



光學字元辨識古籍之全文轉置經驗： 以明人文集為例

林巧敏*  蔡瀚緯**

【摘要】

因應資訊技術的發展，加上數位人文研究對於全文內容分析的使用需求，運用光學字元辨識技術（OCR）將文本內容轉置為全文，可促進全文檢索與內容探勘使用。為瞭解利用 OCR 辨識軟體轉換古籍全文的可行性，本研究運用古籍文本進行實測分析，探討古籍運用 OCR 辨識的成效以及影響辨識率的原因。研究選取 40 種明代文集進行分析，研究結果顯示古籍版式與影像品質皆會影響 OCR 辨識率，尤其版式文字過於擁擠和影像品質不佳，較不利於 OCR 處理，進而歸納出六種常見的辨識錯誤字形樣態，可提供典藏機構進行類似古籍版本全文轉置作業規劃之參考。

關鍵詞

光學字元辨識 全文資料庫 特藏古籍 古籍數位化 數位典藏

壹、前言

古籍數量浩繁，歷代累積的典籍中蘊藏前人智慧的精華，但傳統古籍

* 國立政治大學圖書資訊與檔案學研究所教授
ORCID 0000-0002-9309-9884

通訊作者 E-mail: cmlin@nccu.edu.tw

** 中國信託商業銀行股份有限公司專員
E-mail: bestwiwi79@gmail.com

載體形式為紙質，紙質文獻不僅需要克服保存問題，也不耐經常翻閱使用，致使紙質文獻的資訊傳播受到限制。近代的數位化技術賦予古籍新生命，將古籍內容藉由數位化方式提供使用，可減少翻閱對於原件的傷害，也能藉由資料庫檢索設計快速查詢所需的古籍內容。

早期進行古籍的數位轉置是仰賴人工逐字繕打，但人力成本過高而難以普及運用，古籍數位化處於研究探索階段，其後在「數位典藏國家型科技計畫」主導下，各典藏機構開始將古籍文獻採數位影像掃描方式，建置古籍資料庫運用後設資料 (metadata) 檢索，進而連結古籍內容影像方式提供使用。近年因應資訊技術的發展，加上數位人文研究對於全文內容分析的使用需求，運用光學字元辨識 (Optical Character Recognition, OCR) 技術將文本內容轉換成逐字全文，以提供全文檢索與內容探勘使用，正逐漸受到重視。OCR 最初是運用光學透射原理，善用光線的穿透性，將文字線條加以分析，藉此辨識出文字符號 (Mariner, 2010)。現階段運用 OCR 的目的，並非解決所有文字輸入問題，而是善用這項技術節省人力及時間的耗費 (林巧敏、陳志銘, 2017)。如果能在既有的數位典藏基礎上，運用 OCR 技術將數位典藏之古籍影像透過字元辨識，轉為內容全文字碼，可帶來數位人文研究需要大量全文進行內容分析的重要里程碑發展。

目前臺灣地區建置的全文資料庫，以中央研究院歷史語言研究所建置之「漢籍電子文獻資料庫」規模最大，早期採逐字繕打的方式，30 餘年來累計收錄歷代典籍已達 1,224 種古籍集冊 (中央研究院歷史語言研究所, 2019)。但 OCR 技術的進展，使得漢籍資料庫開始運用 OCR 辨識，協助進行全文的輸入，包括中華電子佛典協會的「漢文電子大藏經」、香港迪志文化公司的「文淵閣四庫全書電子版」、北京書同文公司的「四部叢刊電子全文檢索版」等 (顧力仁, 2001; 陳金木, 2008)。OCR 辨識古籍尚無法達到百分百的精確，需要進行人工校對。目前對於進行人工校對或是人工逐字輸入的效率問題，雖有不同觀點的討論空間 (顧力仁, 2001, 2002; 林巧敏, 2017)，但如果是將 OCR 運用在固定的字體與版面，OCR 的辨識率比起逐字繕打，仍然大有可為。

因此，本研究希望探究古籍進行 OCR 辨識的合宜版式以及辨識過程問題，採用實測分析方式，以國家圖書館 (簡稱國圖) 典藏之明代文人文集 (簡稱明人文集) 為例，挑選不同的古籍版式，比較影響 OCR 辨識精確

率的版式與文字型態，以協助類似版式古籍進行全文轉置優先挑選文本的決策判斷。研究目的在於：

1. 探討影響 OCR 辨識古籍精確率的因素，分析辨識率較佳的版面形式，提供明人文集全文轉置的選擇依據。
2. 整理 OCR 辨識明人文集錯誤的文字字形，提供 OCR 識別字形修正的參考。

貳、文獻探討

一、全文辨識技術之發展與應用

OCR 的概念起源於 1929 年，由德國科學家 Tausheck 提出運用光線、文件及模板的組合，將文件放置於光線及模板之間，試圖操控光線投射影像至模板上。由於光線具穿透性，會穿透文件空白部份而遭黑色部份阻擋，使得投射的結果，可顯示的是遭阻擋的黑色部份，也就是文件中的字形部份，此為 OCR 技術的開端 (Mori, Suen, & Yamamoto, 1992)。

1950 年代，歐美各國從事文字辨識的研究，剛開始僅限簡單的英文及數字辨識；1950 年代中期，日本加入文字辨識研究，將辨識的語言擴大至日文及漢字的範圍；1960 年代，IBM 公司開發出辨識相似文字的技術，可辨識出 1,000 種印刷中文字形。隨著技術的進展，辨識字形的方法越來越多元，各種複雜結構的文字，開始進行相應的辨識軟體開發研究 (潘朝陽，1994)。

不同語言的辨識軟體，依語言的複雜程度，有不同程度的技術要求，西方文字在辨識上具有較高的精確度，是因為西方文字符號結構較簡單，且字母符號種類較少，相較之下，東方文字的符號結構則複雜許多，且文字符號數量逾萬字，複雜程度絕不是英文的 26 個字母可比擬。中文字採 OCR 辨識的困難，在於中文字存在下列特性 (潘朝陽，1994；曾逸鴻、林裕淵，2007)：

(一) 中文字數量過於龐大，包括繁體、簡體及特殊異體字，以康熙字典收錄字為例超過 4 萬多字，辨識區分的難度較高。

(二) 印刷字體種類多元，包括新細明體、標楷體、隸書體等，還可

能有手寫字體。

(三) 中文字筆畫多，筆畫複雜程度會影響字形的辨識。

(四) 中文字存在許多字形相似字，容易造成混淆，不易由外觀辨識出文字差異。

近年來辨識中文字的技術有所成長，基本的過程是需要進行文字前處理(pre-processing)、特徵抽取(feature extraction)，然後比對辨識(matching)後，經由字辭後處理(post-processing)，對照字辭資料庫內容，將可能辨識錯誤的文字，校正成較為通順的詞彙(潘朝陽，1994)。目前尚未有任何OCR軟體能做到百分之百的精確辨識，因此需要進行人工校對，確保輸出文字的正確性。

但進行字辭後處理是有效提升OCR辨識率的技術，可將OCR辨識結果透過軟體字辭庫及語言模式(language model)的訓練，將錯誤辨識偵測出來，並校正成正確詞彙。後處理系統包含兩部份：錯誤偵測(error detection)及錯誤校正(error correction)，主要步驟是發現辨識錯誤詞彙，比對字辭庫儲存詞彙後，加以校正成符合語言邏輯的詞彙(張俊盛、陳舜德，1995；Holley, 2009)。

針對無法辨識的詞彙，可提供多種相似詞彙，供使用者判斷選擇，協助軟體進行智慧學習。校正時，根據選擇資料及辨識結果，對每項候選詞彙產生一個信心值(confident value)，利用系統學習機制的處理，將低於門檻的候選詞彙刪除，善用語言模型的輔助，選擇符合語言邏輯且具高信心值的候選詞彙，視為是正確詞彙並加以校正(Sun, Liu, Zhang, & Comfort, 1992; Mariner, 2010)。

對於OCR辨識的內容進行偵測及校正，能提升OCR辨識的精確度，後處理系統的訓練，需要藉助大量的詞彙以及上下文關聯性的建立，才能提升辨識詞彙的正確性。因此，辨識不同年代類型文件時，應配合內容所記載的語言模式及字辭用語，輸入至OCR後處理系統中，才能提升文件辨識的精確度。

二、影響OCR辨識率因素探討

影響OCR精確度的因素有很多種，包含文件影像品質、掃描解析度、

文字與背景的對比、細部文字特徵的分析、字辭庫詞彙量及 OCR 軟體的設計等。除此之外，語言的複雜度也有所影響，中文相比於英文及日文，由於文字數量遠大於兩者，造成中文的辨識難於其他語言文字。雖然，影響 OCR 辨識精確度的因素有很多，但精確度的高低主要仍由人來判斷，是由計畫人員或軟體開發者，依據辨識的文件結果，計算 OCR 辨識率的高低 (Holley, 2009)。影響精確率的因素，說明如下 (Sun et al., 1992; Patel, Patel, & Patel, 2012)。

(一) 掃描設備

影像掃描儀器的解析度上限，造成解析度受到侷限，無法使用更清晰的影像進行 OCR 辨識；加上掃描儀器及環境不利因素（如灰塵），容易產生文件本身沒有的雜訊。影響程度在於：

1. 解析度：影像解析度越高則越清晰，如果原有的影像解析度不夠，或是設定掃描解析度足夠，但因操作過程不佳，皆會造成影像不精細，導致進行影像字形辨識率結果不佳。數位典藏的影像至少需要達到清晰辨別的程度，雖然解析度決定了影像的細緻度，但一味地提升解析度，並無法直接反映到品質提升上，反而會造成檔案容量的擴大，因此，Chapman 與 Kenney (1996) 提出完整資訊截取 (full information capture) 的概念，認為影像解析度的選擇，除了考慮 OCR 軟體辨識率，也需要考慮影像品質是否耗費成本，在多種因素考量下，選擇最適當的解析度影像，進行 OCR 辨識。
2. 掃描儀器：掃描儀器的好壞，直接影響影像掃描的結果，若儀器支援的解析度有範圍，則無法提供高解析的影像掃描，容易產生字形呈現不完整，或文字線條不連接的情況。此外，掃描儀器的老舊，容易在掃描過程產生故障、毀壞等問題，造成掃描影像不完整，無法產生可辨識的文字。而髒亂的環境容易產生灰塵，會將灰塵掃描至影像中，造成影像產生雜訊，擾亂 OCR 的辨識 (Badoiu, Ciobanu, & Craitoiu, 2016)。

(二) OCR軟體

OCR 軟體的設計，直接影響辨識的精確度。針對 OCR 軟體的功能，分述區塊分割、字辭資料庫及後處理系統，加以說明：

1. 區塊分割：一般的印刷字體，由於鉛字分隔固定，區塊分割的技術影響不大，但對於書寫型文字或手寫文字，由於文字呈現上沒明確的分隔空間，文字相連的情況層出不窮，區塊分割的技術就顯得重要。OCR 軟體依據開發辨識語言的不同，可個別加強區塊分割的能力，但不論何種語言，其辨識皆需要區塊分割的協助。Al-A'ali 與 Ahmad (2007) 因阿拉伯文為草寫及連字的特性，需事先進行水平輪廓投影 (horizontal projection profile) 及垂直輪廓投影 (vertical projection profile)，分割出個別段落及文字區塊。接著，採用動態游標 (dynamically cursor) 的移動擴張，捕捉分割區域的所有特徵，善用游移捕捉技術，能確保阿拉伯文的所有筆畫，皆能被系統捕獲辨識。
2. 字辭資料庫：字辭資料庫的詞彙數量，影響 OCR 辨識的精確度，字辭庫所儲存的詞彙，最好是符合辨識文件的語言模式及使用詞彙。Holley (2009) 在執行澳洲國家圖書館典藏報紙計畫時，運用澳洲當地字辭典輔助 OCR 軟體字辭庫的建立，但除了建立字辭資料庫外，也可建立因外觀相似，經常造成辨識錯誤的混亂字表 (confusing table)，當辨識錯誤產生時，透過混亂字表的比對，判斷是否為錯誤詞彙並加以校正，可提高辨識的精確度 (曾元顯，2004；Balk & Ploeger, 2009; Holley, 2009)。
3. 後處理系統：後處理系統是根據區塊的分析結果，辨識各區塊內的文字走向、筆畫連續性等，以此特徵為依據，與系統字辭庫儲存詞彙進行辨識比對。後處理系統內有混淆詞彙表，透過辨識所累積的經驗，將經常辨識錯誤的部份納入詞彙表中，可搭配字辭庫協助進行辨識更正，但詞彙表若不夠詳細可能造成錯誤辨識的遺漏。將自動辨識的詞彙與字辭庫內的詞彙進行歸類，不論正確詞彙與否一併連結，檢索進行時會根據輸入詞彙跳出相類似的詞彙，藉由使用者的挑選，能協助建立後處理系統的混淆字表，字表越詳細則越能提

升 OCR 辨識的精確度 (曾元顯, 2004)。

(三) 辨識文本

文字字體、文字字型、文字大小、版面編排及保存狀況等,皆會造成 OCR 辨識的影響。分述如下:

1. 文字字體: 中文字體種類多元, 包含隸書、楷書、草書、行書等, 因朝代不同產生型態略有差異的字體。如果將相同文字的不同字體皆儲存至字辭庫中, 除辨識時間會拉長外, 系統的儲存空間也會過量。因此, 最好的解決方式, 是針對辨識文本建立專門的字體庫, 或運用演算法辨識文件字體的差異 (曾逸鴻、林裕淵, 2007)。Cojocaru, Colesnicov, Malahov, & Bumbu (2016) 掃描 18 至 20 世紀羅馬尼亞的印刷書, 發現該批書籍包含各時期的字體, 字體的穿插會影響 OCR 辨識的成效。為解決字體差異的問題, 建立不同字體模板及字體對照表, 透過字體交叉比對, 輔以軟體字辭庫資料, 將不同字體辨識出來, 並在輸出時轉置成相同字體。藉由字體對照表的建立, 能方便不同字體的字母比對, 協助辨識涵蓋多種字體內容的文本。
2. 文字字型: OCR 軟體不擅長進行手寫辨識, 原因在於手寫字型不同於印刷字型固定, 由於書寫者習慣的差異, 容易造成文字黏在一起, 使得字型呈現上沒有規律性。OCR 運用空白區間以辨識不同文字, 但過於靠近的文字會造成辨識困難, 使得辨識單一文字混亂。藉由智能字元辨識 (intelligent character recognition) 的技術, 運用神經網絡 (neural network) 的分割手法, 將過於靠近的文字, 自動調整切割區塊, 可協助辨識不同的手寫文字 (Mariner, 2010)。
3. 文字間距大小: 文字前處理需要將文字與背景相互分離, 文字間距與大小是影響分離的關鍵。文字大小往往決定了空白區間的大小, 文字越大區間空白也越大, 較有利於文字與背景的分離; 反之, 空白區間太小, 要靠高解析掃描或線條粗化的輔助, 以確保文字與背景能順利分離 (Zhu, Tan, & Wang, 2001)。
4. 古籍物件狀況: 年代久遠的文本, 可能因保存不佳等因素, 造成紙

質變形、文字扭曲等現象，或因印刷技術的不純熟，造成印刷品質不佳、掃描透頁、字跡暈開等現象，也可能因年代差異，有用字過時、文字過小及版面不規律等狀況，此皆會造成 OCR 辨識問題（Balk & Ploeger, 2009）。尤其，古籍歷經時間考驗，往往保存狀態不甚完好，迭有紙張破損、霉斑、字跡褪色或是紙張黃化、髒污等問題，嚴重影響 OCR 辨識的進行。

三、明代古籍字體版式

明代刻書依特色可大略區分三個階段，明代初期承襲元代風氣，字體呈現樸實樣貌，多為黑口趙體字；明代中期受復古運動的影響，復刻宋版本的風氣盛行，字體形態轉變成較為方正，多為白口歐體字；明代後期演變成橫輕豎重、板滯不靈的方正字體，多為白口匠體字（李清志，1985；潘美月，1985）。

古籍因種類、版式的差異，會產生不同的風格樣貌，其差異因素包含裝訂形式、版式配置、刻字字體、印頁行款、刻印品質等。依照版本的差異進行研究，能方便判別刻印古籍的時代背景，近一步從事過往歷史的研究。但古籍版式、字體的差異，也是影響 OCR 應用於古籍辨識的阻礙，由於 OCR 辨識的精確度受到影像的解析度、字體及版面形式等因素影響，若想加強 OCR 的辨識率，需要透過軟體調整的方式，針對不同字體進行學習辨識。版式雖然與辨識文字沒有直接相關，但是版式會阻礙 OCR 辨識的流暢，軟體會將版式線條樣貌視為雜訊，進而降低古籍內容的辨識率。

（一）古籍版面格式類型

古籍版面格式與現代書籍有很大的差異，版面是由一印頁經折疊形成雙面所組成，每張印頁皆由相同的格式組成，不同的朝代、地區、單位刻印的文本會產生略微差異的版式樣貌。因在印頁位置不同，有不同版式部位的稱呼，主要包含版面（框）、界格（欄）、天頭、地腳、版心、魚尾、象鼻、書耳（圖 1）（駱偉，2004；劉兆祐，2007）。

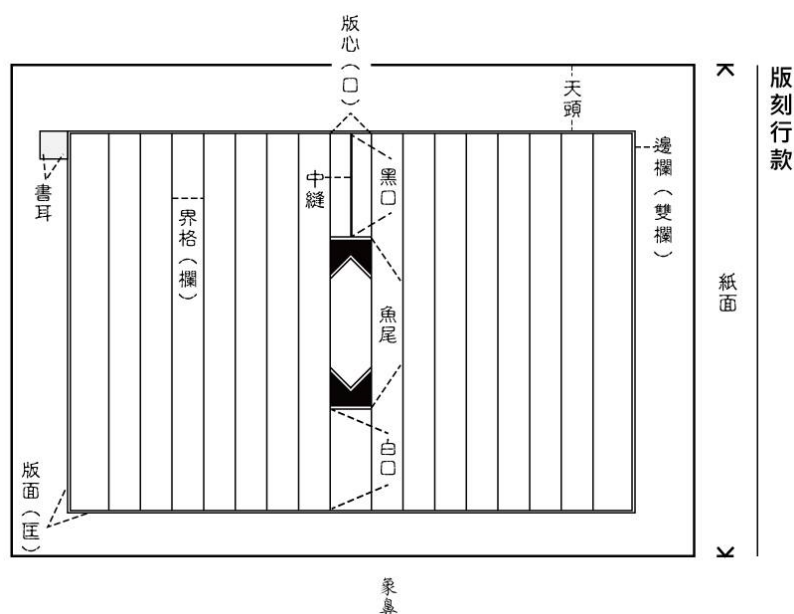


圖 1 古籍版式種類

資料來源：駱偉（2004）。簡明古籍整理與版本學（頁 13）。澳門：澳門圖書館暨資訊管理協會。

古籍印頁圍繞四周的黑線，稱為版框，亦可稱作邊欄。古籍邊欄內用來區分行段的直線，稱為界行，亦可稱作界格。明代古籍的界行不一，常見行數為 18 行至 22 行不等；每行常見字數為 16 字、18 字、20 字、22 字等，依不同的刻印版本而有不同的數量差異。

古籍版面格式的種類非常多元，每個位置的版式種類又有些許差異，加以組合後會產生各種獨具特色的版式風格，可作為評判古籍版本的依據。除了版面格式外，古籍所使用的字體也會隨著政局、社會、文化等因素影響而有所不同，通常刻印使用的字體為當時較為流行使用的字體，因此不同朝代古籍有適合當時使用的字體，可作為古籍版本學研究的參考。

（二）明代古籍字體及版式

依據刻本字體及版式的差異，將明代古籍分成三個時期，包含明代前

期、明代中期及明代後期。明前期指明洪武至明弘治（1368-1505）這段期間；明中期指明正德至明隆慶（1506-1572）；明後期指明萬曆至明崇禎（1573-1644）這段期間。由於受政治、社會、經濟等不同因素影響，使得各時期刻書的版式及字體皆有明顯差異，文集版式圖例如下：

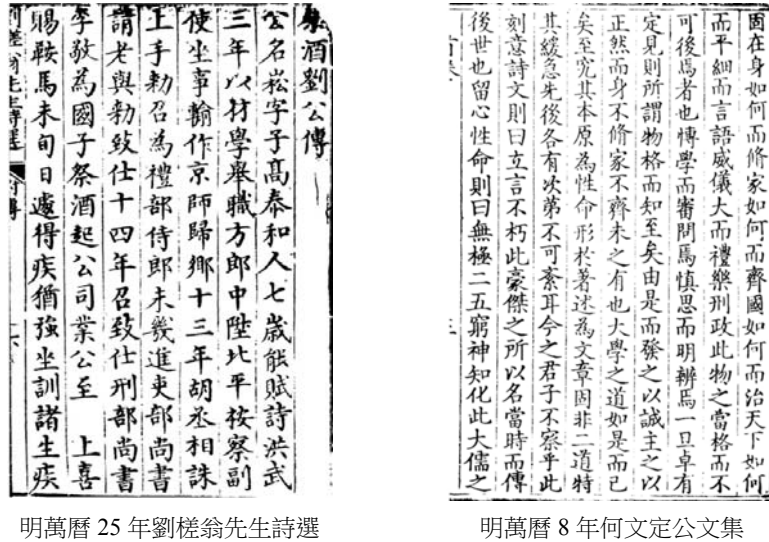


圖 2 明人文集字體及版式圖例

明代古籍各時期版式及字體特性說明如下（李清志，1985；周駿富，1985；駱偉，2004，頁 50；黃永年，2005，頁 118-139）：

1. 明代前期：明代前期歷經洪武至弘治這段期間，該時期所出版的刻書較為稀少，主要在於政治狀況尚未穩定，也因此承襲了元代墨守成規的出版風氣，基本上延續了元刻本的風格，主要特色為使用與元刻本相同的趙體字。趙體字為趙承旨的書法字體，具有古拙樸實之氣，由於承接元刻本風格的關係，連帶影響前期不論在刻印或是書壇上，皆流行使用趙體字。版式部份，不同於元刻本常用的白口、細黑口，主要發展成黑口、大黑口的樣式；魚尾以雙黑魚尾居多，邊欄以雙邊居多，以四周雙邊及左右雙邊兩種為主；刻書裝訂的形式採包背裝。

2. 明代中期：明代中期歷經正德至隆慶這段期間，該時期刻印的發展起了很大的變化，最明顯的差異在於使用字體，由前期的趙體字更改成較為整齊的歐體字，歐體字指歐陽詢的書法字體，特色是方潤整齊、稜角峻厲，由於形狀方整，可達疏密且不雜亂的效果，成為明代中期最為流行的印刷體。由於該時期文壇正經歷「前後七子」倡導的復古文學運動，在「文必秦漢，詩必盛唐」的主張下，文人紛紛崇尚宋代時期的刊本。版式部份仿照南宋浙本，不同於前期大黑口，轉變成以白口為主；魚尾則由雙魚尾轉變成以單魚尾為主；邊欄呈現左右雙邊。
3. 明代後期：明代後期歷經萬曆至崇禎這段期間，該時期刻印的發展又與中期呈現不同的風格，其中以字體的轉變最為明顯。明代後期轉變成更加整齊的匠體字，由於具有方方正正、橫細豎粗的風格，又被稱為「方體字」，亦稱作「宋體字」。明代後期以萬曆時期出版的數量最為龐大，其風格在於字體使用較為方正的匠體字。版式部份，版心以白口為主，單魚尾居多，邊欄呈現上以四周雙邊或左右雙邊為主。此時期印刷事業興盛，刻印數量與版式相較於前述時期顯得豐富多元。

由於明代刻印事業發展興盛，留存許多古籍文本，可提供明代社會研究的重要史料，但也由於印刷事業發展的興盛，造成古籍版式的組合形式多元且複雜。綜合文獻探討，OCR 辨識的精確度會取決於文本的保存狀況、文本的字體及文本的版式，在未來挑選辨識文本進行數位化時，如果選擇較能單純呈現高辨識率的文本，在評估古籍版式上除需考量年代因素外，刻印地點及刻印單位也是必須考量的因素。

參、研究設計

本研究為探討 OCR 軟體對於不同版式古籍文字的辨識率，以單一 OCR 軟體進行不同版式的辨識比對。OCR 軟體選擇較為新穎的商務型 OCR 軟體「ABBYY FineReader 14」，本研究以商務型軟體進行測試，能降低對於資訊人員的仰賴，因本文重點不在探討如何進行 OCR 軟體校正，而是以典藏機構角度，測試能否以市售取得之 OCR 進行全文轉置的作業

分析，以探討適合 OCR 優先處理的文本和轉置過程問題，至於 OCR 修正技術已有資訊技術領域文獻探討，非本文關注重點，此亦為本研究範圍限制。

研究分析古籍以國家圖書館珍藏明人文集為處理標的，主要是因為該文本有研究需求，亟欲轉為全文內容提升館藏服務。希望藉由便利取得之 OCR 軟體，比較不同古籍版式的辨識率，並瞭解古籍版式及影像品質對於辨識程度之影響，找出辨識成效較佳的古籍版式，提供進行全文轉置優先選擇之判斷參考。根據文獻分析瞭解古籍版面的行數與行字數可能造成辨識率影響，其他諸如魚尾數、黑白口等版式差異與 OCR 辨識影響不大。因此，本研究在判斷古籍版式差異與辨識率影響程度時，暫不討論魚尾及象鼻等版式變項，僅著重於比較行數及行字數的差異。

一、研究對象

「明人文集」是研究明代歷史的重要文獻，收錄眾多作者集結而成的合集及個人文章匯整的文集，是對於明代政治制度、文學流向、學術思想、社會變遷的研究，不可或缺的重要史料。國家圖書館典藏 6,000 多部明版書，是國內極為重要的文化資產，有進行內容分析之發展性（國家圖書館，2020）。故本研究以國家圖書館珍藏之明代古籍為對象，明人文集是概括性的稱呼，並非專指古籍的集冊名。基於前述文獻分析可知，明代後期版刻數量多、版式多元，因此，本研究取樣分析之明人文集，以明代後期萬曆及嘉靖年間刊刻之集冊為主，主要考量為該時期印刷出版業達到顛峰，各式各樣不同版式古籍蜂湧而出，透過以此期出版古籍進行辨識比對，可蒐集到以不同版式進行 OCR 辨識的分析數據。本研究立意選樣之古籍書目資訊如下（表 1）：

表 1
分析測試之 40 本明人文集清單

編號	國圖館藏題名	編號	國圖館藏題名	編號	國圖館藏題名
1	誠意伯劉先生文集	16	練公文集	31	陸子餘集
2	王文成公全書	17	西菴集 9 卷	32	竹澗先生文集
3	蟻螻集	18	何文定公文集	33	望雲集

（續下表）

(接上表)

編號	國圖館藏題名	編號	國圖館藏題名	編號	國圖館藏題名
4	鳳池吟稿	19	方簡肅公文集	34	楓山章先生文集
5	空同集	20	芻蕘集	35	瓊臺會稿
6	劉棧翁先生詩選	21	滄溟先生集	36	邊華泉集
7	四溟山人詩	22	弇州山人讀書後	37	甫田集
8	白沙子全集	23	皇甫少玄集	38	蘇門集
9	備忘集	24	徐迪功集	39	震澤先生集
10	夢山存家詩稿	25	遜志齋集	40	徐文靖公謙齋文錄
11	覆瓿集	26	何大復先生集		
12	校刻具茨先生詩集	27	類博稿		
13	西隱文稿	28	唐荊川先生文集		
14	文清公薛先生集	29	翰林羅圭峯先生集		
15	薛考功集	30	康齋先生文集		

二、研究工具

OCR 軟體分為一般型 OCR 辨識軟體及商務型 OCR 辨識軟體兩類。一般型 OCR 辨識軟體指應用 OCR 技術從事簡單文字符號的影像辨識，通常放置於網路上提供使用者便捷且快速的辨識功能，如 Online OCR、i2OCR 等；商務型 OCR 辨識軟體指廠商應用 OCR 技術專門開發來辨識物件內容的軟體，如 ABBYY FineReader、丹青文件辨識系統等。兩者相比，由於技術廠商在不斷執行辨識作業中，會強化修正 OCR 系統的辨識能力，因此，商務型 OCR 辨識軟體會比一般型 OCR 辨識軟體具有適應不同類型字體變化的能力，能直接反應在 OCR 辨識的精確率上。

本研究挑選商務型 OCR 辨識軟體進行古籍影像的辨識測試，是因先行簡易測試過一般型 OCR 辨識軟體的辨識率過低，僅有三至五成辨識率，無法滿足本研究測試需求，因此挑選商務型 OCR 辨識軟體作為本研究的測試工具。選擇使用的 OCR 軟體為 ABBYY 開發的 FineReader 14，雖然該 OCR 軟體不像丹青文件管理系統是專門設計為辨識中文漢字使用的 OCR 軟體，但參酌網路使用經驗評論 (ABBYY Production, 2017; ABBYY, 2020)，因操作簡易且可取得試用版免費使用，因而決定選擇使用 ABBYY FineReader 14 作為本研究測試的工具。

採用 ABBYY FineReader 14 符合本測試需求之功能特性為：

- (一) 提供文件影像與辨識文本的直接比較。
- (二) 針對辨識古籍影像能提供影像切頁、傾斜校正、雜訊去除等初步制式化前置處理作業。
- (三) 辨識會自動切劃文字區塊與影像區塊，若系統切割方式錯誤也能採用人工方式劃分辨識區塊。

因本研究分析是以 ABBYY 軟體以及明代文集為測試標的，研究結果期許可提供類似版式文本及運用類似辨識軟體實務作業之參考。

肆、研究結果分析

一、古籍影像辨識分析

本研究以隨機抽樣挑選古籍影像，將所選之 40 冊文集加註文集編號，將該文集編號除以該文集擁有卷冊、別冊、附錄冊數，相除後餘數值為所要隨機抽樣的文集卷次。由於古籍影像有可能因原版刻文字模糊，或者掃描過程處理缺失導致影像品質不佳，造成該批古籍影像品質有清晰程度上的落差，先由目視判斷挑選 5 頁品質較佳之影像，以避免因影像品質落差過大而造成辨識率判別極端落差的問題，總計 40 冊文集各挑選 5 頁古籍影像、合計 200 頁的影像，藉由測試分析瞭解古籍版式對於影像辨識的差異，甚至探知影像品質優劣對於 OCR 辨識率的影響。

將 40 冊文集以編號標示，並記錄不同古籍之版面格式差異，根據文獻分析得知文字編排方式對於辨識率有影響，故進行古籍版面行字數分析。

(一) 行數與行字數分析

測試分析之 40 冊明代古籍版式行數分布在 8 至 12 行之間，其中行數為 10 者占整體樣本數最多；其次為行數 9 者，行數 12 之古籍樣本最少，僅占 2.5% (表 2)。分析 40 冊古籍進行 OCR 辨識結果，單純以行數值和平均辨識率比較，可發現行數值介於 8-9 的辨識結果，比行數值介於 10-11 者的辨識率高。初步判斷係因行數少者，文字間左右的行間空白較為明顯，比較有利於文字與背景的分離 (Zhu, Tan, & Wang, 2001)。雖然行數 12 的

辨識率不差，但在明人文集此行數樣本數偏低，如檢視該樣本頁面，顯然影像品質對於辨識率的影響更大，故不能全然以行數為判斷依據。

表 2

古籍影像版式行數與辨識率統計表

行數	古籍頁數	百分比(%)	平均辨識率
8	10	5	0.690
9	55	27.5	0.621
10	100	50	0.572
11	30	15	0.518
12	5	2.5	0.743
合計	200	100.0	0.588

接續分析 40 冊明代古籍版式行字數，其數值介於 16 字至 24 字之間，取樣古籍的行字數 20 字者，占整體樣本數最多，其次以行字數 18 字者為次（表 3）。如果觀察古籍版式行字數與辨識率關係，可以發現行字數若介於 19-24 字的辨識結果是呈現較差的狀況，行字數若是 18 字則辨識率較佳，字數 16 雖然辨識率比字數 18 辨識率差，但僅占 5% 的樣本數，檢視此 10 頁影像品質，顯然辨識率受到影像品質的影響更大。因此，僅根據行字數與辨識率的關係，初步可理解當行字數少則間距較明顯，也會有助於字元的辨識（圖 3）。

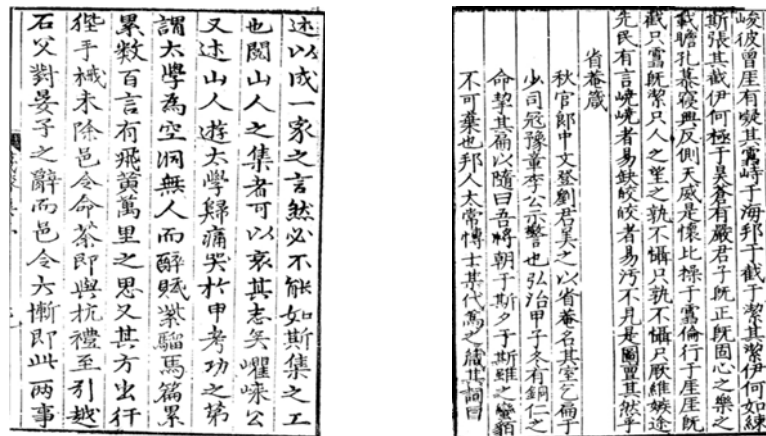
表 3

古籍影像版式行字數與辨識率統計表

行字數	古籍頁數	百分比(%)	平均辨識率
16	10	5	0.547
18	40	20	0.706
19	20	10	0.538
20	90	45	0.578
21	15	7.5	0.518
22	20	10	0.564
24	5	2.5	0.396
總計	200	100.0	0.588



行數較少之行間空白明顯，較為容易辨識字形（左為 8 行，右為 12 行）



行字數少則文字間距較為明顯，也較為容易辨識字形（左圖為 16 字，右圖為 22 字）

圖 3 不同行數與行字數影像辨識率說明圖

（二）影像品質分析

由於影像品質在此次受測的國圖古籍樣本中差異大，也是目前已有數位典藏的古籍影像常見問題。因此，影像品質是影響 OCR 辨識率不可忽

略的因素，本研究參考中央研究院數位典藏制訂之「檔案數位化影像製作規範書」檢核影像品質優劣標準，進行本研究古籍影像品質優劣的判斷依據，分級如下（表 4）。

表 4

影像品質檢核評估項目

項目次序	檢核項目	評分標準	說明
01	影像歪斜不正或陰影遮掩	各項目符合者標註 1，不符合者標註 0，起始數值為 1，將所符合項目分數相加，所得為影像品質數值。數值區間為 1 至 5 級，級數越低則影像品質越優，反之則越劣。	古籍採對折或是反折的方式裝訂成冊，因此經掃描的古籍影像單頁具有兩印頁古籍。若兩印頁間有明顯不平行、或有明顯摺頁陰影，則符合該檢核項目。
02	影像色澤不均		古籍影像若掃描色澤不均，會影響影像內容呈現。樣本中多有色澤不均等情形，故判斷標準為色澤不明顯處若占整體 4 成以上，則視為符合此檢核項目。
03	影像文字線條不連續或斷裂		古籍內容文字若線條不連續，會影響文字辨識呈現。選定樣本內容多少有線條斷裂等情形，故判斷標準為內容文字有斷裂、不連續等情形若占整體 4 成以上，則視為符合此檢核項目。
04	影像文字周圍具斑點汗漬		古籍文字周遭若有斑點汗漬等情形，會影響文字辨識呈現。選定影像多數已經後處理過，但仍有部分影像由於原件過於老舊，以致仍有殘餘汗點於影像上，判斷標準為周遭有汗點文字占整體 4 成以上，則視為符合此檢核項目。

將符合檢核項目之數值標註為 1，不符合檢核項目標註為 0，意指影像具有該項目的情況則標註 1，反之則 0。由於本研究所辨識的明代古籍影像，皆由國家圖書館珍藏提供，該批古籍善本之數位化作業於同一時期進行，透過檢核項目數值的分析，可分析該時期的數位化作業產生的影像共同問題。將選取樣本經過逐頁詳細檢核，統計影像品質狀況分布如圖 4。

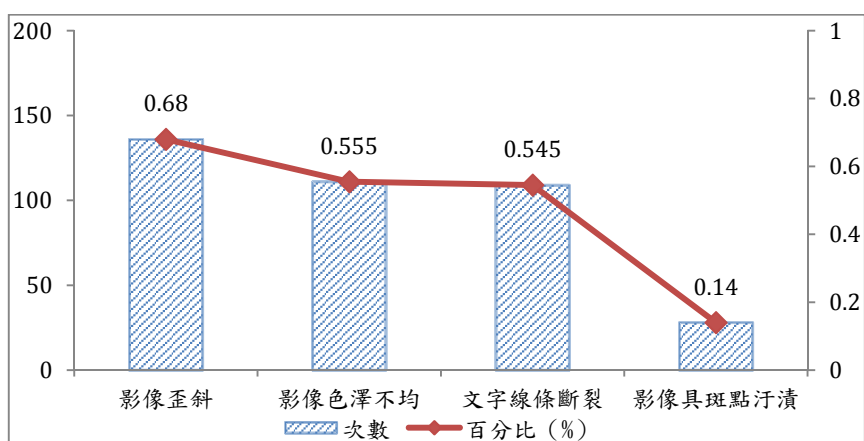


圖 4 影像品質檢核結果之次數與百分比統計圖

結果顯示，影像歪斜不正、影像色澤不均與影像文字線條不連續三項檢核的發生次數相差不大，在 200 頁古籍中發現有 136、111 及 109 次數，分別占整體樣本數的 68%、55.5% 及 54.5%，而影像具斑點汙漬的項目符合次數較少，僅占整體樣本的 14%。顯示該時期古籍原件品相維持良好或是能經過數位化過程在影像處理斑汙點問題，但對於影像歪斜不正、影像色澤不均與影像文字線條不連續問題，顯然缺乏後製修正和檢核校正處理，如果日後進行同一數位化時期所掃描的其他未被抽樣的古籍影像，也有可能發生類似問題，需要先進行影像的後處理才能使用 OCR 進行全文辨識。

(三) 行數、行字數與影像品質對辨識率的影響

為了探討古籍版式字數、行段數以及影像品質對於 OCR 辨識率之間是否具有相關性，使用獨立樣本 T 檢定協助判斷自變項（行數、行字數以及影像品質）中不同差異值與依變項（辨識率）是否有顯著差異；並以多元迴歸分析探討變項間的關係是否具有預測性，試圖找出變項間的因果關係。

1. 獨立樣本 T 檢定

獨立樣本 T 檢定是檢測自變項內組別的差異，由於三種自變項數值多

樣，無法直接區分成兩種組別，因此採用概略劃分數值的方式。將 200 個古籍影像樣本的自變項（行數、行字數、影像品質）按數值排列並取出樣本中位數，以各自變項的中位數作為區分不同組別的中間值（中位數分別為 10、20 與 3）。行數以中位數 10 為間值，大於等於 10 行的數值視為組別 A，小於 10 行的數值視為組別 B；行字數以中位數 20 為間值，大於等於 20 字的數值視為組別 A，小於 20 字的數值視為組別 B；影像品質以中位值 3 為區間值，大於等於品質 3 的數值視為組別 A，小於品質 3 的數值視為組別 B。進行樣本 T 檢定結果如下（表 5）。

表 5

各變項與辨識率獨立樣本 t 檢定結果

自變項	組別	個數	平均辨識率	標準差	t 值	顯著性(雙尾)
行數	A(≥ 10)	135	.56643	.137545	-2.999	.003
	B(< 10)	65	.63180	.157717		
行字數	A(≥ 20)	130	.56206	.137745	-3.443	.001
	B(< 20)	70	.63524	.153381		
影像品質	A(≥ 3)	136	.52294	.122111	-13.092	.000
	B(< 3)	64	.72523	.090905		

結果顯示自變項之行數不同組別間具有顯著差異（ $t=-2.999$, $p=.003<.05$ ）；變項行字數不同組別間具有顯著差異（ $t=-3.443$, $p=.001<.05$ ）；變項影像品質不同組別間具有顯著差異（ $t=-13.137$, $p=.000<.05$ ），顯示三種不同組別之自變項所對應的辨識率具有顯著的差異。亦即，行數較大的組別（組別 A）所得辨識率會比行數較小的組別（組別 B）之平均數還來得低；行字數較大的組別（組別 A）所得辨識率會比行字數較小的組別（組別 B）之平均數還來得低；影像品質較差的組別（組別 A）所得辨識率會比影像品質較優的組別（組別 B）之平均數還來得低。

可知三種自變項數值高低會影響 OCR 辨識率且有顯著差異，如果在進行古籍字元辨識時，可挑選古籍行數低、行字數少與影像品質相對佳的影像，其 OCR 辨識結果會有較佳的辨識率。

2. 多元迴歸分析

運用多元迴歸分析統計方法，可獲得不同自變項之影響性資訊，本研究選擇三種可能影響 OCR 辨識率的自變項，作為多元迴歸分析的分析變項。以古籍物件而言，古籍的版面格式具有差異性，將其版式差異選擇「行數」及「行字數」兩種變項，作為古籍版式的預測變項；以數位影像來說，雖然古籍版式差異會直接顯示在所掃描的影像上，但影像品質的好壞會影響數位影像的呈現，因此訂定「影像品質等級」作為數位影像測試的預測變項。採用標準迴歸分析將所有自變項納入回歸方程式中進行係數分析，以 SPSS 軟體取得分析數值如下（表 6）。

表 6

行數、行字數、影像品質與辨識率標準迴歸分析結果

自變項	多元相關係數 (R)	決定係數 (R^2)	ΔR^2	ANOVA F 值	F 值顯著性	T 值	T 值顯著性	容忍度	VIF
影像品質						-14.468	.000	.984	1.016
行字數	.744	.554	.547	80.994	.000	-3.850	.000	.698	1.432
行數						.407	.684	.695	1.439

結果顯示多元相關係數 (R) 為 0.744，決定係數 (R^2) 為 0.554，顯示三種自變項影像品質、行字數、行數可以解釋依變項辨識率的 55.4%；根據 ANOVA 分析結果，可得三者自變項最少有一種會對依變項有顯著性影響 ($F=80.994$, $p=.000<.05$)；係數分析中影像品質與行字數皆具有顯著性影響 ($t=-14.468$, $p=.000<.05$; $t=-3.850$, $p=.000<.05$)，反之行數不具有顯著性影響 ($t=.407$, $p=.684>.05$)，顯示影像品質與行字數對於 OCR 辨識率具有因果預測相關性，反之行數對於 OCR 辨識率較不具因果預測相關性；共線性診斷中三種自變項的容忍度皆高於 .01 且 VIF 值皆小於 10，代表三種自變項不存在多元共線性的問題，變項間不具相互影響性而多元迴歸分析成立。

綜合統計分析結果，可知三種自變項皆與辨識率呈負相關，亦即若希望辨識率越高，則在古籍選擇上盡量挑選行數、行字數較低的版式，才較

能獲得較佳的辨識效果；影像品質上也盡量挑選品質較佳的數位影像，透過簡單目視得以輕易判斷影像優劣，若影像未經後處理加工，使得內容文字具有色澤不全、雜訊過多等情形，則不考慮進行 OCR 辨識，挑選古籍影像辨識上以版面清晰、文字清楚不模糊為優先考量。透過相關性分析，雖然所挑選的三個自變項皆與辨識率呈現顯著性負相關，但進行多元迴歸分析時，僅有「影像品質」與「行字數」有預測相關性，代表若想預期提升 OCR 辨識率結果，在挑選古籍影像時應優先考量「影像品質」與「行字數」兩項，版式的考量上可以忽略「行數」的多寡，因為對於提升辨識率比起其他兩個變項較無顯著影響性。

二、常見辨識錯誤文字分析

本研究將古籍影像經 OCR 辨識結果，逐頁逐字檢視標記，將辨識文字與古籍影像相互比較，發現在原古籍中有 1,565 個字被 OCR 辨識錯誤為 1,094 個誤判文字。將這些容易被誤判文字特徵加以歸納分析，可供未來 OCR 進行調整修正的參考。本研究將辨識錯誤的文字類型依其形體特徵與原因，歸納為六大類：字形部件類似、字形外觀相近、字形拆開辨識、文字筆畫差異、繁體以簡體辨識、古今異字差異等，列舉說明如下。

(一) 字形部件類似

部件為構成中文漢字的基本單位，是介於「筆畫」與「部首」之間的存在，對中文漢字來說具有表達形體、表達詞義、體現字音、指示位置、替代文字等功能（黃沛榮，2009；莊德明、鄧賢瑛，2009）。此類型辨識錯誤通常發生在兩者文字具有相同部首、不同組成部件的文字間，或不同部首、相同搭配部件的文字間，例如將「孔」辨識成「扎」，其右邊部件「乚」相同；將「例」辨識成「側」，其左右部件「亻」與「亻」等相同情形。

由於中文漢字是由不同的部件組合而成，部件的隨意組合可建構出不同型態的中文漢字，而具代表性的部件為各文字的「部首」，部份部首等同單一部件，例如「口」、「土」、「山」等，但並非是將部首與部件劃上等號，有些部首是由兩種或以上的部件所組合而成的，例如香部是由「禾」與「日」兩種部件所組成、鼻部是「自」、「田」、「升」三種部件所組成（黃沛榮，

2009)。因此，不能單純地藉由部首的劃分來判斷部件種類。

由於中文漢字是由一種以上部件組合而成，為避免分類爭議，本研究在分類判斷上參考語言文字規範（GF0014-2009）與漢字構形資料庫的部件拆分原則，進行部件判斷時遵守：「符合理據」即字形結構符合理據者才得視為拆分部件；「重疊不拆」即筆畫有交疊者視為單一部件而不細拆分；「不成他字」即部件可組成為他字部件者才視為拆分部件（中華人民共和國教育部，2009；莊德明、鄧賢瑛，2009）。本研究除了參考以上兩種部件拆分判斷標準外，鑒於部首與部件間的關係，若兩者文字具相同部首則優先歸類於此；若不具相同部首但具相同主要部件也歸類於此，例如「比」與「此」雖然部首不同，但部件「匕」為整體文字的主要部件，因而歸類於此。此類型辨識錯誤字形舉例如下（表7）。

表 7

字形部件類似之錯誤文字例示對照表

次序	影像 文字	辨識 文字	次序	影像 文字	辨識 文字	次序	影像 文字	辨識 文字
1	也	地	11	北	此	21	仲	忡
2	也	咆	12	北	比	22	任	件
3	今	含	13	卯	印	23	任	住
4	仍	奶	14	句	勺	24	休	林
5	孔	扎	15	召	各	25	先	洗
6	比	此	16	可	珂	26	全	企
7	水	冰	17	台	合	27	匈	勾
8	仙	灿	18	右	古	28	印	即
9	令	冷	19	弘	私	29	各	洛
10	以	似	20	瓜	爪	30	合	各

(二) 字形外觀相近

將古籍文字與 OCR 辨識文字比對後，如果兩者間在構成組合上雖無相同部件，但其文字外觀樣貌上有相似性，導致可能產生辨識錯誤等情況者，皆歸類於此類型，包含「帶」辨識成「蒂」，將「萬」辨識成「禹」等相同情形。此類型辨識錯誤通常發生於兩者文字不管在外觀形式或是構字筆劃規則上皆有其相似性，迫使 OCR 軟體無法藉由文字外形去判斷文字種類，使得辨識上容易產生此類型的錯誤發生。此類型辨識錯誤字形舉例如下（表 8）。

表 8

字形外觀相近之錯誤文字例示對照表

次序	影像 文字	辨識 文字	次序	影像 文字	辨識 文字	次序	影像 文字	辨識 文字
1	七	匕	11	大	六	21	予	于
2	乃	巧	12	子	卞	22	云	么
3	人	八	13	子	于	23	云	亏
4	入	人	14	子	手	24	五	玉
5	八	人	15	子	予	25	五	丑
6	匕	也	16	子	干	26	元	无
7	上	卜	17	小	卜	27	元	尤
8	久	允	18	中	巾	28	內	丙
9	也	心	19	丹	舟	29	內	向
10	大	火	20	之	工	30	兮	今

(三) 字形拆開辨識

部件可依是否具備獨立成字的特性，分為成字部件與非成字部件，成字部件為可單獨存在且具字義的部件，例如「另」、「吉」、「唱」中的部件

「口」；非成字部件為不可單獨存在且不具字義的部件，例如「疾」、「病」、「疼」中的部件「疒」（中華人民共和國教育部，2009）。將古籍影像文字與 OCR 辨識文字相互比較，發現兩者間除有部分部件相同外，且相同部件即使單獨存在仍具有文字意涵，在辨識時由於部份部件屬於成字部件，而產生部份或整體部件被分開辨識的情況，皆歸類於此類型，例如「家」辨識成「豕」、「志」辨識成「士」與「心」等相同情形。

由於古籍書寫採由右至左、由上至下的方式，使得在辨識影像文字時因書寫方向而產生錯誤，尤其是發生在文字由兩至三個部件組成、構成方式採上下分布，且有一至二個具單獨字義的成字部件時，容易發生此類型的辨識錯誤。因部件間的空白區塊，使得軟體在判別文字時產生錯誤判斷，產生單一文字被辨識成兩個文字的情況。此類辨識錯誤字形類型舉例說明如下（表 9）。

表 9
文字拆開辨識之錯誤字形例示對照表

影像文字	辨識文字	說明
吾	五 口	「吾」由兩個獨立部件所構成，辨識容易產生辨識成兩種文字。
昏	氏 日	「昏」由兩個獨立部件所構成，辨識時容易產生辨識成兩種文字。
否	不 口	「否」雖然由兩個部件相連接，但辨識影像若線條不連續容易產生辨識成兩種文字。
天	一 大	「天」雖然為單一部件，但可拆成兩個不同字，當辨識時影像線條若不連續容易產生辨識成兩種文字。
葦	甘 耳	「葦」雖然為三個部件所構成，但上方部件「艹」與「口」由於過於接近，造成形式差異的辨識錯誤，下方「耳」則占整體字比例較大，而沒產生錯誤部件的辨識。

（四）文字筆畫差異

筆畫是構成中文漢字的最小書寫單位，依照筆畫的複雜程度可區分成

三種級別的筆畫：一級筆畫為最為簡單的筆畫，包含最基本的點（丶）、橫（一）、直（丨）、撇（丿）、捺（㇏），為單方向的書寫；二級筆畫相較於一級筆畫增加了方向的改變，多了轉折與鉤等書寫元素；三級筆畫為最為複雜的筆畫類型，相較於二級筆畫又添加了轉彎、鉤折等書寫元素（于惠泉，2008）。將古籍影像文字與 OCR 辨識文字相互比較，如果是兩者文字差異僅在細節筆畫上有所不同，則歸類於此類型，例如將「刀」辨識成「力」、「九」辨識成「丸」等。

歸於此類是在判斷標準上採一筆畫的差異，若文字間僅一筆畫有所不同則歸類於此，但若差異不只一筆畫則不歸類於此，例如將「玉」辨識錯誤成「五」，雖然目視感覺僅相差一筆畫，但其實「玉」中間筆畫為一級筆畫「橫」（一）與「點」（丶），而「五」中間筆畫為二級筆畫「橫折」（乚），兩者具有兩種筆畫的差異，因此不歸納於此類型。此類型辨識錯誤通常發生在筆畫較少的字形上，造成因素可能是影像品質不佳，使得數位影像上有斑點、黑漬等，造成 OCR 軟體將斑污點視為文字線條的一部分，導致辨識產生錯誤。此類型辨識錯誤字形類型舉例如下（表 10）。

表 10

文字筆畫差異之錯誤字形例示對照表

次序	影像文字	辨識文字	次序	影像文字	辨識文字	次序	影像文字	辨識文字
01	一	二	11	刃	刀	21	王	士
02	九	丸	12	千	干	22	天	大
03	几	凡	13	大	太	23	天	夭
04	刀	力	14	子	了	24	夫	大
05	力	刀	15	尸	戶	25	夭	夫
06	又	叉	16	干	千	26	夭	天
07	下	卞	17	互	互	27	心	必
08	于	干	18	什	仕	28	木	本
09	兀	元	19	今	令	29	王	玉
10	凡	几	20	公	么	30	令	今

(五) 繁體以簡體辨識

中文字碼區分成簡體字與繁體字，經簡化而產生的字形，可能由於外觀形式與筆劃較少的繁體字有所相似，而產生辨識成不同字義但有類似樣貌的簡體字，包含「淡」辨識成「谈」、「卒」辨識成「车」等。此類型錯誤所產生的簡體字會遵循與原影像文字相似樣貌的特性，而辨識為該簡體字所對應的繁體字，由於其簡化的特性，必定比原影像文字的筆劃還更為多劃。此錯誤辨識類型分析舉例說明如下（表 11）。

表 11

繁體以簡體辨識之錯誤字形例示對照說明表

影像文字	辨識簡體	對應繁體	說明
尤 ----- 充	龙	龍	兩種文字由於與簡體字「龙」具有形式差異上的關係特性，容易被辨識成錯誤字形，而「龙」所對應繁體字「龍」在構造上比起兩文字複雜許多。
同	间 ----- 閭	間 閭	相同文字與簡體字「间」、「閭」有形式差異上的關係特性，容易被辨識成錯誤字形，而兩簡體字所對應的繁體字「間」「閭」則有部件差異的關係特性。
衣 ----- 求	农	農	兩種文字與簡體字「农」有形式差異上的關係特性，容易被辨識成錯誤字形，但兩者與對應繁體字「農」並無關係特性，基本上不會辨識錯誤成該繁體字形。
惟 ----- 淮	谁	誰	兩種文字由於與簡體字「谁」具有形式差異上的關係特性，容易被辨識成錯誤字形，而「谁」與對應繁體字「誰」變化性不大，使得兩種文字不管是對應「谁」或者「誰」皆具有部件差異的關係特性。

(六) 古今異字差異

明代時期所流行的字形與現代使用的字形有所差異，此類泛稱的異體

字是指雖然兩者讀音、意義上相同，但寫法樣貌呈現上有所不同的漢字(王雅萍，謝筱琳，2011)。由於版刻文字的古今文字外觀的差異，影響 OCR 在辨識上無法找出相似字碼，而辨識錯誤成與古體字的樣貌相似，但不屬於相同意義脈絡的現今字形，例如將「參」之異體字「叅」辨識成「恭」，將「窗」之異體字「窻」辨識成「蔥」等。

這種類型錯誤通常發生在古今字形樣貌形式差異過大，使得 OCR 在辨識判別上依據線條、外觀等因素，辨識成相似的現今字形，由於古異體字多數不通用於現今，使得多數輸入法並無異體字的書寫方式，為能列舉說明，故將異體字樣貌參考國際電腦漢字及異體字知識庫 (<http://chardb.iis.sinica.edu.tw/>) 與教育部異體字字典 (<http://dict2.variants.moe.edu.tw/variants/rbt/home.do>) 收錄之異體字影像例示說明如下(表 12)。

表 12

古今異字差異之錯誤字形例示對照說明表

古異體字	現今字形	錯誤辨識文字	說明
踈	疏	疎、踩、睬	所辨識錯誤文字三者與異體字有部件差異的關係特性。
叅	參	恭	所辨識錯誤文字「恭」與異體字有部件差異的關係特性。
窻	窗	葱	所辨識錯誤文字「葱」與異體字有部件差異的關係特性。
绿	綠	緣	所辨識錯誤文字「緣」與異體字有形式差異的關係特性。
懽	歡	灌	所辨識錯誤文字「灌」與異體字有部件差異的關係特性。
獯	猿	狙	所辨識錯誤文字「狙」與異體字有部件差異的關係特性。
鷓	鴟	鷄	所辨識錯誤文字「鷄」與異體字有部件差異的關係特性。

綜合前述分析古籍影像文字與辨識文字之間的關聯性，依照文字特性所歸納之六種古籍文字辨識錯誤類型，在總數 1,094 個 OCR 誤判字中，以「字形部件類似」、「字形外觀相近」兩類最多(表 13)。由於同一誤判字

有可能，在歸納特徵時可分屬不同錯誤類型，故採計分別計算方式，例如：「也」會辨識錯誤成「地」與「心」，「萬」辨識錯誤成「禹」與「寓」。當「也」辨識為「地」、「萬」辨識為「寓」是因相同部件，屬於辨識錯誤類型中的「字形部件類似」；但「也」被辨識成「心」、「萬」視為「禹」時，則屬於辨識錯誤類型中的「字形外觀相近」。計量時採分開歸類方式，故錯誤次數加總為 1,353 次，百分比以錯誤次數除以總字數 1,094 字，以瞭解該類型錯誤占總字數百分比。

表 13

辨識錯誤文字類型次數及百分比 (N=1094)

錯誤辨識類型	辨識錯誤次數	占整體百分比(%)
字形部件類似	734	67.1
字形外觀相近	380	34.7
字形拆開辨識	78	7.1
文字筆劃差異	65	5.9
繁體以簡體識別	64	5.9
古今異字差異	32	3.0

根據上述統計可知「字形部件類似」與「字形外觀相近」兩類型比例最多，即使其他四種類型，也與部件組成及字形樣貌脫離不了關係，可知經常辨識錯誤的字形與辨識出的文字具有部件與樣貌的關聯性，藉由這樣的關聯性可協助 OCR 軟體進行後端字集庫的升級，將具有相似部件或字形樣貌的文字列為參考字集，日後辨識時可以提供軟體校正判斷先自動選擇相關值較高的文字，能有助於提升 OCR 整體辨識的精確性。但不同年代文本異體字與字形顯然有所差異，故本研究以明代文集所整理之文字錯誤類型較適用於相近時期文本特性，並非建立通用型之字集庫。

伍、結論與建議

本研究根據實測過程分級並輔以業界專家諮詢意見，歸納運用 OCR 轉製古籍全文結果如下：

一、古籍版式與影像清晰度會影響 OCR 對於明代古籍文字辨識率

本研究為瞭解商務型 OCR 辨識古籍的成效，利用可採購取得之 OCR 進行測試分析，發現即使是市售有口碑之 OCR 產品對於古籍的辨識結果，仍是 3 成到 8 成不等的正確率，辨識正確率受到古籍版式和影像清晰度影響，以本實驗結果為例，明人文集版式行數介於 8-9 行，以及行字數為 18 字之辨識結果較佳，而影像品質則是與辨識正確率呈現正相關。亦即未來進行古籍 OCR 辨識，在古籍選擇上盡量挑選行數、行字數較低的版式，才較能獲得較佳的辨識效果；尤其「影像品質」與「行字數」能預測相關性，代表若想預期提升 OCR 辨識率結果，在挑選古籍影像時應優先考量「影像品質」與「行字數」兩項，至於「行數」的多寡，對於提升辨識率上比起其他兩個變項較無顯著影響性。

二、OCR 經常辨識錯誤的明代古籍文字有部件相同或字形相似的特性

將古籍影像內容與 OCR 辨識文字逐一比對，以文字學角度依據中文漢字有相同部件之字形，或是將相似樣貌之字形進行整理分析，本研究歸納列出六種經常辨識錯誤的類型，其中以「字形部件類似」與「字形外觀相近」類型所佔比例最多，可發現經常被辨識錯誤的文字與 OCR 識別產生的文字多具有部件與樣貌關聯的特性，藉由這樣的關聯性可協助 OCR 軟體進行後端字集庫的升級。由於 OCR 辨識核心所儲存的是字形的特徵值資料，並非文字本身。因此，事先準備相似文字分群對於 OCR 本身辨識過程沒有直接幫助，能有幫助的部份是在於後續協助辨識校正工作上，將相似文字分群可以幫助進行更為精細的校正處理，後續可以幫助系統在遇到相同錯誤時，能直接進行內容文字的校正。

三、採用古籍影像進行文字辨識可選擇影像品質較佳者優先進行

本研究結果與多數文獻探討論點相同，顯示影像品質的優劣程度對於 OCR 辨識的精確率產生高度影響性，進行 OCR 辨識之影像盡量挑選品質

較佳者，若影像未經後處理加工，使得影像內容有色澤不全、雜訊過多等情形，則不考慮進行 OCR 辨識，挑選古籍影像辨識上以版面清晰、文字清楚不模糊為優先考量。因此，典藏機構未來執行珍藏古籍數位全文化計劃時，可先對影像品質進行目視檢驗，如果影像未達標準，可先考慮用人工輸入方式轉置全文，而非選擇 OCR 辨識，因為影像品質好壞幾乎決定了辨識的精確程度。若古籍影像皆無內容模糊不清等情形，在選擇優先數位化順序上則考慮單行字數較少的古籍，單行字數少表示文字空白處較寬廣，文字大小也較為合適，對於 OCR 全文化的作業效率才能有所提升。

四、OCR 技術運用於古籍辨識的困難是古籍文字複雜和版式變化的問題

目前 OCR 辨識技術已可以辨識處理很多過往階段無法進行的項目，例如字體字形差異、版面格式差異等，但仍然需要透過技術人員針對辨識標的特性進行軟體調整，才能將辨識率明顯提升。但目前技術仍然有許多狀況尚無法克服，單以手寫字體為例，過於潦草且變化沒有規律的字體，仍然難以透過技術方式進行辨識突破，僅能以人為判讀進行逐字繕打，這部分的侷限也阻礙了典藏機構進行典籍全文化的選擇，即使許多古籍正文部份為版刻字體，但前後之序跋內容可能為藏書家或閱讀者之手寫字體，OCR 辨識在處理上無法完全應付，故目前典藏機構寧可採用人工繕打的方式，也不打算使用 OCR 辨識。

由於機構典藏之古籍有各種版式樣貌，物件狀況不一，不同字體、版式與影像清晰度會影響 OCR 辨識結果。雖然將典藏物件依版式整理並分別進行全文處理有助於 OCR 辨識的進行，但對機構而言，會添加分類整理的人力，但若不整理直接將數位化全文處理交由廠商，又會增加廠商作業處理費用提升的情況。因此，對於典藏機構進行全文轉置作業規劃，本研究根據實測過程觀察，並輔以諮詢業界專家意見，建議數位典藏全文化作業流程可從現有數位典藏影像中挑選需要全文化處理之物件影像；先評估物件影像狀況，若有影像不佳影響文字內容的判讀，或是缺漏狀況嚴重無法讀取內容之情形，需要先委託廠商進行重新掃描；若影像狀況佳者，且字形方正而雜訊較少，可採用 OCR 辨識。雖然近年來數位人文研究者對於漢籍全文存在使用需求，但古籍字形辨識的發展緩慢，研究者只能從

已有全文的資料庫開展研究主題，有全文的數量相對於浩瀚的古籍內容仍然較為侷限，如何提高 OCR 辨識率並發展低成本的全文轉置作業，是數位人文研究之際需要完成的基礎建設。本研究基於國家圖書館明人文集之全文轉置經驗，提供研究分析成果，期許對於運用 OCR 辨識類似古籍和選擇全文化的作業決策能有一些參考。未來進一步研究可針對不同 OCR 軟體及不同時代古籍版本進行比較分析，以累積運用 OCR 進行全文轉製更多面向之實證分析。

(接受日期：2020年12月14日)

參考文獻

- 于惠泉(2008)。《漢字造型規律及書寫技能》。鄭州市：河南美術。
- 中央研究院(2010)。《國際電腦漢字及異體字知識庫》。檢自：
<https://chardb.iis.sinica.edu.tw/>
- 中央研究院歷史語言研究所(2019)。《漢籍電子文獻資料庫》。檢自：
<http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm>
- 中華人民共和國教育部(2009)。《現代常用字部件及部件名稱規範》。北京市：國家語言文字工作委員會。
- 中華民國教育部(2017)。《教育部異體字字典》。檢自：
<https://dict.variants.moe.edu.tw/variants/rbt/home.do>
- 王雅萍、謝筱琳(2011)。《漢籍全文數位化工作流程指南》。臺北市：行政院國家科學委員會。
- 李清志(1985)。明代中葉以後版刻特徵。在吳哲夫編，《古籍鑑定與維護研習會專集》(頁96-121)。臺北市：中國圖書館學會。
- 周駿富(1985)。明代前期版刻特徵。在吳哲夫編，《古籍鑑定與維護研習會專集》(頁83-95)。臺北市：中國圖書館學會。
- 林巧敏(2017年11月)。古籍全文數位化經驗分享。在國家圖書館主持，《國家圖書館通用型古籍數位人文研究平台成果發表會》。國家圖書館主辦，臺北市，中華民國。
- 林巧敏、陳志銘(2017)。古籍風華再現：關於古籍數位人文平台之建置。《國家圖書館館刊》，106(1)，111-132。

- 國家圖書館 (2020)。古籍與特藏文獻資源。檢自：
<http://rbook.ncl.edu.tw/NCLSearch/>
- 張俊盛、陳舜德 (1995)。雜訊通道模型在 OCR 後處理之應用。影像與識別，3(3)，98-109。
- 莊德明、鄧賢瑛 (2009)。漢字構形資料庫的研發與應用。檢自：
<http://cdp.sinica.edu.tw/service/documents/T090904.pdf>
- 陳金木 (2008)。電子全文資料庫與學術研究—以《四部叢刊電子全文檢索版》為例。明道通識論叢，5，120-135。doi:10.6954/MJGE.200811.0120
- 曾元顯 (2004)。應用於資訊檢索的中文 OCR 錯誤詞彙自動更正。中國圖書館學會會報，72，23-31。
- 曾逸鴻、林裕淵 (2007)。中文文件影像中之特殊字體偵測。科學與工程技術期刊，3(4)，29-39。doi:10.7117/JSET.200712.0029
- 黃永年 (2005)。古籍版本學。南京：江蘇教育出版社。
- 黃沛榮 (2009)。漢字教學的理論與實踐。臺北市：樂學。
- 劉兆祐 (2007)。認識古籍版刻與藏書家。臺北市：臺灣學生書局。
- 潘美月 (1985)。明代官私刻書。在吳哲夫編，古籍鑑定與維護研習會專集 (頁 122-136)。臺北市：中國圖書館學會。
- 潘朝陽 (1994)。OCR/中文 OCR 技術。光學工程，47，48-53。
doi:10.30011/OE.199409.0008
- 駱偉 (2004)。簡明古籍整理與版本學。澳門：澳門圖書館暨資訊管理協會。
- 顧力仁 (2001)。中文古籍全文資料庫建置比較研究。國家圖書館館刊，90(2)，197-216。
- 顧力仁 (2002)。永樂大典數位化相關問題之探討：兼論資訊科技對古籍整理的影響。圖書館學與資訊科學，28(1)，33-48。
- ABBYY Production. (2017). ABBYY FineReader 14。Retrieved from
https://help.abbyy.com/static/guides/finereader/14/Guide_ChineseTraditional.pdf
- ABBYY. (2020). ABBYY Expert Talks. Retrieved from <https://www.abbyy.com/expert-talks/>
- Al-A'ali, M., & Ahmad, J. (2007). Optical character recognition system for Arabic text using cursive multi-directional approach. Journal of Computer Science, 3(7), 549-555. doi:10.3844/jcssp.2007.549.555
- Badoiu, V., Ciobanu, A. C., & Craitoiu, S. (2016). OCR quality improvement using

- image preprocessing. *Journal of Information Systems & Operations Management*, 10(1), 1-13.
- Balk, H., & Ploeger, L. (2009). IMPACT: Working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems & Services: International Digital Library Perspectives*, 25(4), 233-248.
doi:10.1108/10650750911001824
- Chapman, S., & Kenney, A. R. (1996). Digital conversion of research library materials: A case for full informational capture. *D-Lib Magazine*, 2(10). Retrieved from <http://www.dlib.org/dlib/october96/cornell/10chapman.html>
- Cojocaru, S., Colesnicov, A., Malahov, L., & Bumbu, T. (2016). Optical character recognition applied to Romanian printed texts of the 18th-20th century. *Computer Science Journal of Moldova*, 24(1), 106-117.
- Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4). Retrieved from <http://www.dlib.org/dlib/march09/holley/03holley.html>
- Mariner, M. C. (2010). Optical Character Recognition (OCR). In Bates, M. J. & Maack, M. N. (Eds.), *Encyclopedia of Library and Information Sciences* (3rd ed.). (pp.4037-4044). Boca Raton, Fla: CRC Press.
- Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE*, 80(7), 1029-1058. doi:10.1109/5.156468
- Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseract: A case study. *International Journal of Computer Applications*, 55(10), 50-56. doi:10.5120/8794-2784
- Sun, W., Liu, L. M., Zhang, W., & Comfort, J. C. (1992). Intelligent OCR processing. *Journal of the American Society for Information Science*, 43(6), 422-431. doi: 10.1002/(SICI)1097-4571(199207)43:6<422::AID-ASI3>3.0.CO;2-Z
- Zhu, Y., Tan, T., & Wang, Y. (2001). Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1192-1200. doi:10.1109/34.954608



Full Text Conversion Experience in Optical Character Recognition of Ancient Books: An Example of Ming Dynasty Literati Collections

Chiao-Min Lin* Han-Wei Tasi**

【Abstract】

Due to the development of information technology and the need for content analysis of digital humanities research, the use of optical character recognition technology (OCR) to convert contents into verbatim texts can facilitate full-text search and content exploration. In order to understand the feasibility of using the OCR software to convert the full text of the ancient books, this study used the ancient texts to conduct a measured analysis to explore the effectiveness of OCR identification and the reasons for the impact of text recognition. The study selected 40 different layouts and glyphs of Ming Dynasty ancient books for analysis. The results show that the ancient book layout and image quality would affect the OCR recognition rate. When the layout is too crowded and the image quality is blurred, it is not conducive to OCR recognition. This study summarized six common types

* Professor, Graduate Institute of Library, Information and Archival Studies, National Chengchi University

ORCID 0000-0002-9309-9884

Principal author for all correspondence E-mail: cmlin@nccu.edu.tw

** Coordinator, CTBC Financial Holding Co., Ltd.

E-mail: bestwiwi79@gmail.com

of identification error glyphs, which can provide the collection agencies to carry out the plan of the full text conversion of similar ancient books.

Keywords

Optical character recognition, Full-Text database, Old rare books, Ancient book digitization, Digital archive

【Summary】

1.Introduction

To cope with the development of information technology and the needs for digital humanities research on full-text content analysis, the application of optical character recognition (OCR) technology to convert contents into full texts could facilitate the use of full-text search and content exploration. The OCR of Chinese ancient books are discussed in this study. With measured analysis, Ming Dynasty literati collections in National Central Library are selected different ancient book formats to compare the factors in the accuracy of OCR recognizing formats and text types, in order to assist in selecting similar ancient books for full text conversion.

2. Research Methods

Commercial OCR software, “ABBYY FineReader 14” is applied in this study to test the text recognition of ancient books. The commercial software test could better conform to the current technological ability of institutional repository, rather than authorizing IT staff for the self-development. ABBYY FineReader could provide preliminary preprocessing of image cut page, tilt correction, and noise removal to help the experimental operation.

Ming Dynasty literati collections in National Central Library are used as the test target in this study, as Ming Dynasty collections present the academic

research needs for full-text recognition and the publishing peaked in Ming Dynasty, with emerging ancient books with different formats. The analysis of ancient books in the period could collect the OCR data of different formats. 40 volumes of collections are randomly sampled, and each volume is picked 5 pages for OCR.

3. Results

3.1 Line number and word number analysis

The line number of ancient books appears in 8-12 lines. In the comparison of line number and recognition rate, line number of 8-9 reveal higher recognition rate than the line number of 10-11, as text space between lines, with few line numbers, is more obvious and could benefit OCR separating text and background.

The 40 volumes of ancient books show the word number of 16-24 in each line. By observing the relationship between word number in each line and recognition rate, the gap between words is obvious in the format with few word numbers and the recognition result is better.

3.2 Image quality analysis

The image quality judgment is referred to “image specifications for archive digitalization” made by Academia Sinica. More than a half of ancient book samples in this study appear “image skew”, “uneven image color”, and “discontinued image text line, while the problem of “image with spots and stains” is less.

3.3 Effects of line number, word number, and image quality on recognition rate

Independent-Sample T test is used for judging the difference of independent variables (line number, word number, image quality) in the

dependent variable (recognition rate). The results show that the group with more lines appears lower recognition rate than the group with fewer lines, the group with more words shows lower recognition rate than the group with fewer lines, and the group with worse image quality presents lower recognition rate than the group with better image quality. In other words, ancient books with the formats of fewer lines and words, with higher recognition rate, could appear better recognition effect. Without post-processing, image text with uneven color and excessive noises are not suitable for OCR.

3.4 Analysis of common error recognition

After the ancient book images going through OCR, the verbatim text is compared with the ancient book images to classify text misidentified by OCR, according to the physical characteristics, including similar font radical, close font appearance, font decomposition recognition, difference in text stroke, traditional Chinese recognition with simplified Chinese, and difference between ancient and modern text. Among total 1094 misidentified words, most of them appear on “similar font radical” and “close font appearance”. Apparently, text composition and font appearance are the major factors in recognition rate.

4. Discussion and Conclusion

4.1 Ancient book formats and definition of images would affect the recognition rate of ancient book text with OCR. For ancient book OCR, the formats with lower line number and word number in a line would be first selected for the better recognition effect. To enhance the OCR result, “image quality” and “word number in a line” could be first considered in the selection of ancient book images, while “line number” does not show significant effects on the promotion of recognition rate.

4.2 Frequently misidentified ancient book text presents the characteristics of same radical or similar font. It is discovered in this study that frequently misidentified font reveal radical and appearance correlations with recognized

text. Such correlations could assist OCR software in upgrading the back-end word database and list words with similar radical or font appearance in the reference word set for providing words with higher relevance to enhance the OCR accuracy.

4.3 Images with higher quality could be selected for text recognition of ancient book images. When converting ancient book contents into full texts in the future, the image quality could be visually inspected. When the image quality does not achieve the standard, it can be evaluated to input verbatim text manually, rather than using OCR, as the image quality would determine the recognition accuracy.

Romanized & Translated Reference for Original Text

于惠泉（2008）。*漢字造型規律及書寫技能*。鄭州市：河南美術。【Yu, Hui-Quan (2008). *Han zi zao xing gui lv ji shu xie ji neng*. Zhengzhou: Henan mei shu. (in Chinese)】

中央研究院（2010）。*國際電腦漢字及異體字知識庫*。檢自：

<https://chardb.iis.sinica.edu.tw/> 【Academia Sinica (2010). *Guo ji dian nao han zi ji yi ti zi zhi shi ku*. Retrieved from <https://chardb.iis.sinica.edu.tw/> (in Chinese)】

中央研究院歷史語言研究所（2019）。*漢籍電子文獻資料庫*。檢自：

<http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm> 【Institute of History and Philology, Academia Sinica (2019). *Han ji dian zi wen xian zi liao ku*. Retrieved from <http://hanchi.ihp.sinica.edu.tw/ihp/hanji.htm> (in Chinese)】

中華人民共和國教育部（2009）。*現代常用字部件及部件名稱規範*。北京市：國家語言文字工作委員會。【Ministry of Education, The People's Republic of China (2009). *Xian dai chang yong zi bu jian ji bu jian ming cheng gui fan*. Beijing: guó jiāyǔ yánwén zìgōng zuòwěi yuán huì (in Chinese)】

中華民國教育部（2017）。*教育部異體字字典*。檢自：

<https://dict.variants.moe.edu.tw/variants/rbt/home.do> 【Ministry of Education, Republic of China (2017). *Jiao yu bu yi ti zi zi dian*. Retrieved from <https://dict.variants.moe.edu.tw/variants/rbt/home.do> (in Chinese)】

- 王雅萍、謝筱琳（2011）。*漢籍全文數位化工作流程指南*。臺北市：行政院國家科學委員會。【Wang, Ya-Ping, & Xie, Xiao-Lin (2011). *Han ji quan wen shu wei hua gong zuo liu cheng zhi nan*. Taipei: National Science Council, Executive Yuan. (in Chinese)】
- 李清志（1985）。明代中葉以後版刻特徵。在吳哲夫編，*古籍鑑定與維護研習會專集*（頁 96-121）。臺北市：中國圖書館學會。【Li, Qing-Zhi (1985). Ming dai zhong ye yi hou ban ke te zheng. In Wu, Zhe-Fu (Ed.), *Gu ji jian ding yu wei hu yan xi hui zhuan ji* (pp. 96-121). Taipei: Library Association of China. (in Chinese)】
- 周駿富（1985）。明代前期版刻特徵。在吳哲夫編，*古籍鑑定與維護研習會專集*（頁 83-95）。臺北市：中國圖書館學會。【Zhou, Jun-Fu (1985). Ming dai qian qi ban ke te zheng. In Wu, Zhe-Fu (Ed.), *Gu ji jian ding yu wei hu yan xi hui zhuan ji* (pp. 83-95). Taipei: Library Association of China. (in Chinese)】
- 林巧敏（2017 年 11 月）。古籍全文數位化經驗分享。在國家圖書館主持，*國家圖書館通用型古籍數位人文研究平台成果發表會*。國家圖書館主辦，臺北市，中華民國。【Lin, Chiao-Min (2017, November). Gu ji quan wen shu wei hua jing yan fen xiang. In National Central Library (Chair), *Guo jia tu shu guan tong yong xing gu ji shu wei ren wen yan jiu ping tai cheng guo fa biao hui*. National Central Library. Taipei, Republic of China. (in Chinese)】
- 林巧敏、陳志銘（2017）。古籍風華再現：關於古籍數位人文平台之建置。*國家圖書館館刊*，106(1)，111-132。【Lin, Chiao-Min, & Chen, Chih-Ming (2017). Revitalizing the splendor: The construction of digital humanities platform on Chinese ancient books. *National Central Library Bulletin*, 106(1), 111-132. (in Chinese)】
- 國家圖書館（2020）。*古籍與特藏文獻資源*。檢自：
<http://rbook.ncl.edu.tw/NCLSearch/> 【National Central Library (2020). *Gu ji yu te cang wen xian zi yuan*. Retrieved from <http://rbook.ncl.edu.tw/NCLSearch/> (in Chinese)】
- 張俊盛、陳舜德（1995）。雜訊通道模型在 OCR 後處理之應用。*影像與識別*，3(3)，98-109。Chang, Jun-Sheng, & Chen, Shun-Der (1995). Za xun tong dao mo xing zai OCR hou chu li zhi ying yong. *Images & Recognition*, 3(3), 98-109. (in Chinese)】

- 莊德明、鄧賢瑛（2009）。漢字構形資料庫的研發與應用。檢自：
<http://cdp.sinica.edu.tw/service/documents/T090904.pdf> 【Zhuang, De-ming, &
Deng, Xian-Ying (2009). *Han zi gou xing zi liao ku de yan fa yu ying yong*.
Retrieved from <http://cdp.sinica.edu.tw/service/documents/T090904.pdf> (in
Chinese)】
- 陳金木（2008）。電子全文資料庫與學術研究——以《四部叢刊電子全文檢索版》為
例。《明道通識論叢》，5，120-135。【Chen, Chin-Mu (2008). Electronic version
of text retrieving of collection and academic research: Take electronic version of
text retrieving of collection of the four traditional divisions of texts as an
illustrative example. *MingDao Journal of General Education*, 5, 120-135. (in
Chinese)】 doi:10.6954/MJGE.200811.0120
- 曾元顯（2004）。應用於資訊檢索的中文 OCR 錯誤詞彙自動更正。《中國圖書館學
會會報》，72，23-31。【Tseng, Yuen-Hsien (2004). Error correction of Chinese
OCR texts for information retrieval. *Bulletin of the Library Association of China*,
72, 23-31. (in Chinese)】
- 曾逸鴻、林裕淵（2007）。中文文件影像中之特殊字體偵測。《科學與工程技術期
刊》，3(4)，29-39。【Tseng, Yi-Hong, & Lin, Yu-Yuan (2007). Special typeface
identification in Chinese document images. *Journal of Science and Engineering
Technology*, 3(4), 29-39. (in Chinese)】 doi:10.7117/JSET.200712.0029
- 黃永年（2005）。《古籍版本學》。南京：江蘇教育出版社。【Huang, Yong-Nian
(2005). *Gu ji ban ben xue*. Nanjing: Jiangsu jiao yu chu ban she. (in Chinese)】
- 黃沛榮（2009）。《漢字教學的理論與實踐》。臺北市：樂學。【Huang, Pei-Rong
(2009). *Han zi jiao xue de li lun yu shi jian*. Taipei: Le xue. (in Chinese)】
- 劉兆祐（2007）。《認識古籍版刻與藏書家》。臺北市：學生書局。【Liu, Zhao-You
(2007). *Ren shi gu ji ban ke yu cang shu jia*. Taipei: Student Book. (in
Chinese)】
- 潘美月（1985）。明代官私刻書。在吳哲夫編，*古籍鑑定與維護研習會專集*（頁
122-136）。臺北市：中國圖書館學會。【Pan, Mei-Yue (1985). Ming dai
guan si ke shu. In Wu, Zhe-Fu (Ed.), *Gu ji jian ding yu wei hu yan xi hui zhuan ji*
(pp. 122-136). Taipei: Library Association of China. (in Chinese)】
- 潘朝陽（1994）。OCR/中文 OCR 技術。《光學工程》，47，48-53。【Pan, Chao-Yan
(1994). OCR/Chinese OCR technology. *Optical Engineering*, 47, 48-53. (in

- Chinese)】 doi:10.30011/OE.199409.0008
- 駱偉 (2004)。《簡明古籍整理與版本學》。澳門：澳門圖書館暨資訊管理協會。
- 【Luo, Wei (2004). *Jian ming gu ji zheng li yu ban ben xue*. Macao: Macao Library and Information Management Association. (in Chinese)】
- 顧力仁 (2001)。中文古籍全文資料庫建置比較研究。《國家圖書館館刊》，90(2)，197-216。【Ku, Li-Jen (2001). Zhong wen gu ji quan wen zi liao ku jian zhi bi jiao yan jiu. *National Central Library Bulletin*, 90(2), 197-216. (in Chinese)】
- 顧力仁 (2002)。永樂大典數位化相關問題之探討：兼論資訊科技對古籍整理的影響。《圖書館學與資訊科學》，28(1)，33-48。【Ku, Li-Jen (2002). Exploration of the relating problems in digitization of Yung Lo Encyclopaedia; The impact of information technology to the Chinese ancient books. *Journal of Library and Information Science*, 28(1), 33-48. (in Chinese)】
- ABBYY Production.(2017). *ABBYY FineReader 14*。Retrieved from https://help.abbyy.com/static/guides/finereader/14/Guide_ChineseTraditional.pdf
- ABBYY.(2020). *ABBYY Expert Talks*. Retrieved from <https://www.abbyy.com/expert-talks>
- Al-A'ali, M., & Ahmad, J. (2007). Optical character recognition system for Arabic text using cursive multi-directional approach. *Journal of Computer Science*, 3(7), 549-555. doi:10.3844/jcssp.2007.549.555
- Badoiu, V., Ciobanu, A. C., & Craitoiu, S. (2016). OCR quality improvement using image preprocessing. *Journal of Information Systems & Operations Management*, 10(1), 1-13.
- Balk, H., & Ploeger, L. (2009). IMPACT: Working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems & Services: International digital library perspectives*, 25(4), 233-248. doi:10.1108/10650750911001824
- Chapman, S., & Kenney, A. R. (1996). Digital conversion of research library materials: A case for full informational capture. *D-Lib Magazine*, 2(10). Retrieved from <http://www.dlib.org/dlib/october96/cornell/10chapman.html>
- Cojocaru, S., Colesnicov, A., Malahov, L., & Bumbu, T. (2016). Optical character recognition applied to Romanian printed texts of the 18th-20th century. *Computer Science Journal of Moldova*, 24(1), 106-117.

Full Text Conversion Experience in Optical Character Recognition of
Ancient Books: An Example of Ming Dynasty Literati Collections

- Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4). Retrieved from <http://www.dlib.org/dlib/march09/holley/03holley.html>
- Mariner, M. C. (2010). Optical Character Recognition (OCR). In Bates, M. J. & Maack, M. N. (Eds.), *Encyclopedia of Library and Information Sciences* (3rd ed.). (pp.4037-4044). Boca Raton, Fla: CRC Press.
- Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE*, 80(7), 1029-1058. doi:10.1109/5.156468
- Patel, C., Patel, A., & Patel, D. (2012). Optical character recognition by open source OCR tool tesseraact: A case study. *International Journal of Computer Applications*, 55(10), 50-56. doi:10.5120/8794-2784.
- Sun, W., Liu, L. M., Zhang, W., & Comfort, J. C. (1992). Intelligent OCR processing. *Journal of the American Society for Information Science*. 43(6), 422-431.
- Zhu, Y., Tan, T., & Wang, Y. (2001). Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1192-1200. doi:10.1109/34.954608