過後格式統一的影像，便可在影像中標註關鍵字，幫助使用者從影像中找到想要的資訊。

1.3 相關研究

1.3.1 廣西大學—「古今圖書集成索引&全書圖像」

網址：http://gjtsjc.gxu.edu.cn/

Fig. 1-2 「古今圖書集成索引&全書圖像」查詢方式

「古今圖書集成索引＆全書圖像」是由廣西大學古籍所開發完成，部名的查詢方法有三種：

1. 按原書的經線三級分類(彙考→典→部)查找。即點擊選擇左邊列表項中所列出的 6 個彙編名下的的典名，中間的空白框將出現該典所轄的所有部名。
2. 在上方“部名”的空白框中，直接輸入部名，再點擊“查詢”鍵，中間的空白框出現部名。
3. 點擊“[由古異名或今名查詢部]”，彈出一個包含古今參見小窗體。古今參見搜集了某些部具有的異名或相應的現代概念的內容。從中找到所需的部名。

通過上述方法找到的部名，都列表顯示在中間空白框。點擊框內的部名（若有），將得到該部名所轄緯目在書中的位置。結果顯示在“查詢結果”的框內。框內顯示出該部的“彙考”、“總論”、“列傳”、“藝文”、“選句”、“紀事”、“雜録”、“外編”等緯目在精裝本和綫裝本中的起始册頁欄塊行碼。並非每一部都具有所有的緯目，某部没有的緯目，該緯目顯示“此部無這一緯目。

「古今圖書集成索引＆全書圖像」影像的呈現方式則是將影像以欄為單位切割成三欄，顯示的影像為其中一欄，並註明查找的部名是在第幾欄的第幾塊之第幾列。如圖 Fig. 1-3 所示：

Fig. 1-3 「古今圖書集成索引＆全書圖像」影像呈現方式

「古今圖書集成索引＆全書圖像」雖然有全書圖像，但並沒有全文的資訊，所以僅能根據目錄查詢至部名，無法再往下查詢。且查詢結果僅以影像呈現，資料並無經過分析處理，影像的「塊」與「列」意義不明，對於使用者來說只是將紙本閱讀轉換成電子檔閱讀，並無太大區別。

1.3.2　故宮&東吳—「數位古今圖書集成」

網址：http://192.83.187.228/gjtsnet/index.htm

Fig. 1-4　「數位古今圖書集成」

查詢方式：在首頁中可選擇簡易搜尋或進階搜尋（僅有全文，無影像）

◆　簡易搜尋

Fig. 1-5　「數位古今圖書集成」之簡易搜尋

1. 核選資料庫：簡易搜尋中之資料庫核選以六大彙編下之『三十二典』為單位，您可選擇其一加以搜尋。

2. 輸入檢索字串：資料庫核定後則可『輸入檢索字串』，於空欄中填入欲檢索之關鍵字詞。

3. 檢索結果以『頁』為單位

例：欲檢索「臺灣」相關之地理資料—

可於核選資料庫選擇【山川典（方輿彙編）】，另於檢索字串中鍵入【臺灣】，最後點選『開始搜尋』即可進行資料查詢。

◆ 進階搜尋

Fig. 1-6 「數位古今圖書集成」之進階搜尋

1. 核選資料庫：進階搜尋中資料庫核選同樣以『三十二典』為單位。另可複選資料庫（按住 Ctrl 鍵，以滑鼠左鍵選擇不同資料庫），提供跨資料庫檢索功能。

2. 核選內容分項：選擇文章內容體例。(可複選)

3. 核選搜尋邏輯：可選擇是否運用布林邏輯「且」and、「或」or 組合檢索字串。

4. 輸入檢索字串：於空欄中填入欲檢索之關鍵字詞，如配合布林邏輯則需於兩字串間空 1byte。

5. 檢索結果以『段』為單位

例：如欲檢索與『太極圖』相關之資料—

可於核選資料庫選擇【曆相彙編 • 三典】全部選取，勾選所有內容分項，另於檢索字串中鍵入【太極圖】，最後點選『開始搜尋』即可進行資料查詢。

檢索結果會於內文上方以異色字體標示該頁（段）文章所屬之彙編/典/卷/部，以及其體例、頁碼等書目資料，並可點選「前頁資料」及「後頁資料」之功能鍵瀏覽前、後三頁之內容。

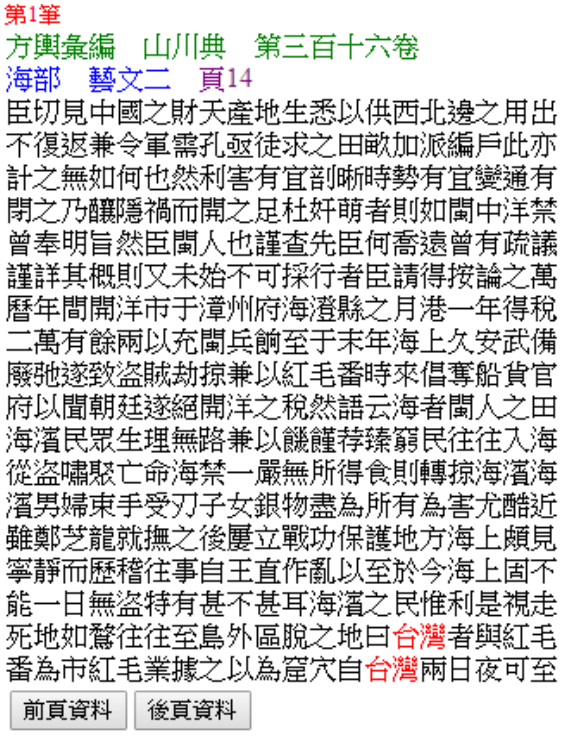
第1筆
方輿彙編　山川典　第三百十六卷
海部　藝文二　頁14
臣切見中國之財天產地生悉以供西北邊之用出不復返兼令軍需孔亟徒求之田畝加派編戶此亦計之無如何也然利害有宜剖晰時勢有宜變通有閉之乃釀隱禍而開之足杜奸萌者則如閩中洋禁曾奉明旨然臣閩人也謹查先臣何喬遠曾有疏議謹詳其概則又未始不可採行者臣請得按論之萬曆年間開洋市于漳州府海澄縣之月港一年得稅二萬有餘兩以充閩兵餉至于末年海上久安武備廢弛遂致盜賊劫掠兼以紅毛番時來倡奪船貨官府以聞朝廷遂絕開洋之稅然語云海者閩人之田海濱民眾生理無路兼以饑饉荐臻窮民往往入海從盜嘯聚亡命海禁一嚴無所得食則轉掠海濱海濱男婦束手受刃子女銀物盡為所有為害尤酷近雖鄭芝龍就撫之後屢立戰功保護地方海上頗見寧靜而歷稽往事自王直作亂以至於今海上固不能一日無盜特有甚不甚耳海濱之民惟利是視走死地如鶩往往至島外區脫之地曰台灣者與紅毛番為市紅毛業據之以為窟穴自台灣兩日夜可至
前頁資料　後頁資料

Fig. 1-7 「數位古今圖書集成」之搜索結果

「數位古今圖書集成」與「古今圖書集成索引&全書圖像」相反，無全書圖像但有全文資訊。透過全文檢索可得檢索關鍵字所在之彙編/典/卷/部，以及其體例、頁碼等書目資料，可惜並無原書影像作為對照，僅能以文字方式呈現結果。

1.4 論文架構

本論文在第二章「研究資料介紹」會對《集成》做一些簡介，以及介紹所使用到的資料，包含《集成》的影像和文本以及其特性。第三章「《集成》影像標準化處理」將對影像做旋轉及切割，透過標準化處理以提高影像的一致性。第四章「《集成》文字檔處理」則是將影像中資訊忠實的對應到文本的方法，也利用到了許多第二章提到的影像與文本的特性，對原本的文字檔做了部分保留與部分修改，並補上原來未有的空行資訊。第五章「文字位置計算」是利用第三章的影像與第四章的文字檔來計算出影像中的文字位置。第六章「結論與未來工作」將總結本論文所完成的貢獻以及此研究方法的前提與假設，並探討各種可能的延伸應用。

Chapter 2　研究資料介紹

2.1　《古今圖書集成》

2.1.1　簡介

《古今圖書集成》原名《古今圖書彙編》，是清康熙時期由陳夢雷編輯而成的大型類書[1]。此書共一萬卷，目錄 40 卷，可概分為 6「彙編」：曆象、方輿、明倫、博學、理學、經濟，此六「彙編」可再分為 32「典」，每典又可再細分為 6117「部」[7]，引用書目高達六千多種，為現存最大部的類書[8]。

2.1.2　版本介紹

- 銅字版（1726 年~1728 年），由清內府用銅活字排印成。
- 鉛字版或扁字版（1884 年），以三號扁體字鉛印。
- 同文版或光緒版（1894 年），光緒帝令上海同文書局石印。
- 中華版（1934 年），上海中華書局以雍正銅活字本影印。
- 文星書店版（1964 年），文星書店出版，現今最常見版本。
- 中華巴蜀版（1986 年），中華書局與巴蜀書社合作，並新增《簡明索引》。

文星書店版為現今最常見之版本，所用的原始版本是 1934 年上海中華書局出版的銅模活字版的照相本，文星採用新式影印技術把 50 萬頁的原版縮成 5 萬頁，編成 100 冊，加考訂文字、介紹文字，及新編索引一冊。本論文即以此版本之《集成》建立而成的全文檢索系統作為研究對象。

2.1.3 集成目錄

彙編	典	部	卷
曆象	乾象	21	100
	歲功	43	116
	曆法	6	140
	庶徵	51	188
方輿	昆輿	21	140
	職方	223	1544
	山川	401	320
	邊裔	542	140
明倫	皇極	31	300
	宮闈	15	140
	官常	65	800
	家範	31	116
	交誼	40	120
	氏族	2696	640
	人事	97	112
	閨媛	17	376
博物	藝術	43	824
	神異	70	320
	禽蟲	317	192
	草目	700	320
理學	經籍	66	500
	學行	97	300
	文字	49	260
	字學	24	160
經濟	選舉	29	136
	詮衡	12	120
	食貨	83	360
	禮儀	70	348
	樂律	46	136
	戎政	30	300
	祥刑	26	180
	考工	155	252

Table 1 《集成》目錄

2.2 《集成》數位化影像資料

本研究所使用的《集成》數位化影像為文星書店版《集成》之掃描影像檔，一共 96887 張影像，《集成》一共 808 冊，但此處只收錄了第 7 冊至第 800 冊之內容，不包含 1～6 冊之目錄以及 801～808 冊之考證。

《集成》影像檔名已經過特殊編排，一共 7 碼：

- 前 3 碼表示冊（007~800）
- 4~6 碼表示頁數
- 第 7 碼表示左頁或右頁（1 代表右，2 代表左）

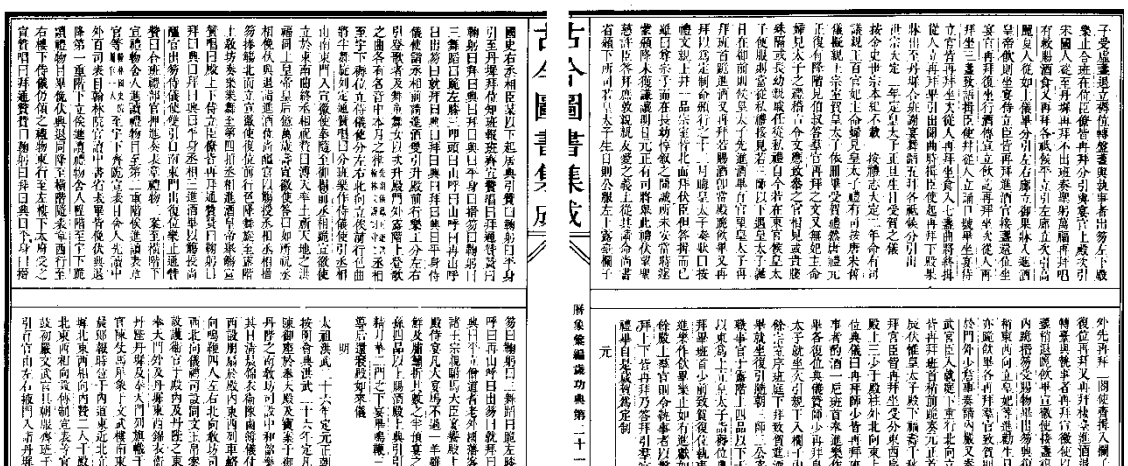

Fig. 2-1 《集成》影像範例 檔名:0170051(右)、0170052(左)

《集成》影像之特性

- 分為上、中、下三個欄位
- 每個欄位皆包含 27 行，一頁共 81 行
- 每行至多 20 字
- 影像中之圖像有固定位置
  - 一圖像固定占據 9 行文字位置
  - 圖像坐落在影像中九宮格其一
  - 以 Fig. 2-2 說明之

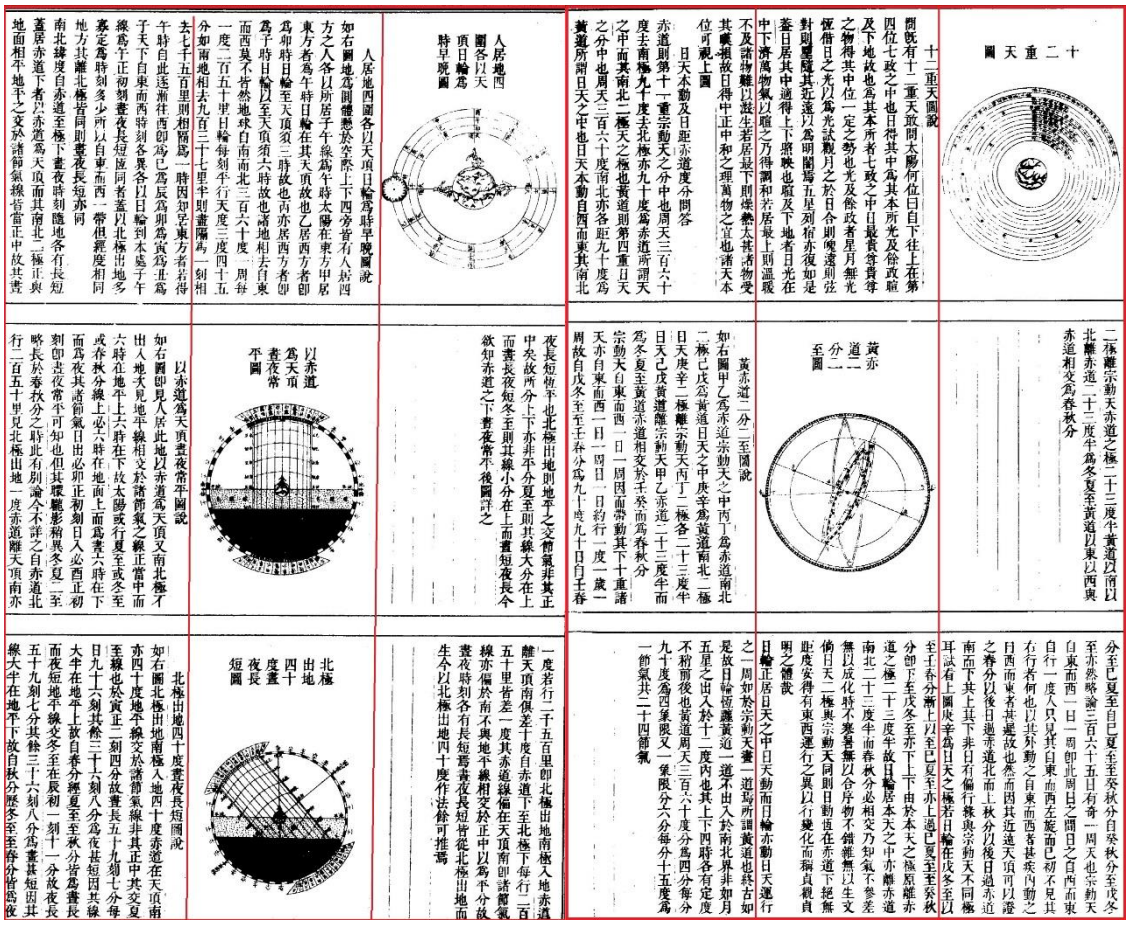

Fig. 2-2 《集成》影像特性

2.3 《集成》數位化文字資料

《集成》全書文字部分皆已數位化成.txt 之文字檔案，編碼方式為 UTF-8，並且按照彙編－典的樹狀結構分類，以卷為單位，《集成》共一萬卷，每卷置於一個檔案，故共計一萬個檔案。其中文字檔的內容每一段皆對應到一張《集成》影像；每一個文字檔會對應到約 10 張影像；每段的第一行以中括號記錄了該段對應之影像檔名。而內文的格式部分，每一行的字數、或是開頭的前置空白、斷句都忠於原書，且若影像中有出現圖像，也會另外以中括號註記該圖像之名稱。

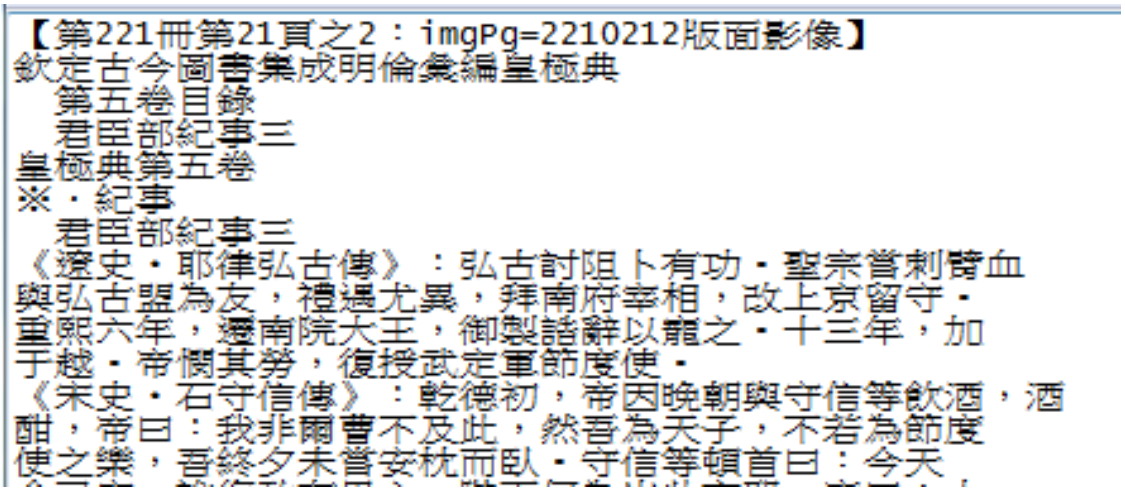

Fig. 2-3 《集成》數位化文字資料

但此文字檔與原書仍有許多不同之處，如：

- 加註了新式的標點符號
- 將標題的部分額外抓出且用「※」標示
- 原書的空行部分在文字檔內並無記錄
- 原書中的稀有字在文字檔內是以字碼的方式呈現，並以兩個「@」住，如：@C08D@

- 原書中小字的部分以小括號括住

Fig. 2-4 為文字檔實例，對照 Fig. 2-5 之原書紙本實例。

Fig. 2-4　　文字檔實例

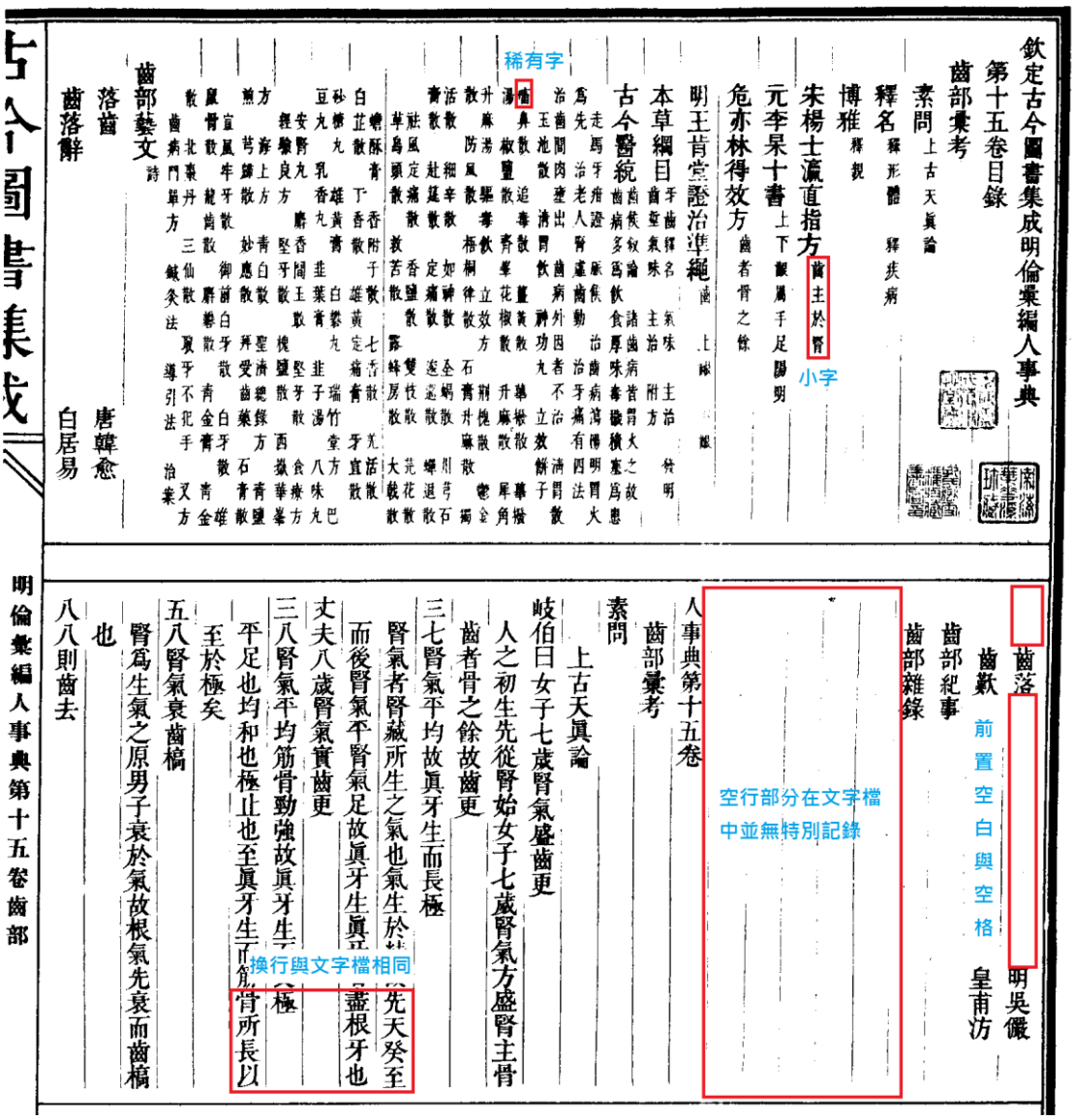

Fig. 2-5　原書紙本實例

因此，為了使文字檔能更接近原書狀態，必須進一步對文字檔作處理，詳細的處理方式在 Chapter 4 會有詳細介紹。

Chapter 3　《集成》影像標準化處理

為了能夠精確的掌握影像中文字的位置，本論文將為《集成》影像做一套標準化處理，目的在於：在不同的影像之間能有相同的文字座標系。以 Fig. 3-1 說明：

Fig. 3-1　《集成》同一頁之左右頁影像

在右頁中找到第一行第一個字為「是」，利用此字之坐標，試圖在左頁找尋找左頁之第一行第一個字，卻發現該位置並無對應文字，原因有兩點：

- 《集成》左頁之文字較靠左，右頁之文字較靠右
- 影像為人工掃描，故會出現影像位置不同、傾斜度不同之情況

為修正此兩種情況，需先將影像之傾斜角度校正，再將文字區塊切割出來，捨棄掉多餘的資訊（書名、書局名稱、頁數等等），如 Fig. 3-2。

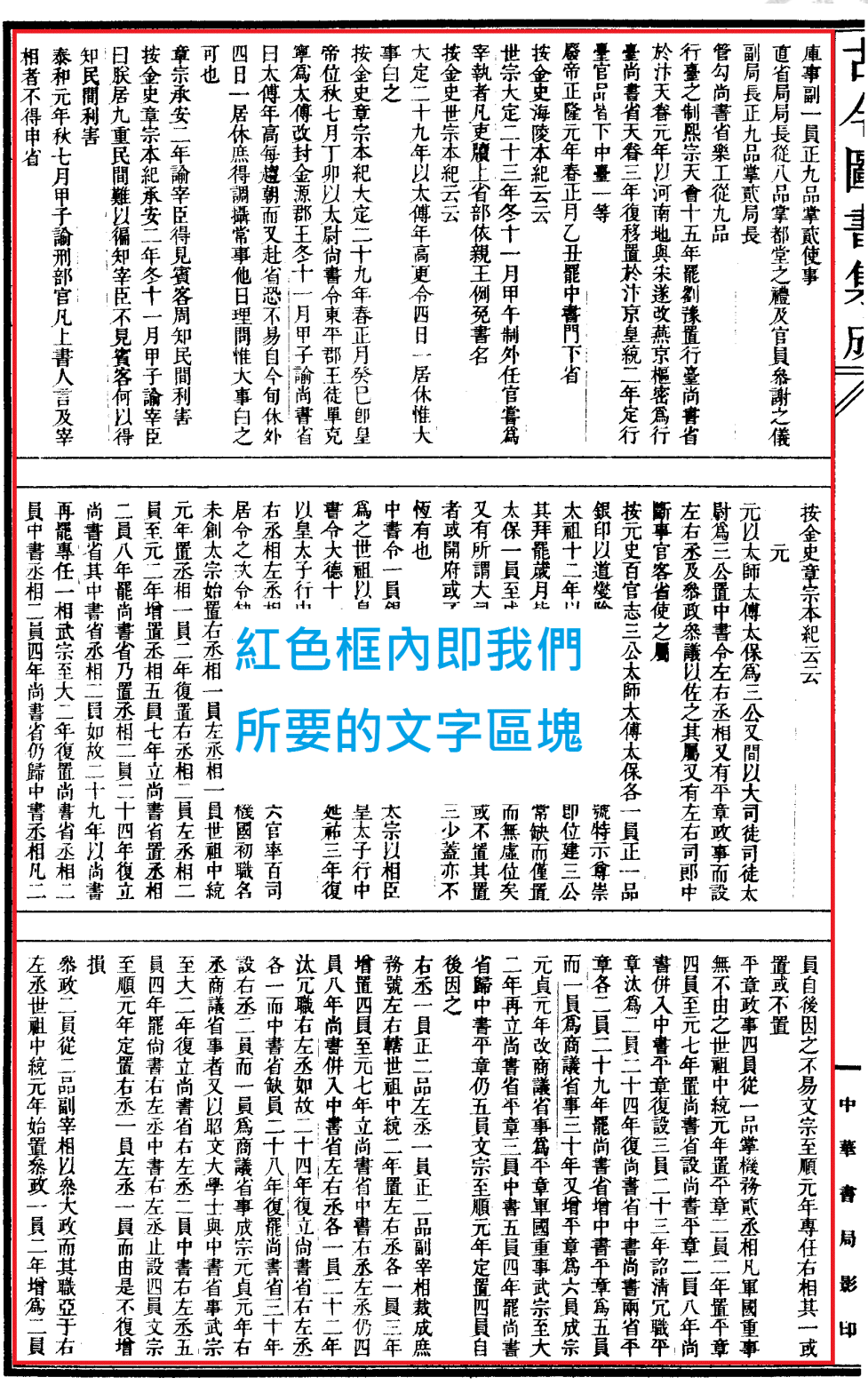

Fig. 3-2 《集成》影像之文字區塊

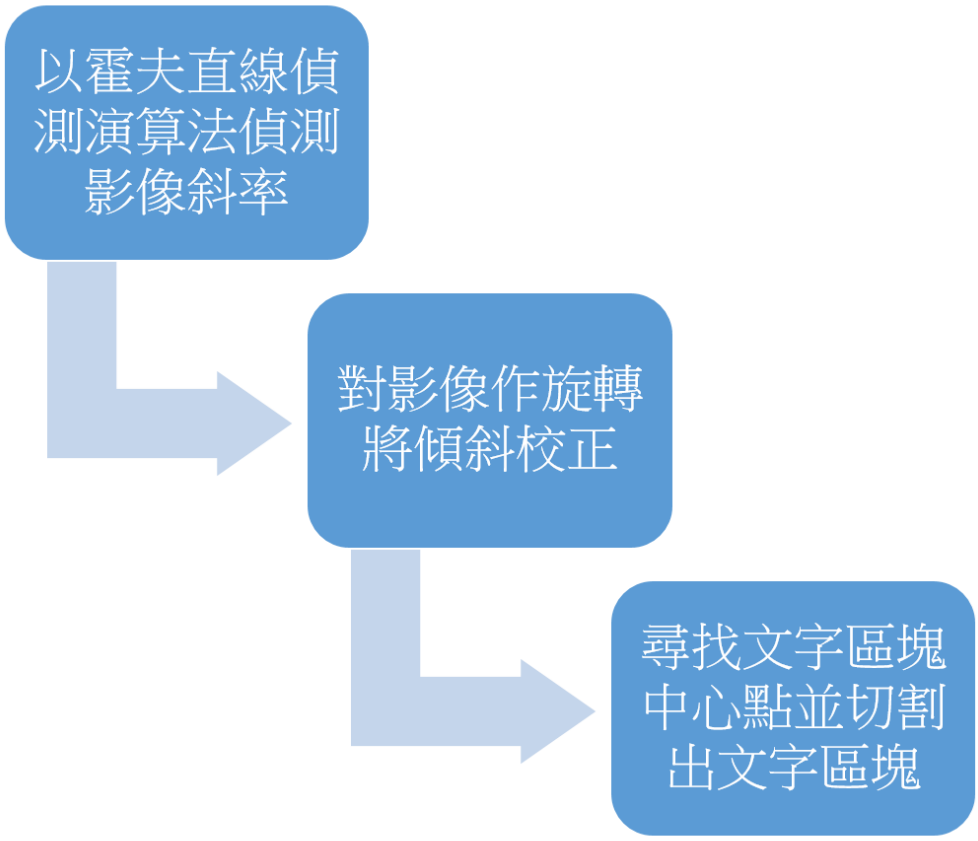

Fig. 3-3 《集成》影像標準化處理流程

3.1 偵測影像斜率

本論文是利用霍夫變換（Hough Transform）[9]來偵測影像中之線段，霍夫變換的概念是：在一個平面直角作標系（x-y）中，一條直線可表示為方程式 $y = kx + b$，而對於直線上的一個點（x_0, y_0）有$y_0 = kx_0 + b$，這表示了在參數平面（k-b）中的一條直線。即，影像中的一個點對應參數平面中的一條直線，而影像中的一條直線對應到參數平面中的一個點，對影像中所有的點作霍夫變換，最後所要偵測的直線一定是對應到在參數平面中直線相交最多的那個點。

OpenCV（Open Source Computer Vision Library）提供了三種不同的霍夫直線偵測方法，分別是：

- CV_HOUGH_STANDARD：代表傳統的霍夫轉換，以(ρ,θ)表示一個線段，ρ 是直線與原點的距離，θ 是直線與 X 軸的角度
- CV_HOUGH_PROBABILISTIC：回傳的線段是影像中的分割線段，而不是整段線段
- CV_HOUGH_MULTI_CALE：多尺度的 CV_HOUGH_STANDARD

由於傳統的霍夫轉換並沒有辦法知道線段的實際位置，所以本研究所使用是 CV_HOUGH_PROBABILISTIC，共有七個參數分別說明之，括號內為本研究所使用之參數值：

- Image：輸入單通道或二值化影像
- Line_storage：儲存偵測結果
- Rho：每條線的鄰近距離（1）
- Theta：允許偵測的角度限制（CV_PI/360）
- Threshold：累積超過閥值的像素點算是一條線（100）
- Param1：最短線段長度（size.width/2）
- Param2：每條被偵測出的直線距離小於此參數則被合併（20）

Rho 設為 1 是希望能盡可能抓出越多線段;最短線段長度設為影像寬度的二分之一，避免將一行文字偵測成直線的可能性，以及避免偵測出影像中文字間的分隔線（印刷上易有誤差）; Param2 只要不會將兩框線合併就好，影響不大。下圖 Fig. 3-4 表示了影像中實際線段被偵測出的情形。

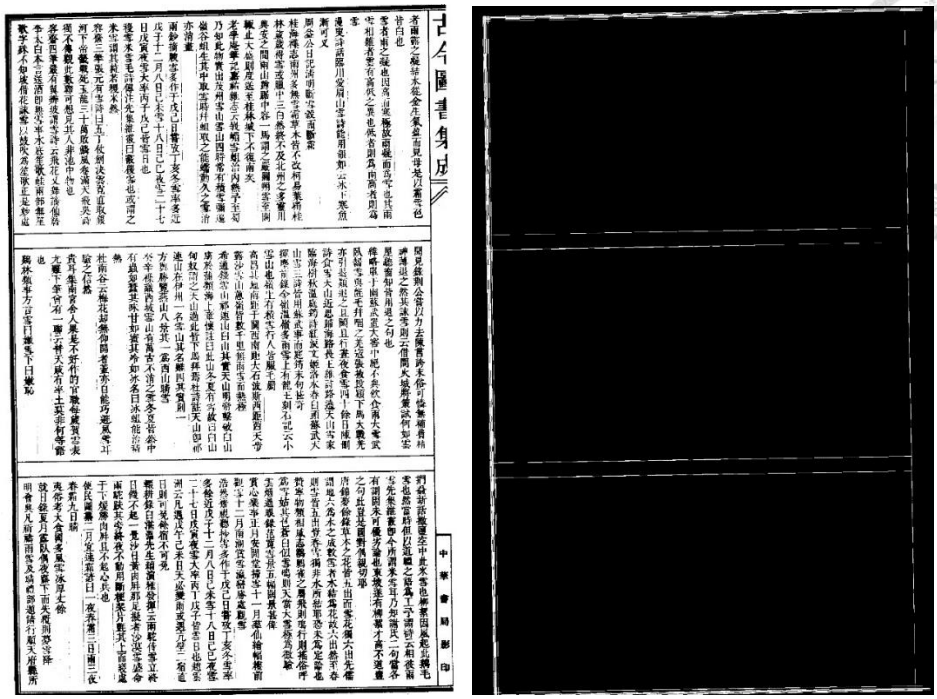

Fig. 3-4　線段偵測實例

透過偵測出的線段，可以很明顯的看出影像的傾斜情況，為了斜率的計算方便，只收集水平方向的線段。又透過觀察發現，最上及最下兩條框線在印刷時很容易有歪曲不直的情形，故此兩條框線在斜率計算上不採記，只採記中間四條框線。

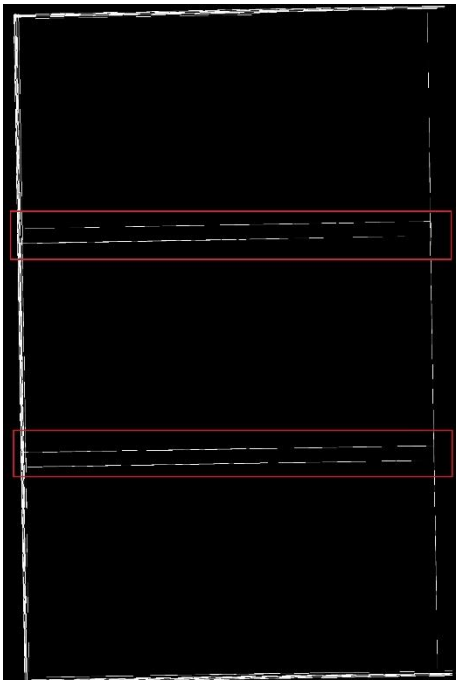

Fig. 3-5　線段偵測實例 2

✧　但有少部分影像（約 500 張）因保存情況不良，影像清晰度較低，中間的線段不易被偵測出，若此情形發生則採記上下之框線。

令採記到的線段與 X 軸的夾角為θ_i，採記到線段數目為n，則影像與 X 軸之夾角 θ 為：

$$\theta = \frac{\sum_1^n \theta_i}{n}$$

3.2　旋轉影像

利用先前算出的影像傾斜度 θ ，生成旋轉矩陣

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) & a \\ \cos(\theta) & \sin(\theta) & b \end{bmatrix}$$

$$a = width * 0.5 \qquad b = height * 0.5$$

Fig. 3-6　旋轉矩陣

其中 a、b 為影像中心點之 x 座標、y 座標，以此為中心點對影像作旋轉的動作，重新計算影像中每個 pixel，得到旋轉後的新影像如 Fig. 3-7。

Fig. 3-7　（左）旋轉前、（右）旋轉後

3.3　文字區塊切割

在已知影像大小都相同的前提下，文字區塊大小已測量出為 1580*2580，剩下的是要找出切割的基準點。對先前校正過斜率的影像使用 OpenCV 中計算最小外

接矩陣的功能 **minAreaRect**，可得到影像整體的輪廓，再利用輪廓的中心點平移至文字區塊的中心點，平移的量為輪廓一半的寬減掉文字區塊一半的寬，如此便可計算出切割基準點，以 Fig. 3-8 說明之。

Fig. 3-8　計算切割基準點示意圖

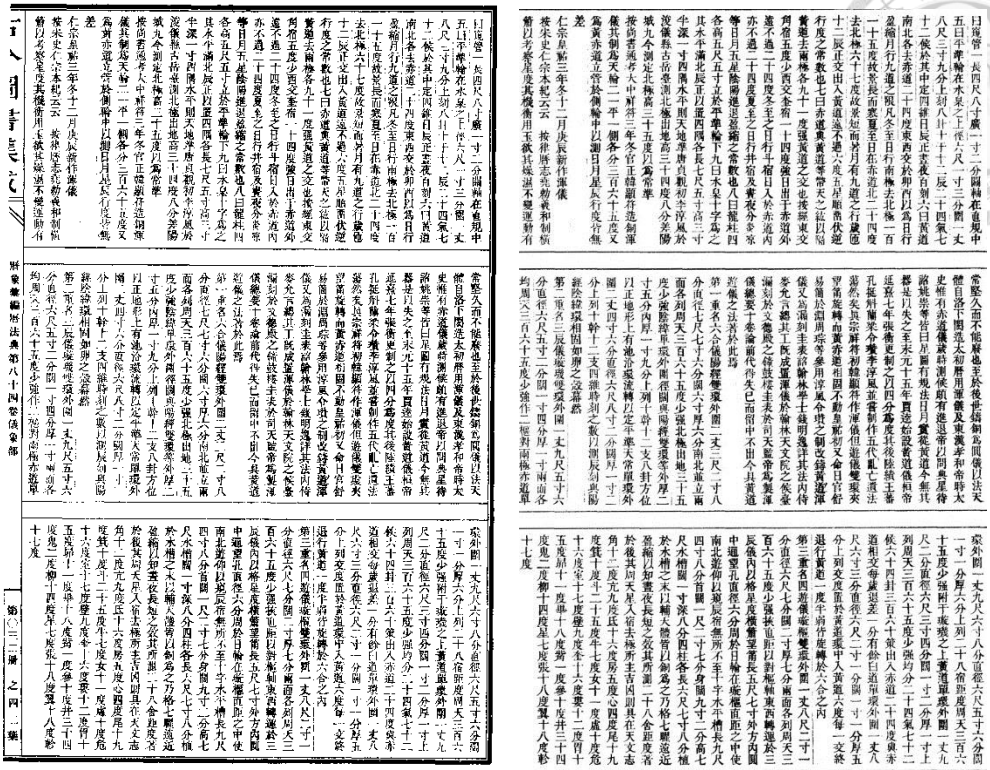

Fig. 3-9　原圖與切割後比較

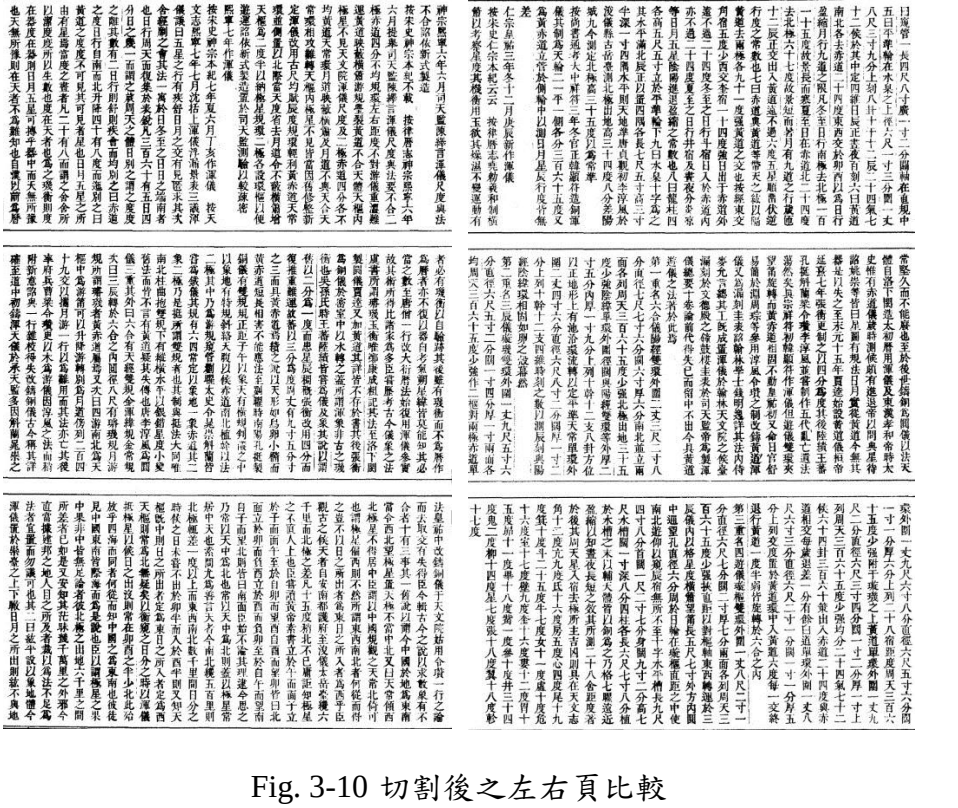

Fig. 3-10 切割後之左右頁比較

透過標準化處理完成後的影像，皆為水平的影像，有一致的斜率，且切割過

後的影像已分不出為左頁還是右頁，每一頁皆有相同的文字座標系。

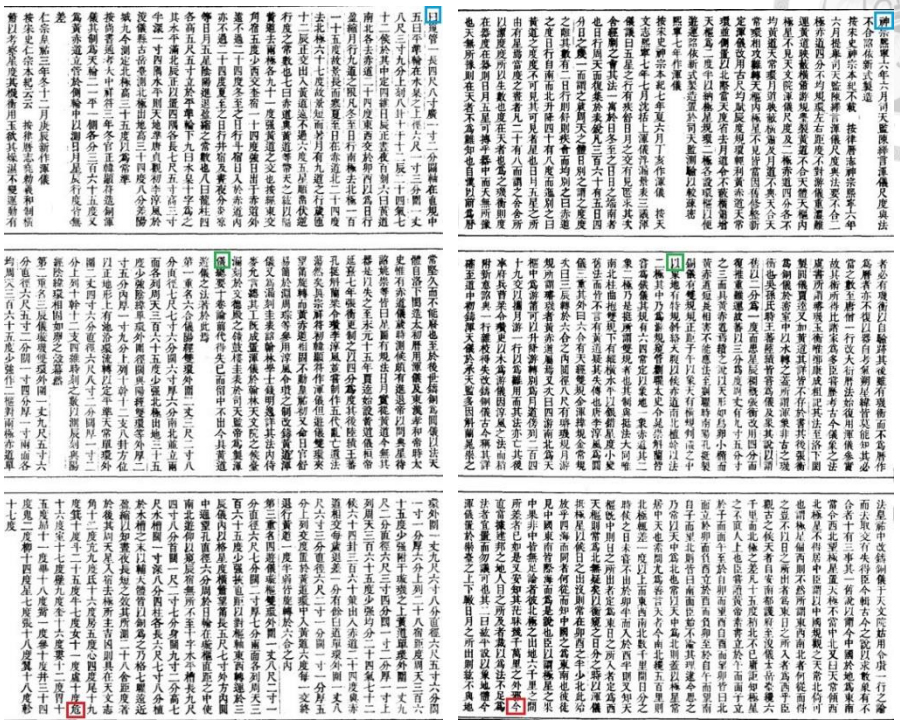

Fig. 3-11 標準化後之影像有相同文字座標系

Chapter 4　《集成》文字檔處理

在 Chapter 3 的部分已介紹過《集成》的數位化文字檔，也說明了文字檔與原書的異同之處，本研究的目的在於利用文字檔來還原原書中文字的狀態，文字檔能越接近原書越好，因此在此章將對文字檔作處理，不屬於原書的部分將之去除，而屬於原書的部分則保留或是利用特殊符號加以註記，最後將原書中有但文字檔中沒有的部分加以補上。

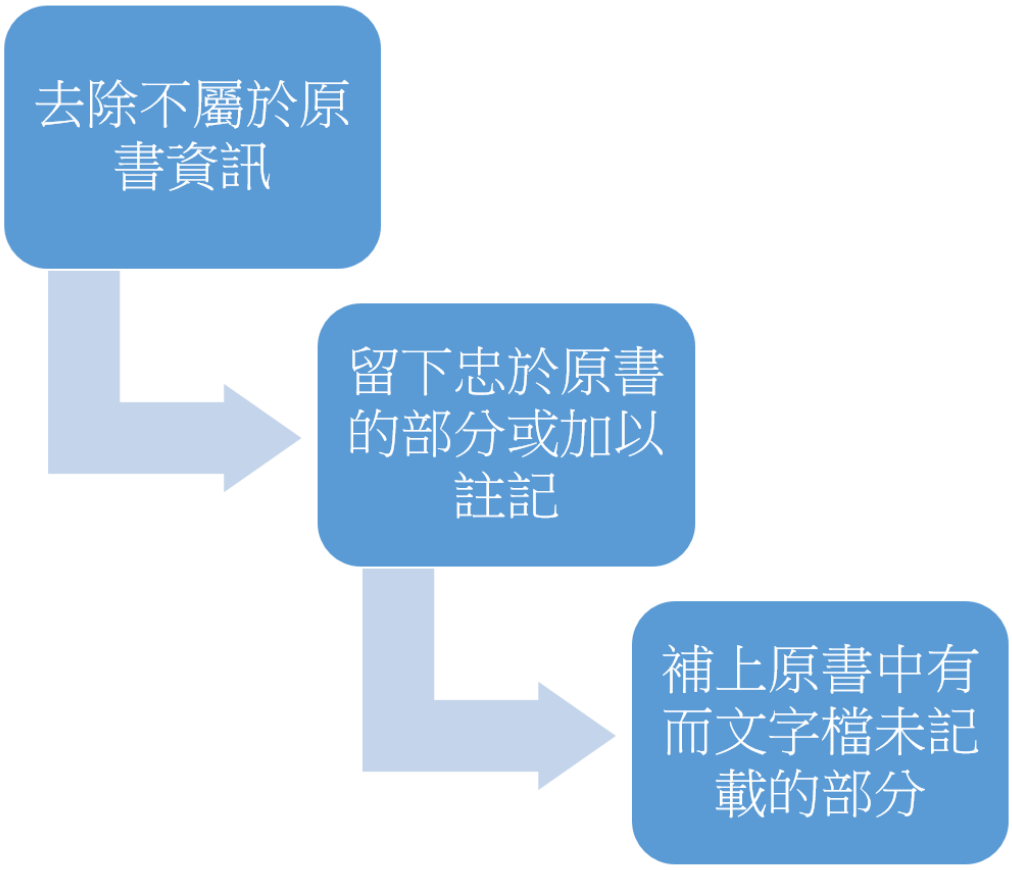

Fig. 4-1　《集成》文字檔處理流程

4.1 去除多餘資訊

文字檔中不屬於原書的部分有：

- 新式標點符號
- 重複的標題，以「※」註記

4.1.1 標點符號去除

標點符號部分：將文字檔中針對內文使用到的標點符號列出，除小括號部分用於記錄小字，故不包含。其中共有句號、問號、驚嘆號、逗號、頓號、分號、冒號、書名號、間隔號十種（書名號有左右符號兩種），將上述標點符號包含全形與半形進行 find and replace，將所有標點符號以空字串（empty string）代替，如此便完成標點符號之去除。

4.1.2 重複標題去除

文字檔中有使用到「※」符號來做註記的只有重複標題的部分，對文字檔以行為單位掃描進行字元比對，若是行中有出現「※」則直接將該行刪除，如此便將文字檔中不屬於原書的部分去除。

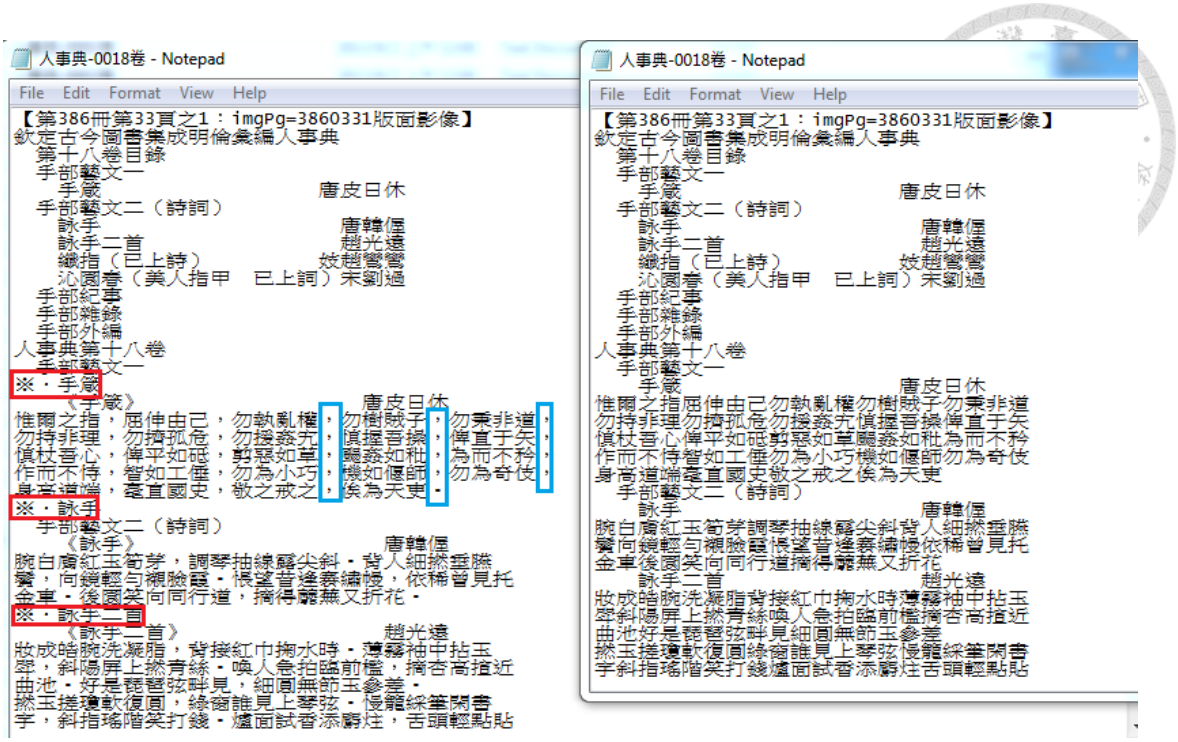

Fig. 4-2　去除多餘資訊後之《集成》文字檔

4.2　保留原書資訊

文字檔中屬於原書的部分有：

- 內文部分
- 小字部分
- 影像檔名
- 稀有字
- 圖像

4.2.1　內文及小字

文字檔內文已經按照原書格式建立，故直接保留；小字部分已以小括號括住，也直接採用原記錄方式保留。

4.2.2　影像檔名處理

影像檔名雖然不是存在於原書中之資訊,但為文字檔與影像對應之重要資訊，故不得去除，原處理方式是以中括號括住影像冊數、頁數與影像檔名，如：【第 386 冊第 33 頁之 1：imgPg=3860331 版面影像】。由於中括號在文字檔其它地方也有使用，且避免其中文字與內文混淆，將檔名部分只留下數字檔名，其於去除。實際做法：每一段之第一行固定為檔名資訊，收集該行所有數字，將末七碼輸出即成。

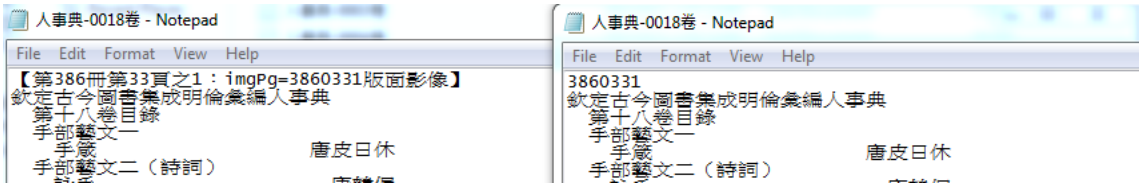

Fig. 4-3　修正後檔名部分

4.2.3　稀有字處理

稀有字在原文字檔內是以字碼方式呈現，並以「@」括住，例：@1826@。此種作法造成每個字的位元數不統一，將會影響到字數上的計算，因此提前處理。實際做法：以行為單位進行字元比對，若是找到「@」則在找到下一個「@」時將兩個「@」與中間的字碼以一個「？」取代。在進行稀有字的處理時發現了一些人工打字的失誤，如：

- 忘了打「@」只打字碼，或是漏打其中一邊的「@」
- 只打「@」卻沒有打中間字碼
- 用了「@」以外的符號，例如：&4343F&、&&3F5C
- 句中有不明符號或英數字

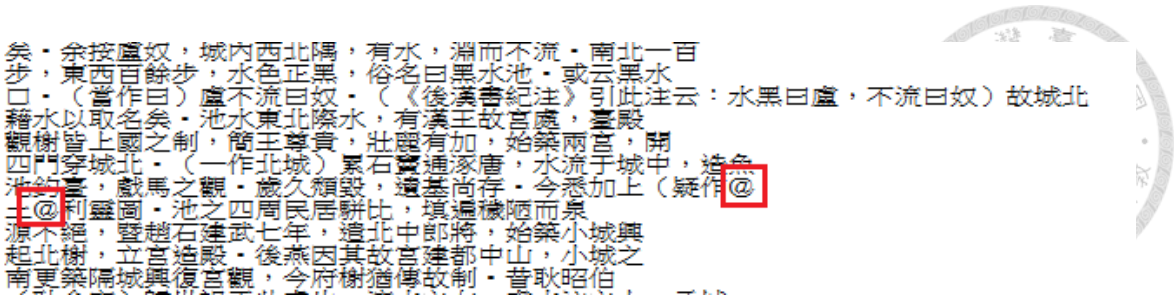

Fig. 4-4　文本錯誤實例 1

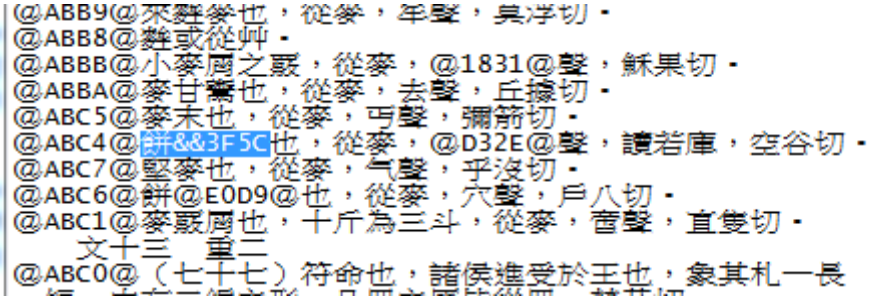

Fig. 4-5　文本錯誤實例 2

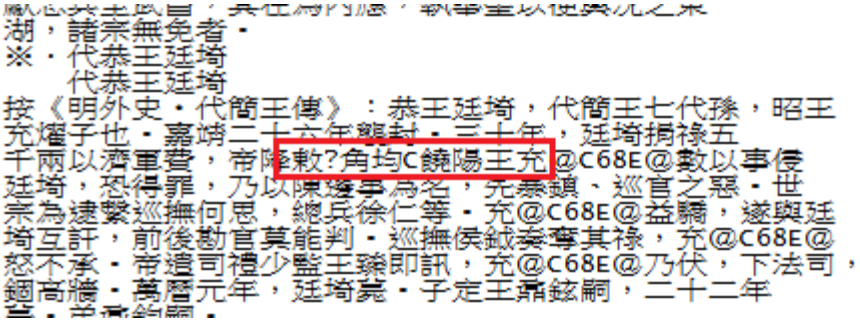

Fig. 4-6　文本錯誤實例 3

凡是在句中發現無法判斷意義之符號或英文字母、數字，統一在該行前以一「$」（dollar sign）加以註記，方便後續人工觀察及處理。

✧　全文處理完成後一共發現 63 處被註記。

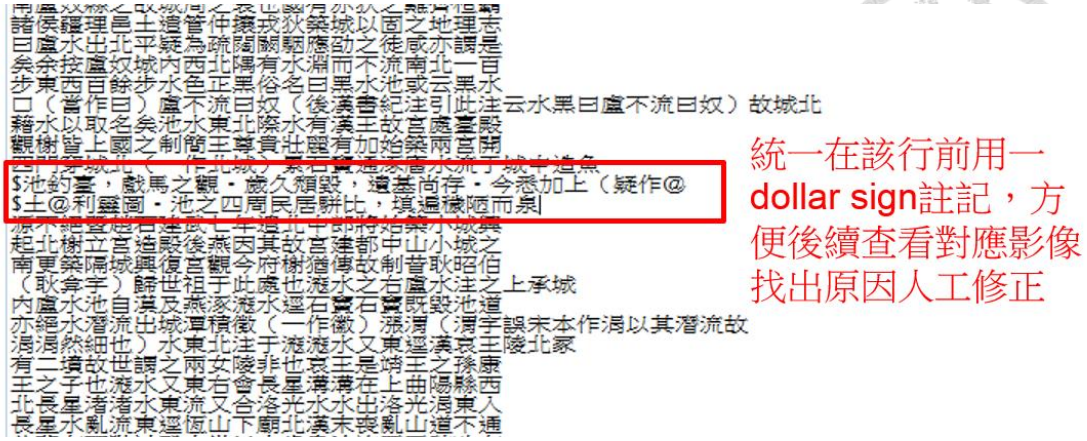

Fig. 4-7　異常文本處理方式

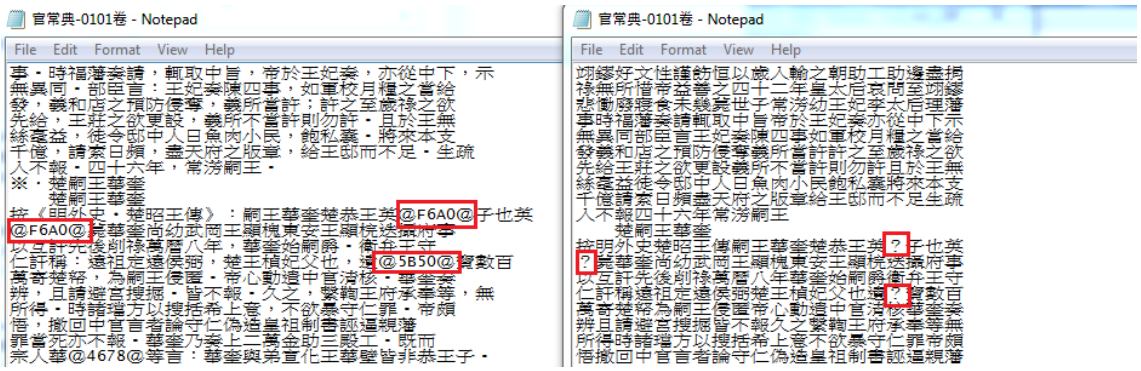

Fig. 4-8　稀有字處理結果

4.2.4　圖像處理

文本中對原書中之圖像的記錄方式是採取：將圖像名稱以一中括號額外表示，再以一行「參考頁面圖像」代表該位置為圖像，一樣以中括號括住。如 Fig. 4-9

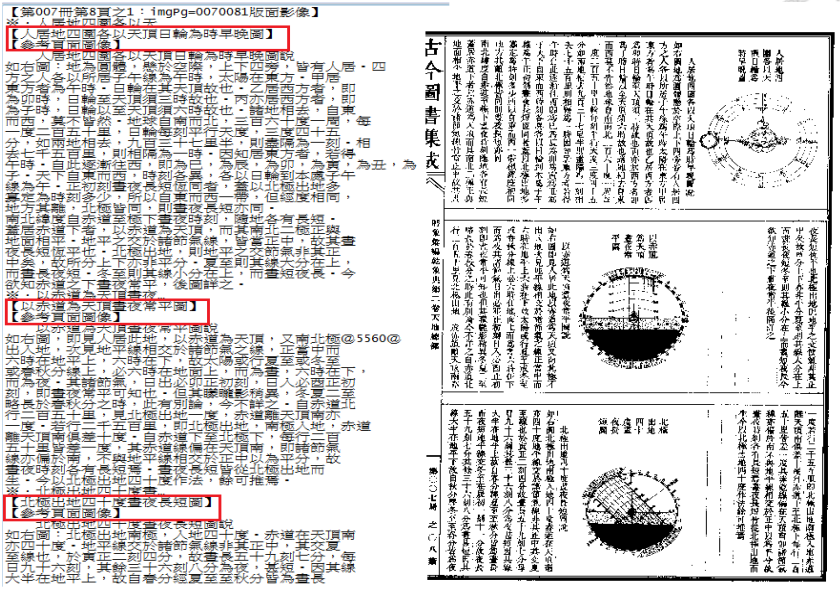

Fig. 4-9　文本中記錄圖像方式

在 Chapter 3 中已介紹過圖像的特性，占據影像中九行的位置，而中括號內之圖像名稱為圖像之一部分，並不屬於內文，所以此處將原文本中兩個中括號替換為九個「*」，以「*」字符來代表此行為圖像區域，更貼近原書情況。如 Fig. 4-10

Fig. 4-10 圖像部分處理結果

✧ 圖像之特例：少數圖像或文本中記錄圖像的方式並未照上述規則，須個別處理，以下圖說明之。

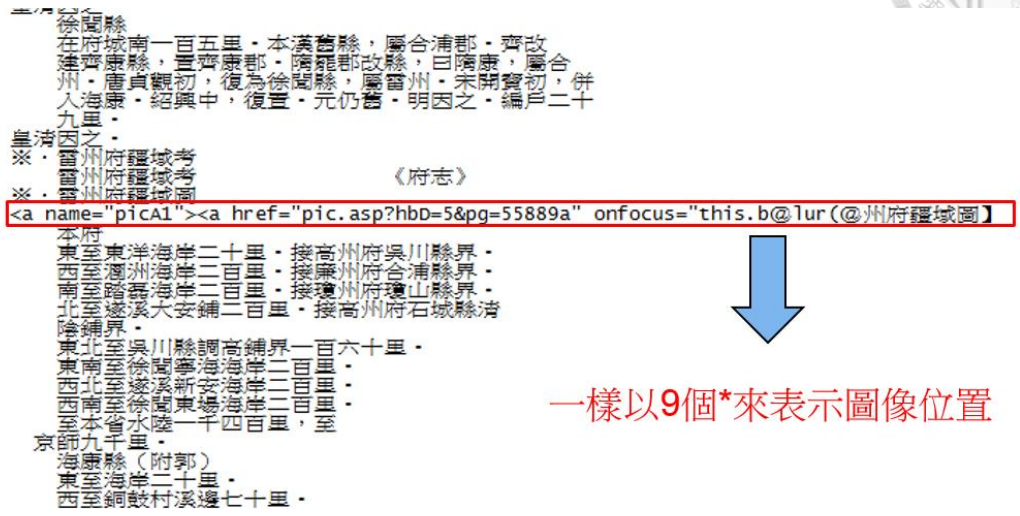

Fig. 4-11 圖像特例 1

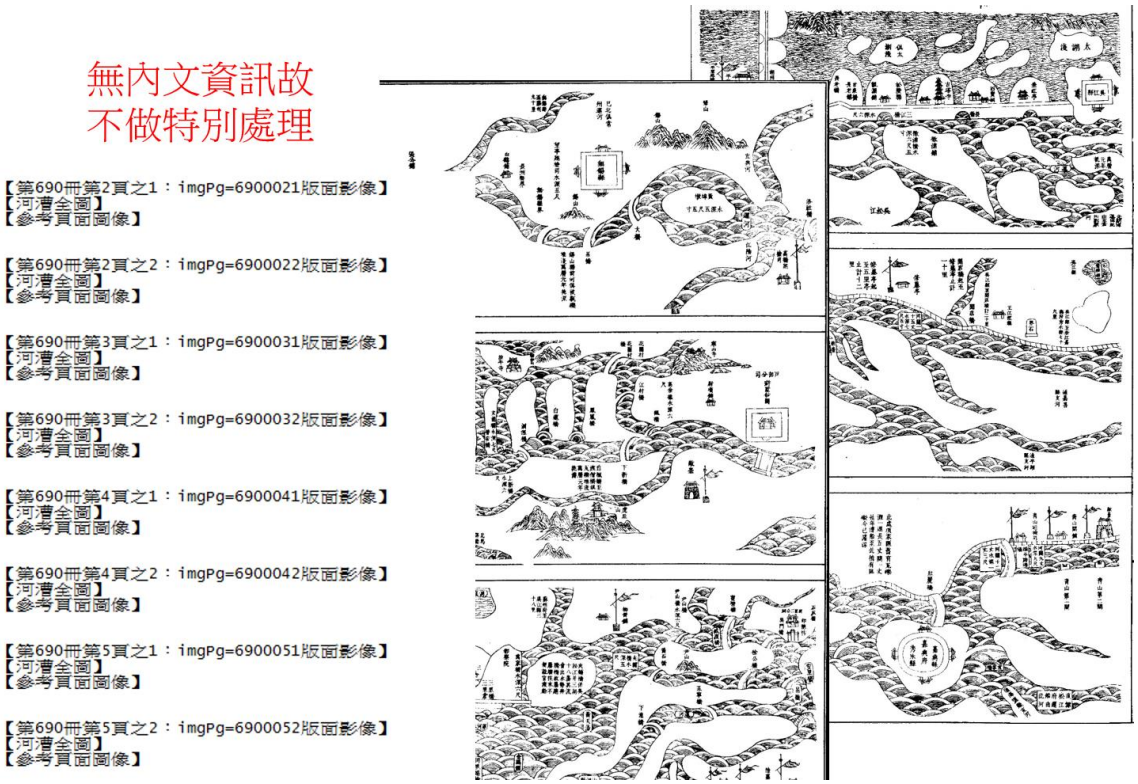

Fig. 4-12 圖像特例 2

特例 2 中整張影像皆為圖像，在文字檔中僅以一張圖表示，但由於影像並無文字資訊，故搜尋不到，因此此處並不做特別處理，同樣以 9 個「*」取代即可。

4.3 補上空行資訊

經前面步驟處理過的文字檔，已經沒有多餘的標點符號、標題，內文空格、斷句也忠於原書，圖像的行數也做了對應，剩下的就是要將原本文字內沒有記錄的空行資訊補上。

要計算影像中的空行，先將先前標準化處理好的影像作去雜訊以及二值化的處理(Threshold 為 128)，再將影像的三個區塊分別對所有 Column 作 scan 的動作，每個區塊的大小是 1580*830，由上而下三個區塊的 y 座標範圍分別是 0~829、875~1694、1750~2579。

利用標準化後的影像，以一個區塊為單位，分別對所有column作scan。

Fig. 4-13 將影像分為三個區塊分別 scan

由於影像已經先二值化，所以 scan 到的像素值只有 0 或 255，若該 column 像素值為 0 的 pixel 數量小於 5，則將該 column 視為空白行的起點，記錄該 column 之 x 座標為 start，持續 scan 直到碰到像素值 0 的 pixel 數量大於等於 5 的 column，將

前一個 column 之 x 座標記錄為 end，可得一空白區域，如 Fig. 4-14。

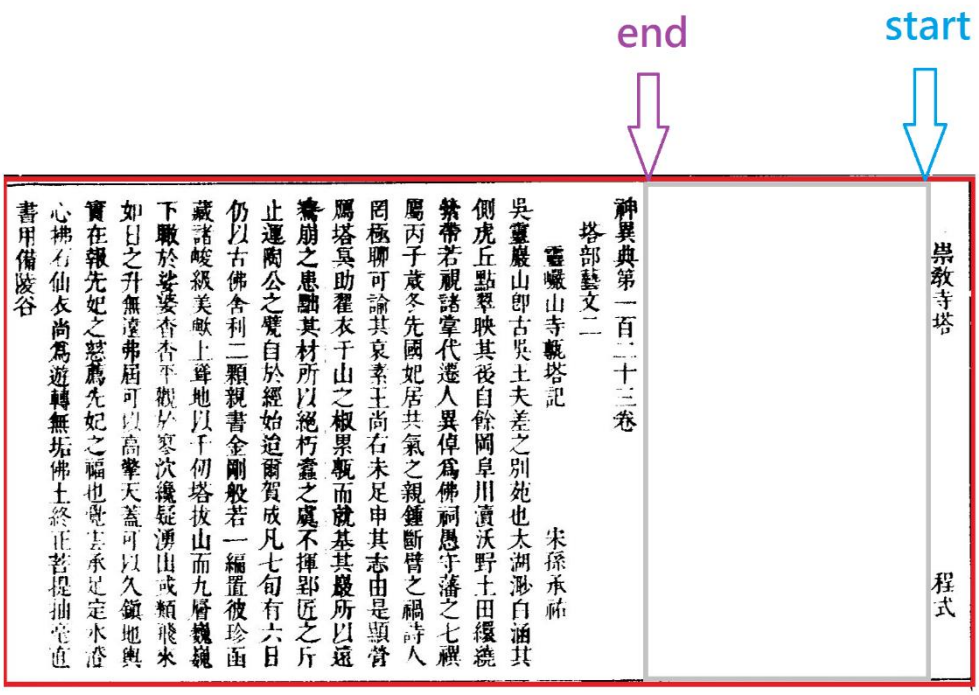

Fig. 4-14 找出空白區域

一個區塊有 27 行文字，區塊寬度為 1580，所以一行的文字寬度為 1580 除以 27 約等於 58.5 個 pixel。檢查此空白區塊寬度（end-start)是否大於 58.5，若非，表示此空白區塊為文字行的空隙，非空行；若是，則將區塊寬度除以 58.5，無條件捨去至整數（因為空行實際上大於一行文字的寬度），可得到此區塊占的空行數。再以 start 的 x 座標算出由第幾行開始，最後以一陣列（size=81）記錄該影像中空行情形，0 表示空行，1 表示非空，實際測試結果如下圖 Fig. 4-15。

Fig. 4-15 測試結果

左圖結果由左至右分別對應了影像由右至左的每一行，刻意找了標準化情形較差之影像（早期印刷技術不夠精確導致行與行間不平行）來做測試，結果仍然正確的計算出了影像中的空行及其位置。接著再利用此陣列以及前面處理好的文字檔，建立一個新的包含空行資訊的文字檔。首先一行一行讀取文字檔的內容，若是文字，則檢查陣列該行的值，若是 1 則照填入新文字檔；若是 0 則以一個「#」表示該行為空行，並持續檢查下一個陣列值直到陣列值 1 才填入該行，否則都填入「#」。若讀取到文字檔內容為圖像區域「*」則不論對應的陣列值為何都填入「*」。整段都處理完成後，檢查扣掉第一行影像檔名總行數是否為 81 行，若否則在該段底下以一「!」做註記，以利後續人工檢查錯誤原因及修正。

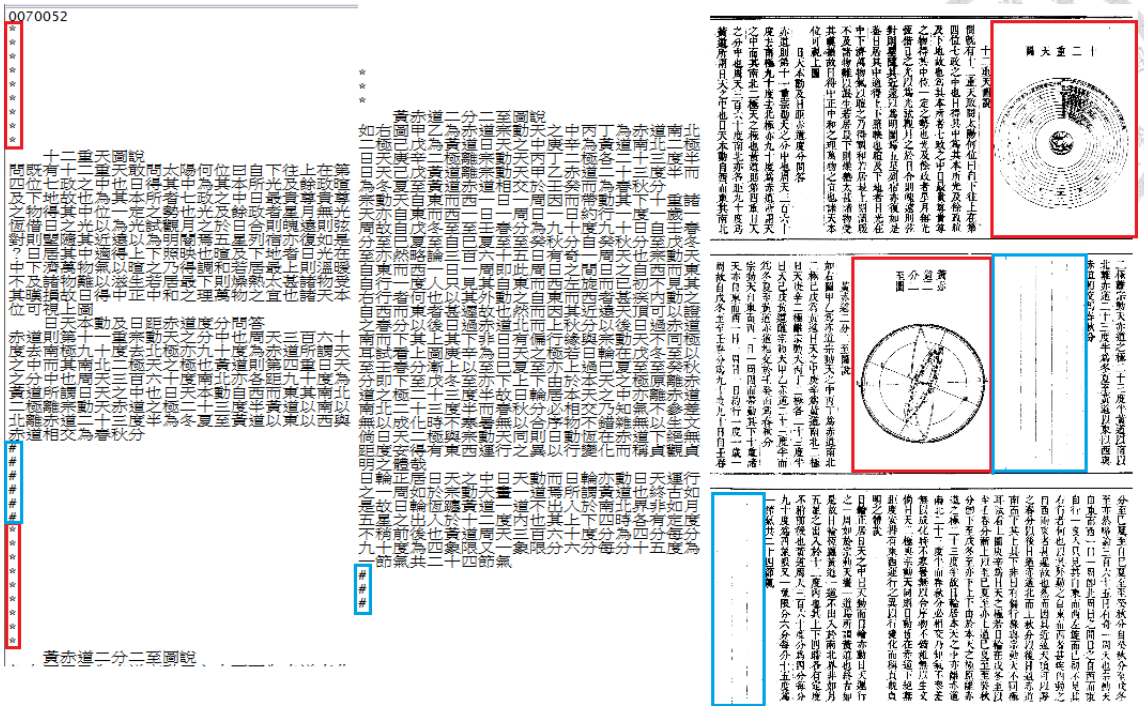

Fig. 4-16 完成對應之文本與其對應之影像

影像總數 96887 張，其中有被「!」共有 1253 張，錯誤率約 1.3%，錯誤的可能原因有：

- 空行算錯(979)
- 人工建立文本時有誤，漏打、多打、將兩行打成一行、將一行打成兩行等(227)
- 圖像特例(47)

錯誤部分需要以人工觀察找出錯誤原因，並修正之。

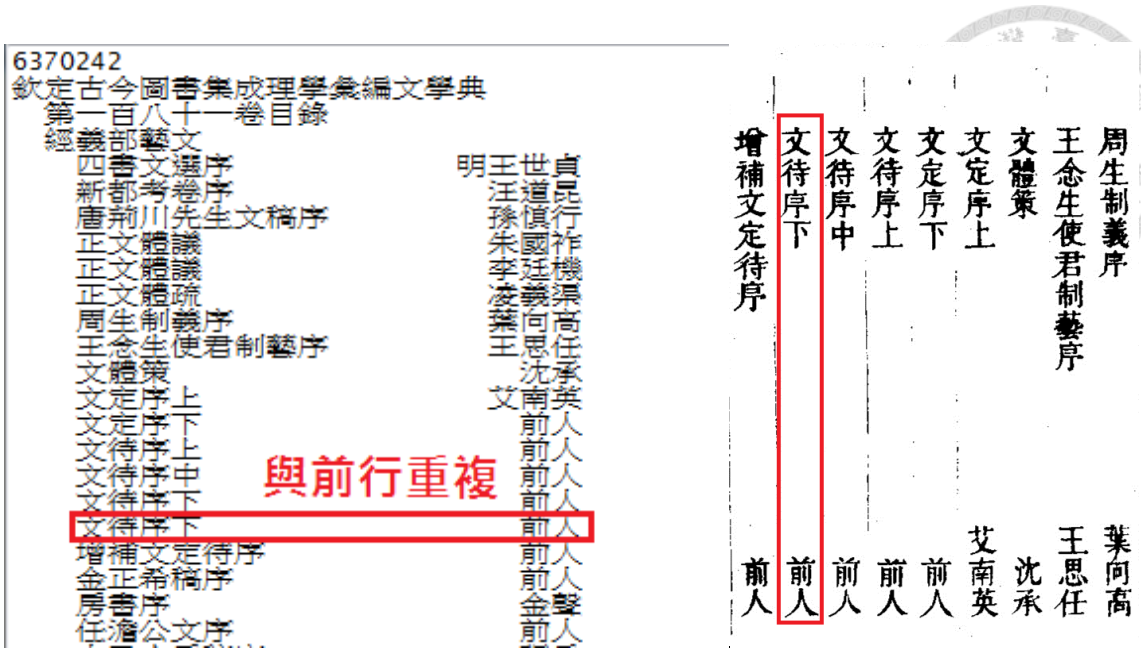

Fig. 4-17 文本錯誤實例 1

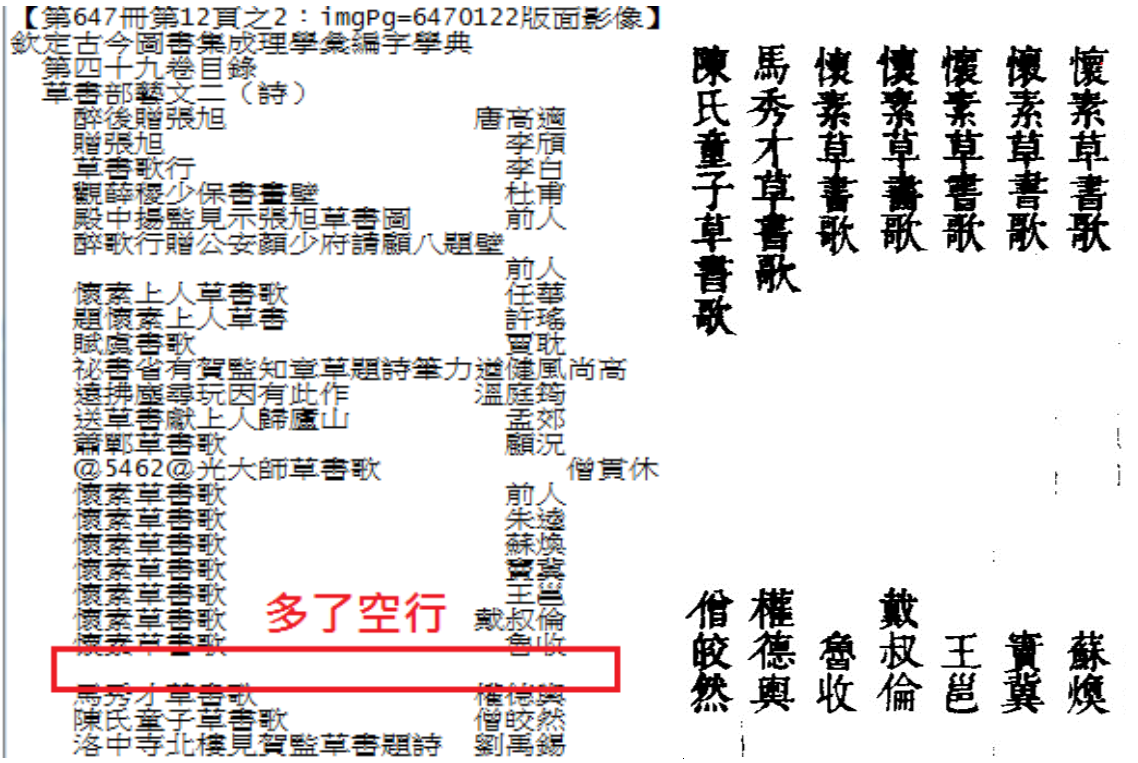

Fig. 4-18 文本錯誤實例 2

Chapter 5　文字位置計算

有了標準化後的影像和與其高度對應的文本，便可以很容易的利用文字在文本中的行數與字數，算出文字在影像中的位置。以下將以實例來計算文字位置：

在全文中搜索關鍵字「蘇軾」，其中一搜索結果如下圖 Fig. 5-1

Fig. 5-1　(左)搜索結果 (右)對應影像

透過搜索到文字段落第一行可知其對應之影像檔，再利用文本計算「蘇軾」二字所在的行數，以及該行的前置字數及空白，以此例來說，「蘇軾」位於文本中之第 28 行，前置字數及空白共 16 字元，28 除以 27 等於 1 餘 1，表示在第二區塊的第一行，一個文字方格的大小約為 58.5*39，第二區塊的起點為（0,875），因此可以算出「蘇軾」二字所在的文字區塊位置起點為（0+58.5*(1-1),875+(16*39)）=（0,1499）以該點往下對兩個文字方格標註如下圖 Fig. 5-2：

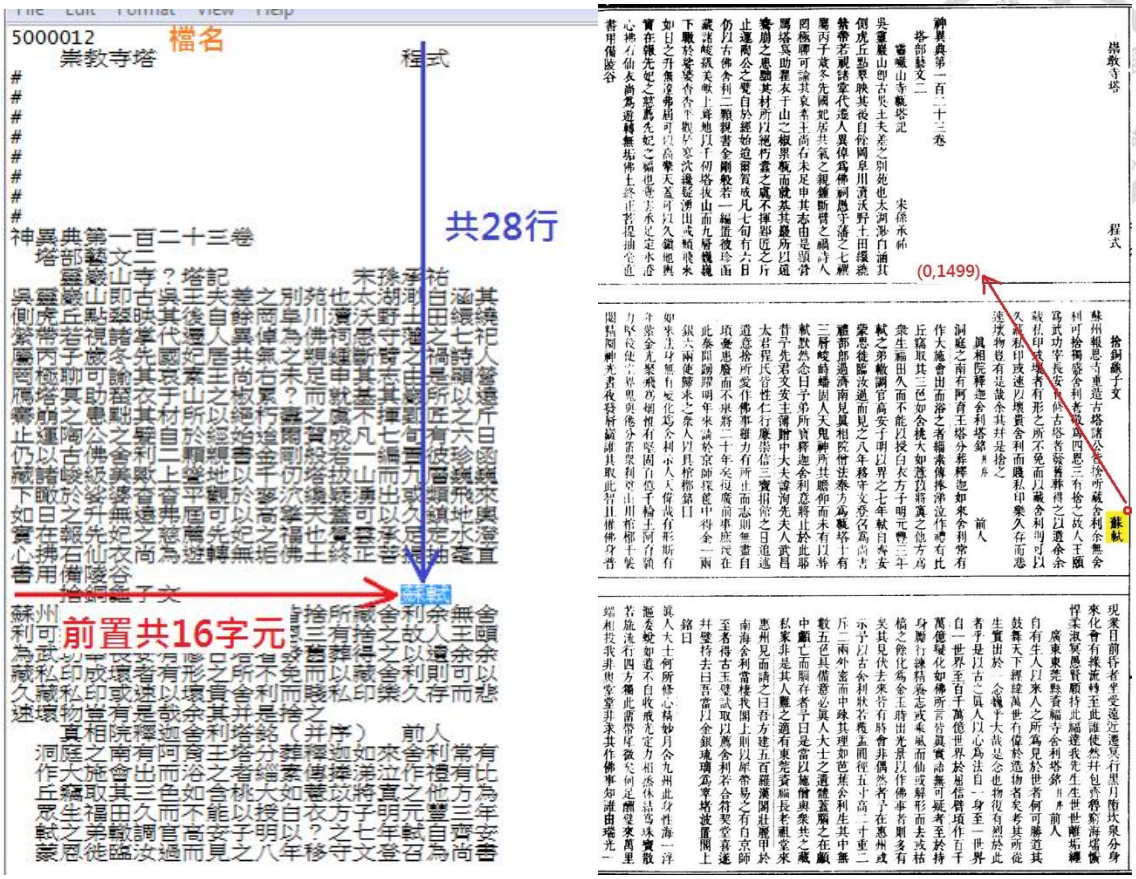

Fig. 5-2　關鍵字於影像中完成標註

即使影像稍有歪斜或偏移，計算出的位置也會在所要找的關鍵字附近。以下為計算公式：

假設關鍵字在文本中位於第 A 行（不包含檔名），前置文字與空白共 B 字元，「/」計算元取整數，「%」計算原取於數，則所在文字區塊起點（x,y）為

$$x = ((A-1)\%27) * 58.5$$

$$y = \left(\frac{A-1}{27} * 875\right) + B * 39$$

若關鍵字跨行則將不同行的部分分開計算，分開標註即可。

Chapter 6　結論與未來工作

在《集成》的序中，雍正帝說：「始之以曆象，觀天文也；次之以方輿，察地理也；次之以明倫，立人極也；又次之以博物、理學、經濟，則格物致知、誠意正心、治國平天下之道，咸具于是矣[12]！」。又說：「故是書之成，貫三才之道而靡所不該，通萬方之略而靡所不究也。我皇考金聲玉振，集五帝、三王、孔子之大成。是書亦海涵地負，集經、史、諸子百家之大成。前乎此者，有所未備；後有作者，又何以加焉[13]？」。這說明了《集成》這本書在歷史上的地位是如此巨大，囊括了中國自古以來的知識菁華。透過現代的技術將此書完成的數位化保存下來，若只透過數位的文字來參考此書未免可惜了，利用本研究的成果可以資訊的方法來觀察《集成》原書中的資訊，更有閱讀此書的實境感。

將計算字數的功能實作於 THDL-based 古今圖書集成，由系統來提示影像中的文字資訊，以達到文本與影像參照閱讀的目標。

《集成》影像與文本對應的成果，除了能夠用來尋找文字資訊，也提供了一定的自動化偵錯能力，透過文本處理、以及行數的對應的過程，標記了出現錯誤的句子或段落，而這些錯誤有許多是建立文本時人為的疏失，利用觀察被標註的部分即可人工修正之。

本論文的研究方法，用到了《集成》本身的特性，例如文字固定行數、圖像固定大小等等，因此並不適用於其他書籍，若是想對其他書籍做類似的對應，需要重新分析該書籍的特性，建立一套新的處理流程。

REFERENCE

引用及參考資料：

[1] 維基百科－古今圖書集成，Available: http://zh.wikipedia.org/wiki/古今图书集成

[2] 裴芹，〈《古今圖書集成》研究〉，2001 年

[3] 古今圖書集成索引＆全書圖像，Available: http://gjtsjc.gxu.edu.cn/

[4] 數位古今圖書集成，Available: http://192.83.187.228/gjtsnet/index.htm

[5] THDL-based 古今圖書集成，Available: thdl.csie.org/L303_GuJinTuShuJiCheng/

[6] 蔡孟竹、曾元顯，〈中文 OCR 文件檢索測試集之製作與應用〉，「教育資料與圖書館學」，第 40 卷，第 3 期，2003 年 3 月

[7] 丁原基，古今圖書集成，Available: http://192.83.187.228/gjtsnet/index.htm

[8] 圖書集成經緯目錄，Available: http://gjtsjc.gxu.edu.cn/jwml.aspx

[9] 維基百科－霍夫轉換，Available: https://en.wikipedia.org/wiki/Hough_transform

[10] 林易徵，〈《古今圖書集成》自動化內容建構與出處擷取〉，碩士論文，國立台灣大學，2013 年

[11] OpenCV 官方網站，Available: http://opencv.org/

[12] 陳夢雷原著，蕭孟能編印，《古今圖書集成及索引》第 001 冊，文星書店，1964 年，p.2-3

[13] 陳夢雷原著，蕭孟能編印，《古今圖書集成及索引》第 001 冊，文星書店，1964 年，p.3

[14] 胡道靜〈《古今圖書集成》的情況、特點及其作用〉，1962 年