

字級及掃描解析度與中文光學文字辨識系統辨識率關係之研究—以蒙恬認識王專業版為例

指導教授：陳昌郎 老 師
研究學生：施軍宇 辛炳宏
黃英銓 羅子堯

摘要

現今這資訊爆炸的社會中，早期記錄保存文字資料的方式已不再適合，尤其在現今講求效率、速度及精確的工作品質下，我們會透過中文光學文字辨識系統將紙本資訊轉為文字檔案，將文字作數位歸檔、編排及整理，除了可防止印刷品老化流失甚至失傳，還可將文字資訊作更有彈性的運用，例如書籍翻新、網路傳播，甚至強化學術界的資訊交流機制。

本研究主要的目的在於研究市售的專業版中文辨識軟體，對於一般書籍中文字辨識能力，本研究採用蒙恬認識王專業版，對常見內文字進行辨識能力測試，並探討其『辨識能力』、『掃描解析度』、『字級大小』以及與『掃描解析度與字級乘積大小』的關係，藉此瞭解最佳辨識效果的字級與掃瞄解析度組合，例如當辨識的文件是 12 級字時，以 300DPI 的解析度掃瞄後再進行辨識，即可得到最好的辨識率，如此一來可讓我們在選擇掃瞄解析度時有客觀的依據，也可節省文稿掃瞄及檔案存取時間。

在早期的研究中，可以得知中文光學文字辨識系統在於字級、掃瞄解析度和辨識結果有著顯著的相關聯性，因此本研究將同一字型，不同字級影像文件加以辨識，取得各字級最好之辨識率，這樣一來當我們拿到有需辨識的文件時，便可以直接對文件加以判斷，得到最好的辨識效果。

壹、緒論

一、問題陳述

從人類開始會思考及記錄的那一刻起，圖與文字便存在於我們日常生活中，藉著圖與文字的流傳及對於歷史文化的傳承，刻畫出一部人類文明的發展史。為了延續文化的傳承、新聞的傳播、知識的傳遞及

學術性文章等，將印刷品上的文字及其他載體上的文字，轉換成一種可供辨識的文字類型，是未來不可缺的趨勢。

現今這資訊爆炸的社會中，早期記錄保存文字資料的方式已不在適合，尤其在現今講求效率、速度及精確的工作品質下，我們所需要的是一套能夠減少人力資源、提高整體效率、快速記錄儲存、降低錯誤率的光學



辨識系統，而這辨識系統可以精確的轉換文字檔案，成為可以被電腦解讀的文字編碼，因此中文光學文字辨識系統在這一點上，便能扮演著一個快速輸入文字的打字員，一位快速轉換編碼的工程設計師，它不僅能快速的把文字檔轉換成可供編排的數位檔案，也可將數量很繁重的文字或歷史性的學術文件整理歸檔，以防止印刷品老化流失或檔案失傳。且在於檔案的類型上，完整的圖片檔案（BMP）的資料量遠大於一般數位文字檔，因此中文光學文字辨識系統也可以減少系統資源的容量，而辨識後的數位文字檔可以快速的經由網路傳遞至各處。

但就現實面來說，中文光學文字辨識系統的辨識率之好與壞，便是影響著是否能超越打字數度，增加文件輸入的效率，而當一份文件需經過中文光學文字辨識系統的處理時，需掃瞄多少解析度，其解析度，是否因文件的字型及字級去調整他的掃瞄解析度。因此本研究希望可以將複雜的像數換算的公式簡化，變成字級對於掃瞄解析度的簡易方式，相對的有了這公式，當有一份需辨識的文件時，可以經由文件字級直接知道需掃瞄的解析度為何，可以更有效率的得到最好的辨識率。

二、研究目的

本研究最主要的目的在於了解，市售的中文光學文字辨識系統對於文章內文、字級與掃瞄解析度影響正確辨識率的關係。由於中文光學文字辨識系統對書籍文字的辨識過程中，掃描器的掃描解析度會影響辨識影像的清晰度，進而對辨識能力造成影響，因此掃描解析度是需探討的課題。此外，字級大小也與掃描後的文字影像畫素息息相關，所以字級大小也不可忽視。由於字級與掃描解析度對中文光學文字辨識系統的辨識能力，在相關文獻中已確定有顯著關係，

因此本研究將進一步探討字級與掃描解析度的交互關係，在各組字級和掃瞄解析度的組合中找出最好的辨識率，在將其訂為一個標準，將所辨識的文件影像縮放置這最佳組合的條件中，探討是否增加其辨識的能力。簡而言之，本研究目的有二，其一為瞭解各字級其最佳辨識率的掃瞄解析度，其二為找出縮放文字影像的最佳範圍，這樣一來，便可以將大字級或是小字級文字，有技巧的縮放成最佳辨識率的文件。

三、問題的重要性

現在市面上的中文光學文字辨識系統，隨著掃描機的普及且研發辨識系統的廠商增加，可以了解到，將繁雜的文字輸入過程轉化成經由辨識系統辨識，是在現今社會中需要被加以利用的。在現今資料來源快速且繁多的資訊時代中，中文光學文字辨識系統在於應用上，卻為少數且專業的人所了解、運用，其問題來自於大部分人，對於文件的儲存輸入還是利用人工一字一字的編輯輸入，這樣子是無法降低真正的文件處理編輯的效率，且浪費人力資源。其中有一個關鍵點其實是在於人們對中文光學文字辨識系統的不信任，他們認為文字經過掃描再加以辨識成數位文字檔的辨識率很低，或是不知道如何運用中文光學文字辨識系統，使它發揮最高辨識率，因此寧願選擇最花費時間的打字輸入方式。

但反觀現今中文光學文字辨識系統在某一些層面上也跟不上多樣字體、多樣創新文字出現的速度，而就文件內文而言，這幾年各種字體、各種字型大量產生，相對於中文辨識系統而言，對於這一些新的內文常用字型是否有同樣的辨識率，如果我們可以對於最常用內文字型其辨識率作一個比較，找出內文字型最適合的辨識方式，或是一個辨識公式，對於需文字辨識的人們提供一個可



以參考的依據，解決之前文章所敘述到的不信任感及無法達到最高辨識率的問題，這樣一來，需要文字辨識的人將可以利用市售的軟件減少人力文件輸入且提高工作效率。

在辨識系統中佔有其重要部分的中文字型，經由銷售軟件的公司販售或網路的流通，最後以單純文字的方式被印載於市面上的印刷品，如果這份印刷文件需要被大量備份或是資料存檔交由網路世界交流，製造書籍與印刷品的製造者便會站在一個文件可以容易被辨識的立足點，來選擇我們文件所要使用的字型，相對於中文光學文字辨識系統的使用者在辨識一份文件時也可由字型來探討，此份文件是否要經過辨識，還是交由其他方法來做一個複製或存取的工作。由此可見如果先對文字『字型』、『字級』和『掃描解析度』的關係有所了解，便可以達到更有效率且良好辨識率的狀況。

整體來討論中文辨識系統在於文字的輸入上的重要性，就以國科會的數位典藏為例，利用中文光學文字辨識系統可以快速的將一些文字備份，使得一些將失去的文字利用辨識系統轉成數位化，讓那些令人感動的唐詩、宋詞或是一些偉大作家的作品保留給後世，且成為可以利用之數位文字檔。因此在製作國際網頁或是文字傳遞時，中文光學文字辨識系統所辨識出的數位文字檔對文化及學術上有著重大的含意。

站在數位典藏的觀點來看中文光學文字辨識系統，數位內容產業的商機龐大，在二十一世紀的國際主流上已經形成一個顯著的商業模式，各國也都爭相投入資源積極發展，相對的，我國政府也積極調整產業升級轉型步調，鎖定數位內容產業為「兩兆雙星」產業之一。行政院在 2002 年 5 月核定通過加強數位內容產業發展推動方案，積極扶植台灣數位內容產業整體發展。同時在成立數位圖書館上，更強化其 24 小時隨時上

網即可找到所需要的書，而且一本書可以重複在網路上點閱，不用怕書不在圖書館或是被借走，更可以將國家內部的重要文獻和世界接軌，讓大量文字快速的經過辨識率高的字級大小及掃描解析度，快速且精準的辨識加以存檔傳遞。

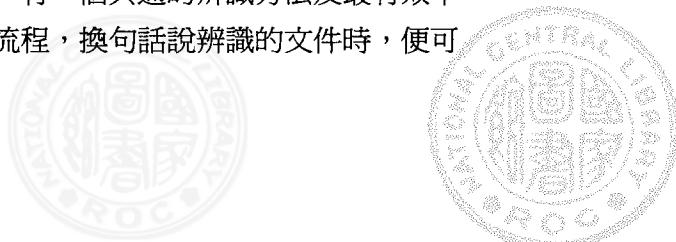
四、研究動機

中文光學文字辨識系統，簡而言之是將紙本或其他平面媒體的文字轉換成一般電腦可以儲純的數位文字檔。

市面上現存的專業版中文辨識系統，普遍的中文光學文字辨識系統，雖然在偵測辨識方面都有宣稱百分之 95 的辨識率，但在現今文字文件大量複製，大量創新的時代，大部分的辨識系統廠商所謂的辨識率只限於 BIG5 常用中文字集 5401 字。

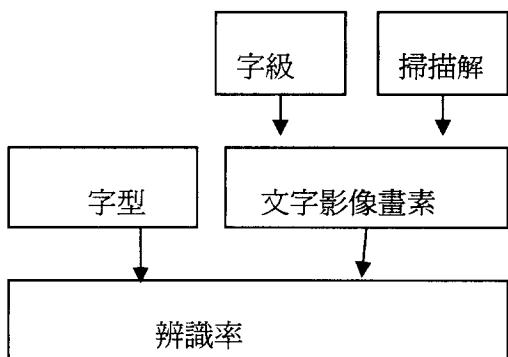
以往在中文光學文字辨識系統有論文研究的前輩們，對於中文光學辨識系統都有其研究，但是大多數研究並沒有提出具體數據，對於各字型及字級相對應的掃描解析度也只有設定在幾個少數特定數值，因此本研究希望在字型、字級及掃描解析度上作一個全盤的分析，以提供更詳細的數據，讓使用中文光學辨識系統的使用者，在使用時有一個依據。因此本研究希望藉由市面上專業的中文光學文字辨識系統，對一般書籍的內文常用字型、字級加以辨識，以分析比較其字型的辨識率、掃描解析度、字級大小與影像畫素之間辨識率，求得量化數據，作為掃描書籍文字時的參考。

這樣一來本研究便可以在中文光學文字辨識系統的文字辨識上，對不同的內文字型、字級，找到一個較適當的掃描解析度，讓需要使用中文光學文字辨識系統的使用者，可以在任何的印刷品或任何文字、文件辨識上，有一個共通的辨識方法及最有效率的工作流程，換句話說辨識的文件時，便可



從其文字字型得知需使用多少掃瞄解析度去加以辨識，且取得最高辨識率。

五、研究架構



圖一 研究架構

六、研究範圍與限制

字體：楷體、明體

字型：新細明體、標楷體

樣本字數：400 字（隨機抽取）

文字顏色：黑色

字間、行間：Microsoft Word XP 預設值

編排方向：橫排

印表機：HP LaserJet 8550 雷射印表機

輸出點陣圖解析度：1200DPI

掃描器：EPSON Perfection 1270 平台

掃描器

影像格式：8bit 灰階.bmp 點陣圖。

中文光學文字辨識軟體：蒙恬認識王專業版 3.0

七、名詞解釋

中文光學辨識系統 (OCR)：

利用光學儀器將印刷品文字或是手寫文字，經由光學儀器輸入成電腦可以

解讀的資料，並經由中文光學辨識系統，辨識成可供電腦編排的數位文字檔。（陳紀安，1982）

BIG 5：

台灣目前使用最為普遍的一種字集，由倚天中文系統成功的在市場上推銷，BIG 5 碼已經凌駕於今天的國家標準碼 (CNS-11643) (交通大學/數學系電算中心，2004)。

解析度：

解析度是指分辨圖像細部之能力，在單位長度內所能分辨的平行線數目來決定，人眼所能分辨的解析度於 30 公分的明視距離下，為 100 到 200 線數 (R.M Romano, 1998)。

八、研究假設

(一) 操作型定義：

辨識率：

蒙恬光學辨識軟體對標楷體、新細明體的 BIG5 常用繁體中文字的辨識率（內碼範圍為 A440~C67E）。辨識率越高，辨識能力越好。

字級：

單位為點，簡寫 pt，一英吋長度有 72pt。常用內文字級為 8、9、10、11、12。

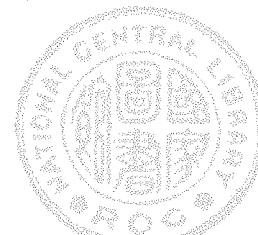
掃瞄解析度：

單位為 DPI。本實驗設定的掃描解析度為 200、300、400、500、600、700、800DPI。

(二) 研究變因

操作變因：字級與掃瞄解析度乘積值分組百分位排行¹

由於本實驗的 29 種字級與掃瞄解析度乘積值，皆各由少數字級及掃瞄解析度情形所構成，為求客觀，本實驗將字級與掃瞄解析度乘積值依下列範圍分為 8 組。



應變變因：辨識率百分位排行

固定變因：詳見研究限制與範圍

假設一

$$Y_s = \alpha + \beta_1 X_s + E$$

Y_s ：辨識率百分位排行

α, β_1 ：參數

X_s ：字級與掃瞄解析度乘積值百分位排行

E：隨機誤差

$$H_0: \beta_1 = 0 \quad i = 1$$

辨識率百分位排行與字級與掃瞄解析度乘積值百分位排行，兩者沒有顯著關係

$$H_a: \beta_1 \neq 0 \quad i = 1$$

辨識率百分位排行和字級與掃瞄解析度乘積值百分位排行，兩者之間有顯著關係

假設二

前提：

假設二的推論，是在假設一的成立 H_a 的情形下，才具有的。因為假設一跟假設二有因果關係。當假設一的 H_a 不成立（辨識率百分位排行與字級與掃瞄解析度乘積值百分位排行，兩者沒有顯著關係），假設二本身也沒有意義。

操作變因：是否縮放為一定的字級與掃瞄解析度乘積值

應變變因：全辨識率

固定變因：印刷紙張，雷射輸出機，掃描器，OCR 辨識軟體

假設二如下

$$Y_s = \alpha + \beta_2 X_t + E$$

Y_s ：全辨識率

α, β_2 ：參數

X_t ：是否縮放為固定的字級與掃瞄解析度乘積值（沒有縮放；縮放為一個固定字級

與掃瞄解析度乘積值）

E：隨機誤差

$$H_0: \beta_2 = 0 \quad i = 1$$

是否縮放為一個固定字級與掃瞄解析度乘積值，對於全辨識率沒有顯著關係

$$H_a: \beta_2 \neq 0 \quad i = 1$$

是否縮放為一個固定字級與掃瞄解析度乘積值，對於全辨識率有顯著關係

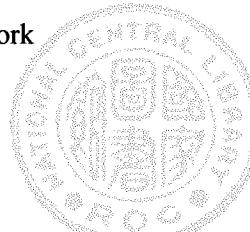
貳、文獻探討

一、何謂中文光學辨識系統

(一) 文字辨識系統的評估

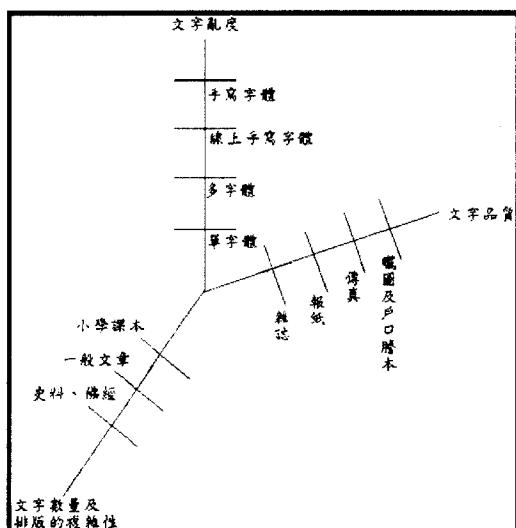
中文光學辨識系統的評估方式以單純的眼光來判斷，現在市面上的中文光學辨識系統軟體，普遍性來說已經可達準確率百分之 95，辨識速度一秒鐘 20 個字。但是這種分析忽略了一個大重點，就是真正處理文字的文件輸入的問題，並非單純的以文字輸入作為評估重點。因此必須以整體來說，以整體的處理效率來做評斷，凡舉文件的分析可分為幾類，自動排版版面分析、英文及中文的文字分離等，同一對文字不同的字型或是不同的字級等，因此單就文件分析的正確性、電腦輔助校稿的方便性因素，以上幾點都是評估的重要條件及指標，以文字辨識率評估的標準，來推導出符合消費者所需求的最高品質光學中文辨識系統、光學中文輸入方式及輸入的可靠文件技術，使得其辨識結果的錯誤頻率降低。

在陳宏興（1989）提出，可靠的辨識技術方面，可劃分為文書畫像的處理方式，前處理技術，特徵抽取技術即便是原理。在知識處理技術上可分為，文章數據的處理技術、短句數據的處理技術及高速文字的對照技術。在系統軟體技術方面，有 Host/Network



連接技術（連線技術）、讀取控制技術及修正、操作、運用技術（人為變因因素）。

在掃瞄技術上可分為，光學變換量子技術、追蹤控制技術及送紙原理技術。以上這四種分類都是可靠輸入法的關鍵技術。



圖二 文件輸入的研究範圍及技術層面

二、掃描機之理論探討

(一) 關於影像掃瞄輸入

市面上的掃描機，依照光學原理可分為反射式與透射式兩種，反射式主要是對於非透明原稿進行掃描，而透射式字主要對於投影片的掃描。一部掃描機能夠運作主要是靠感光元件感光產生類比訊號，再由類比訊號轉換成數位訊號。

在 G.G Field (1999), 中指出影響圖像品質與辨識關係，最大的因素在於其過網解析度與光學影像擷取解析度，因此本研究在選取影像輸入方式上，我們也本組也探討其掃描原理。

1. 影像擷取

印刷圖像經由 CCD 掃描器的掃描擷取，將影像傳到類比/數位 (A/D) 轉換器，把類比電壓轉換成不同的色彩或我們所謂的灰階值，儲放於電腦記憶體中，以便進行顯示

與處理分析。

掃描機工作原理。將光源反射至圖像，原稿上由鏡片把圖像原稿反射之光線折射到鏡頭由鏡頭把影像聚焦至 CCD 晶片上，以電壓方式紀錄；再傳送到 A/D 轉換器，把影像類比電壓轉換成數位資料 (I.Sybil, I.Emil, 1996)。

2. 影像資料格式

一般呈現於電腦上的圖像或照片，我們統稱為數位影像。依存放方式的不同又可分為空間域影像資料格式及頻率域影像資料格式。以下就影像空間域資料格式作一探討。空間域影像資料格式是數位影像在電腦裡最常使用的資料格式。

每張數位影像都是由許多像點所組合而成的，這些點又稱為像素 (Pixels)。通常每一個像素可以呈現出許多不同的顏色。一張數位影像的資料通常是以一個二維陣列來存放，每一個像素的資料，皆相對應於三維陣列中的某一個元素 (element)，元素數值的大小端視像素所呈現的顏色之不同而有所不同。

以黑白數位影像為例，每一個像素只有 2 種變化，最亮的顏色係白色，以數值 1 表示之；最暗的顏色係黑色，以數值 0 表示之；如果將一個存有數位影像資料約二維陣列之每一個元素值讀出，即可輕易地將數位影像顯示在電腦上，這些元素值我們又稱之為「像素值」；以上述方式所存放的數位影像，稱之為「空間域數位影像」，這種一個像素對應二維陣列中一個元素的數位影像資料存放格式稱為「空間域影像資料格式」。而用來存放空間域影像資料的二維陣列，又稱為「二維空間域陣列」。以 256x256 像素的灰階影像而言，每個像素具有 256 (=28) 個灰階，其中 256x256 為影像的大小，稱之為空間解析度；256 個灰階所代表的是明暗度，則每一個像素可以有 256 種變



化，最亮的顏色係白色，以數值 255 表示；最暗的顏色係黑色，以數值 0 表示；介於黑、白之間的顏色稱為灰色，而灰色依明暗的以數值 254 到 1 來表示之。

數位影像除了黑白及灰階兩種類型外，彩色影像也是常見的一種數位影像格式，電腦中是以三原色光（藍、紅、綠）來表現顏色，每一種原色光依其顏色由深至淺分別以數值 0 至數值 255 表示，共有 256 種變化。因此，這三原色光可以表示出 $256 \times 256 \times 256 = 16,777,216$ 種顏色變化。經由以上說明，我們可以瞭解到一張彩色數位影像中的每一個像素，可以有 2^{24} 次方種顏色變化，而每一像素都是由三原色光所組合而成。

三、結論：

以上各節探討了光學辨識系統的流程，其大多離不開以下幾點，光學辨識系統原理、中國字的演進、編碼及文字構成原理、中文字的辨識困難度、內文常用字型、常用字級、掃描器、雷射印表機等。

然而現今中文光學辨識系統普及化，文字的輸入已改由從前的打字輸入，改變成由掃描器辨識得到文字編碼，但是相對的在市面上最需要經由光學輸入的文本，是字數繁多的內文文字。

為了增加工作效率，選定一個合於語言的辨識軟體和了解光學辨識系統流程都是必要的。然而，辨識還是會有問題，因為由以上論文我們可以知道，在辨識流程上文字的字級、字型、掃描方式、列印方式及印刷品輸出方式，都會影響真正的辨識率，而本組將在以下實驗中，對於現今最普及的兩種內文文字作各種解析度（200、300、400、500、600、700、800）及各字級（8、9、10、11、12）加以探討，讓中文光學辨識系統使用者在於利用時，得以有精準的辨識效果。

參、研究方法

一、研究流程

1. 文獻蒐集與探討
2. 指導老師訪談
3. 決定研究方向
4. 文獻蒐集與探討
5. 文獻分析與整理
6. 擬定研究範圍、變項、限制、限定、假定研究命名

二、研究變項

以下為本研究實驗變項：

(一) 自變項

1. 字型：標楷體及新細明體
2. 字級：8, 9, 10, 11, 12 (單位：pt)
3. 掃瞄解析度：200, 300, 400, 500, 600, 700, 800 (單位：DPI)
4. 字級與掃瞄解析度乘積值：等於字級乘以掃描解析度

(二) 應變項

辨識率

(三) 固定變項

參照第一章研究限制與範圍

三、量測工具

(一) 自動比對程式

為避免人工比對所產生的失誤，本實驗小組使用 JavaScript 程式透過網頁介面進行自動化字元比對。

(二) SPSS

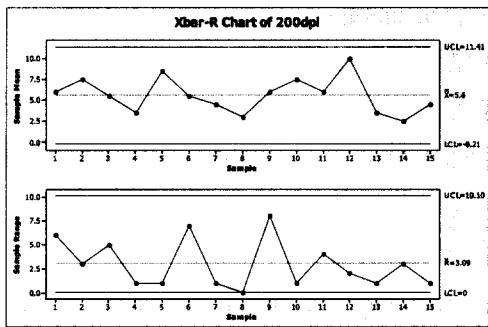
(三) Minitab

四、實驗過程穩定度測試

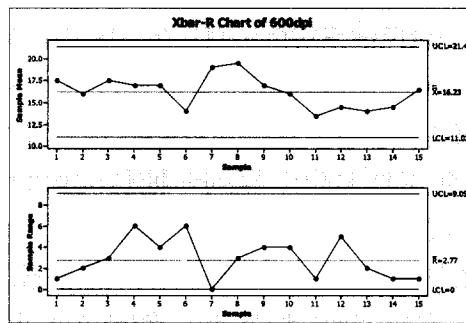
本研究在進行正式實驗前，先實施實驗流程穩定度檢測。檢測時抽取一張測試文字樣張，並以正式實驗流程進行測試，再將結果透過 MINITAB Release 14 的 Variable Chart for Subgroups 功能，以 Subgroup size (子組大小) 為 2 並勾選以下選項輸出圖表作穩定性監控：



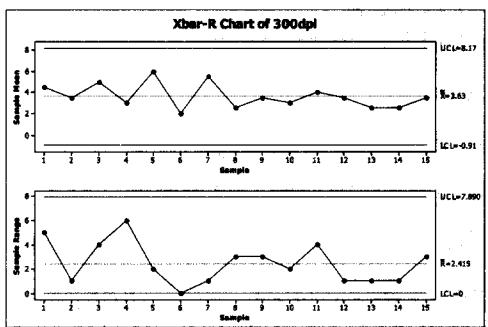
實驗流程穩定度檢測圖：500DPI



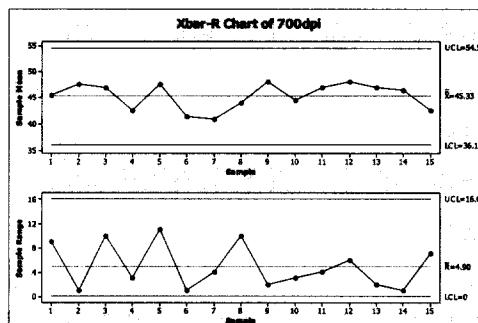
實驗流程穩定度檢測圖：200DPI



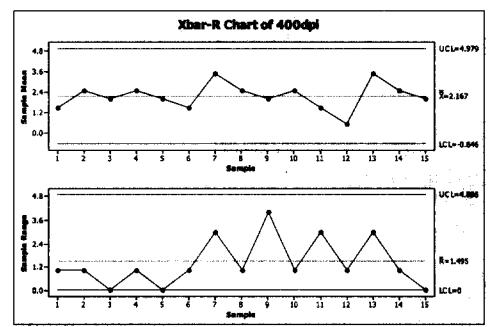
實驗流程穩定度檢測圖：600DPI



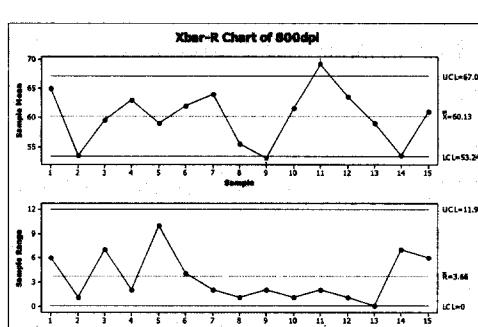
實驗流程穩定度檢測圖：300DPI



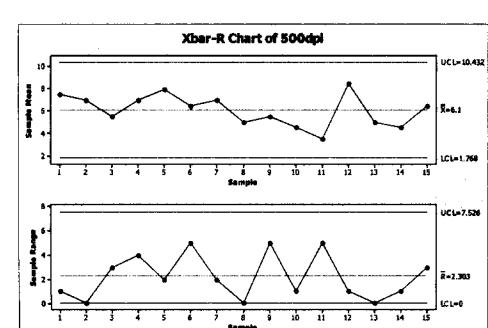
實驗流程穩定度檢測圖：700DPI



實驗流程穩定度檢測圖：400DPI



圖三 實驗流程穩定度檢測圖：100~800DPI



由 MINITAB 的 Variable Chart for Subgroups 可知，測試樣本在 200~700DPI 時，辨識率皆在三個標準差的穩定範圍內，在 800DPI 時，也僅兩個點的平均值超出三個標準差，顯示本研究操作過程是可信的。

五、實驗設計

本研究採單因子，2 階段式實驗設計。



母群體為一般書籍常用內文字

(一) 預備工作：

1. 樣本抽樣：

從 5 包(1500 張)雷射印表機專用紙中，使用 Systematic Random Sampling 方式，選取出 2 (字型) *5 (字級) *30=300 張紙。

從 BIG5-5401 字中，使用電腦程式隨機抽取 400 個字。

2. 樣本輸出

3. 樣本掃瞄

使用 Epson Perfection 1270 平台式掃描器掃瞄以上 300 份樣張。

4. 樣本辨識：

使用蒙恬認識王專業版 3.0，辨識以上 2100 個圖檔。

每個圖檔將產生 1 個辨識結果的文字檔。

將此文字檔另存新檔，以供下一比對步驟使用

5. 比對：

使用自動比對程式，比對出每個圖檔之錯誤字數。產生 2100 個 data

6. 辨識結果比對：

使用自動比對程式。

(二) 數據轉換：

把各字級、掃瞄解析度組合轉換成字級與掃瞄解析度乘積值。

例：8 字級；400DPI (8*400=3200)。

將有以下(標楷體、新細明體分開討論)

29 個種字級與掃瞄解析度乘積值

(三) 標楷體辨識率、新細明體辨識率

T-test

將標楷體跟新細明體這兩字型轉換成 29 組的字級與掃瞄解析度乘積值作 one-sample T-test。預期兩字型無顯著差異。

(四) 第一階段

1. 分組

如第一章所述，本實驗的 29 種字級與掃瞄解析度乘積值，皆各由少數字級及掃瞄解析度情形所構成，為求客觀，本實驗將字級與掃瞄解析度乘積值依下列範圍分為 8 組：

2. 檢定

把 8 組字級與掃瞄解析度乘積值做 one-way ANOVA，預期各組間都有顯著差異。

找出最佳全辨識率的一組字級與掃瞄解析度乘積值，並取其組字級與掃瞄解析度乘積值的中間值，作為最佳的字級與掃瞄解析度乘積值。例：2600~3599 為最佳組，則 $3100 = (2600 + (3599+1)) / 2$ 為最佳字級與掃瞄解析度乘積值

(五) 第二階段

1. 縮放

把 29 個字級與掃瞄解析度乘積值 (1600、1800、2000、2200、2400、2700、3000、3200、3300、3600、4000、4400、4500、4800、5000、5400、5500、5600、6000、6300、6400、6600、7000、7200、7700、8000、8400、8800、9600) 的原始圖檔，縮放成為最佳字級與掃瞄解析度乘積值的比例。

2. 再辨識

將縮放後的圖檔，用上述（實驗設計—預備工作）的辨識、比對方法，取得再辨識後的全辨識率。產生另外新的 2100 份 data。

3. 第二次分組

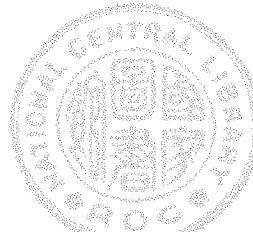
將縮放後的圖檔辨識後的全部辨識率(29 個字級與掃瞄解析度乘積值)，在利用相同的分組方式分成 8 組。

4. 檢定

將掃瞄完後直接辨識的 2100 份 data 與掃瞄完後再經過縮放後辨識的 2100 份 data，做 one-sample T test。希望它們之間有顯著差異。

5. 分析與結果

六、研究目的與研究假設



本研究主要探討內容為中文光學文字辨識系統，對一般書籍常用中文內文字『辨識率』與『字級與掃瞄解析度乘積值』的關係。透過研究方法，將樣本文字利用雷射印表機輸出成實驗樣張，並經過平台掃描器作掃描及中文光學辨識系統作辨識，最後再使用本研究撰寫之比對程式，對樣本文字及辨識結果文字進行核對，取得錯誤字數後，換算為辨識率，與『字型』、『字級』、『解析度』、『字級與掃瞄解析度乘積值』等變項進行分析與探討。兩階段研究假設如第一章所述。

七、一般描述性統計

(一) 各字型

接著再觀察兩字型的異同。標楷體及新細明體的最大值、平均值皆相同，且偏斜情形也都是明顯的負偏斜分佈，顯示兩字型的樣本辨識率大部分皆較接近 100%。但新細明體最小值為 82.5%，異動範圍較標楷體大 7.5%，且標準差也較高，整體來看，新細明體的辨識率較不穩定。

	8	9	10	11	12
200	96.43	97.81	97.76	98.82	98.60
300	99.08	98.87	99.13	99.16	99.09
400	99.29	99.69	99.48	99.17	99.46
500	99.17	99.87	99.87	98.43	98.48
600	99.57	99.47	98.68	97.22	95.94
700	99.13	99.03	97.96	94.96	88.67
800	99.06	98.29	96.68	91.29	84.97

新細明體

	8	9	10	11	12
200	94.42	97.39	97.26	98.07	99.00
300	97.73	99.12	98.77	99.30	99.42
400	98.18	98.50	98.97	99.19	98.80
500	99.07	98.97	98.21	98.44	96.88
600	98.35	98.52	98.01	98.45	96.24
700	98.14	98.04	98.00	97.82	94.17
800	98.20	97.89	96.93	97.48	95.21

標楷體

字型分組	1	2
字型	標楷體	新細明體
樣本數（份）	1050	1050
最大值（%）	100.00	100.00
最小值（%）	90.0	82.5
平均值（%）	97.9174	97.6721
標準差	1.51876	3.21507
偏斜度	-1.902	-2.647

各字型樣本辨識率描述性統計表

(二) 各字級

由各字級的最大值、平均值以及其大小不等的負偏斜情形可知，各字型辨識率整體上也是較接近最大值，但 10 號字偏斜度較



低，較接近常態分佈，字級提高或降低，分佈會較為傾向負偏斜。而離散程度最低的則是標準差為 0.89083 的 9 號字，字級增加或減少，離散度都有增加的趨勢。

字級 (pt)	最大值 (%)	最小值 (%)	平均值 (%)	標準差
8	100.00	92.00	98.2708	1.42713
9	100.00	94.25	98.6750	0.89083
10	100.00	95.50	98.2637	1.03285
11	100.00	89.75	97.6994	2.14683
12	100.00	82.50	96.0649	4.35775

各字級樣本辨識率描述性統計表

(三) 各解析度

至於各解析度的辨識率情形，相對於 300~600DPI，較低解析度或較高解析度的最大值皆未達 100%，且最小值、平均值較低，標準差也較大，顯示高解析度或低解析度的辨識率都會較 300~600DPI 低。至於偏斜情形，各解析度皆為明顯負偏斜，但 400DPI、500DPI 的偏斜程度更大。

解析度 (DPI)	最大值 (%)	最小值 (%)	平均值 (%)	標準差
200	99.75	92.00	97.5542	1.48127
300	100.00	96.00	98.9642	0.64152
400	100.00	94.75	99.0733	0.73553

500	100.00	93.00	98.7367	1.12504
600	100.00	92.50	98.0433	1.40613
700	99.75	86.75	96.5908	3.16992
800	99.75	82.50	95.6008	4.20936

各解析度樣本辨識率描述性統計表

(四) 各字級與掃瞄解析度乘積值值分組

由於本實驗字級與掃瞄解析度乘積值值變項為類別型變項，且本研究樣本的各相鄰字級與掃瞄解析度乘積值值有不等距情形，故改以每 1000 字級與掃瞄解析度乘積值為間隔，從字級與掃瞄解析度乘積值值的最小值起將字級與掃瞄解析度乘積值值分為 8 組等具類別型變項再進行探討。

	8pt	9pt	10pt	11pt	12pt
200DPI	1600	1800	2000	2200	2400
300DPI	2400	2700	3000	3300	3600
400DPI	3200	3600	4000	4400	4800
500DPI	4000	4500	5000	5500	6000
600DPI	4800	5400	6000	6600	7200
700DPI	5600	6300	7000	7700	8400
800DPI	6400	7200	8000	8800	9600

各字級、解析度組合的字級與掃瞄解析度乘積值表

由各字級與掃瞄解析度乘積值值分組的樣本來看，各組的平均值差異較從字型、字級、解析度探討時所觀察到的變化來得



大，以第 2 組與第 8 組為例，兩者平均辨識率相差約 6.74%（小數點下第二位四捨五入），由標準差所描述的離散度也差異得十分明顯。由此字級與掃瞄解析度乘積值分組變項比起字型、字級、解析度等三個變項對常用內文中文字辨識率的影響更高，是否有顯著關聯，尚待假設一的假設檢定。

分組	字級與掃瞄解析度乘積值	最大值 (%)	最小值 (%)	平均值 (%)	標準差
1	1600 2599	99.75 92.00	92.00 97.6951	97.6951 1.43181	
2	2500 3599	100.00 97.00	97.00 98.9760	98.9760 0.57055	
3	3600 4599	100.00 97.25	97.25 99.2146	99.2146 0.52210	
4	4600 5599	100.00 94.75	94.75 98.9108	98.9108 0.79490	
5	5600 6599	99.75 93.00	93.00 98.3633	98.3633 1.03879	
6	6600	99.25	92.50	97.4990	1.22677

	7599				
7	7600 8599	99.00	86.75	94.8708	3.19965
8	8600 9600	98.50	82.50	92.2375	4.91801

各字級與掃瞄解析度乘積值分組辨識率描述性統計表

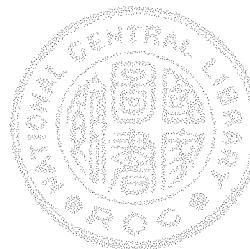
八、假設檢定

字級與掃瞄解析度乘積值分組與辨識率的相關性

這個階段的目的在於檢定假設一是否成立。本研究將 0.05 的顯著水準 (α 值) 針對字級與掃瞄解析度乘積值分組與辨識率透過 SPSS11.5 版進行變異量分析

(ANOVA, Analysis of variance)。由於進行 ANOVA 的應變項 (Dependent Variable) 必須為常態分佈，而本實驗樣本的辨識率有明顯負偏斜傾向，故改採 2100 份樣本的辨識率百分位排行，遇到若干相同值時，採其所占據排行範圍的平均值為其排行，如此一來可取得無偏斜的數據進行分析。此外 ANOVA 的因子 (Factor) 也必須是連續型 (Interval-ratio) 變項，因此本實驗先將字級與掃瞄解析度乘積值分組組別以該組所含樣本數作百分位排行，再進行變異量分析。

經過 SPSS11.5 的 One-Way ANOVA 後，我們可以從其檢測結果的 Sig. 值 (顯著值) 得知該因子對應變項的關聯性，由於 Sig. 值小於 α 值，顯示該因子與應變項有顯著關係，推翻 H_0 ，成立本假設，因此字級與掃瞄解析度乘積值分組與辨識率有顯著關



係，換言之，字級與掃瞄解析度乘積值會影響辨識率。

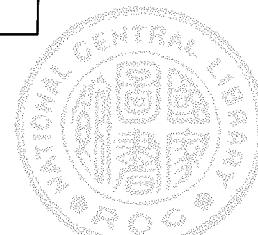
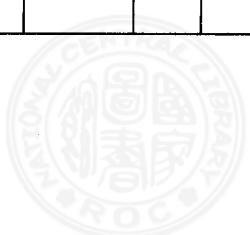
	Between Groups	Within Groups	Total
Sum of Squares	925574.452	815670.114	1741244.566
自由度	7	2092	2099
平均值 Square	132224.922	389.900	
F	339.126		
Sig.	.000		

假設一檢定變異量分析結果表（應變項：辨識率百分位排行（%）因子：字級與掃瞄解析度乘積值分組百分位排行）

另外，由字級與掃瞄解析度乘積值分組百分位排行的平均數折線圖可見，有最佳的字級與掃瞄解析度乘積值分組存在，因次本研究再對 One-Way ANOVA 結果的 Scheffe 多重分析表作進一步探討。Scheffe 多重分析圖表將各字級與掃瞄解析度乘積值分組百分位排行作交互比較，Sig. 小於顯著水準時表示因子的值為 (I) 與 (J) 時，兩者的應變項的值之間有顯著差異，因此可以觀察出當字級與掃瞄解析度乘積值分組百分位排行為 8.58% 和 22.87% 時，應變項的值會有顯著差異，當其為 22.87% 和 37.15% 時也發生同樣情形。以此類推，本實驗的字級與掃瞄解析度乘積值分組百分位排行的相鄰項目

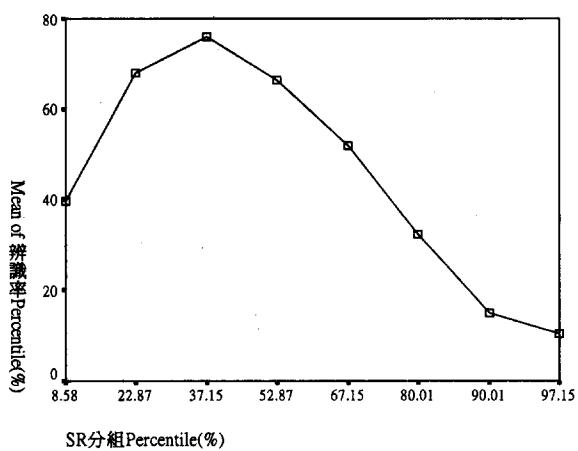
之間，除了第 7 組與第 8 組 (91.01% 及 97.15%) 外皆有顯著差異，並可得知第 3 組 (37.15%) 的應變數（辨識率百分位排行）顯著高於其它組，而 37.15% 為字級與掃瞄解析度乘積值分組第 3 組（詳見圖四十五），也就是 3600~4599 字級與掃瞄解析度乘積值（詳見表十一），此為辨識率最佳的字級與掃瞄解析度乘積值分組，字級與解析度乘積為 3600~4599 字級與掃瞄解析度乘積值之間的字級、解析度組合皆為最高辨識率的組合。

(I) 字級與掃瞄解析度乘 積值分組 百分位排行 (%)		22.87			
(J) 字級與掃瞄解析度乘 積值分組 百分位排行 (%)		37.15	22.87	52.87	37.15
平均差 (I-J)					
Std. Error					
Sig.					
95%	Lower Bound	-14.0397	1.64549	-7.8614 (*)	37.15
		1.6831	1.64549	7.8614 (*)	22.87
		3.7427	1.54361	9.5384 (*)	52.87
			1.54361	-9.5384 (*)	37.15
				-15.3342	.000



	Upper Bound	-1.6831	14.0397	15.3342	-3.7427
--	-------------	---------	---------	---------	---------

Scheffe 多重比較表(假設一的各字級與掃瞄解析度乘積值分組百分位排行的辨識率百分位排行(%))



圖四 各字級與掃瞄解析度乘積值值分組百分位排行的辨識率百分位排行折線圖

未縮放之辨識影像與預先縮放至最佳字級與掃瞄解析度乘積值之辨識影像時的辨識率差異

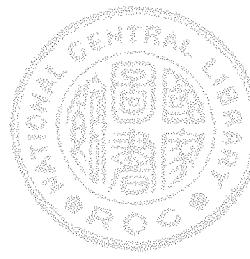
當假設一成立時，本實驗進行第二階段的假設檢驗。本階段的檢驗目的為測試假設二，預先縮放 2100 份全樣本影像至最佳字級與掃瞄解析度乘積值組合後，辨識率是否顯著高於未作縮放處理的 2100 份全樣本。本實驗同樣透過 SPSS11.5 以 0.05 的顯著水準的獨立樣本 T 考驗(Independent-Samples T Test) 對縮放為最佳組合後的 2100 份全樣本辨識率百分位排行及未作縮放處理的 2100 份全樣本辨識率百分位排行進行差異性檢測，由於探討的範圍與實驗一不同，故實驗二對所探討的 4200 份樣本的辨識率作重新排列，以取得未縮放與預先縮放的各 2100 份全樣本辨識率百分位排行再進行分

析。由測試結果可見，Sig. 值小於顯著水準 0.05，兩獨立樣本有顯著差異，且預先縮放處理的樣本群組辨識率平均值大於未預先縮放處理的樣本群組，推翻 H0，成立本假設。因此，預先縮放辨識影像至最佳字級與掃瞄解析度乘積值時的辨識率百分位排行顯著高於未預先縮放處理的影像，前者的辨識率優於後者。

	預先縮放的 全樣本辨識率 百分位排行 (%)	未縮放的 全樣本辨識率 百分位排行 (%)
樣本數	2100	2100
平均值	44.4198	55.6040
標準差	31.29582	24.73190
Std. Error	.68293	.53969
平均值		

假設二的獨立樣本 T 檢定結果二

	未/預先縮放的 全樣本辨識率 百分位排行 (%)	
	Equal variance	Equal variances



		s assumed	not assumed
Levene's Test for Equality of Variances	F	248.071	
	Sig.	.000	
	t	-12.849	-12.849
	自由度	4198	3985.098
	Sig. (2-tailed)	.000	.000
	平均差	-11.1841	-11.1841
	Std. Error	.87044	.87044
	Difference		
	95% Confidence Interval of the Difference	Lower	-12.8907
		Upper	-9.4776

假設二的獨立樣本 T 檢定結果二

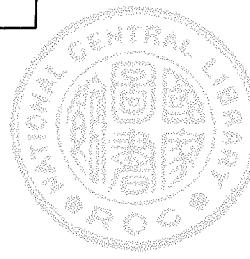
九、真實印刷樣本驗證

由於本研究的主要實驗所採用的樣本是以雷射印表機輸出來模擬印刷品的樣本，因此最後仍需再採用真實印刷品進行結果驗證。檢驗部分的實驗操作過程大致與主要實驗相同，採用的樣本為羅福林、李興才先生合著、中國文化大學出版的印刷工業概論（上）的第 223 頁（隨機抽取），該樣本為雪面銅版紙印刷品，內文字級大小為 11pt，字型為某廠牌明體字。由於真實印刷品文字尚包括英數及標點符號，非本研究所探討的字集範圍，並不列入辨識率計算，而此樣本中為 BIG5 常用 5401 中文字的字數有 602 字，此為本實驗計算辨識率時所依據的單份樣本文字數。另外，本測試採用的最佳字級與掃瞄解析度乘積值分組值為假設一所找出的最佳分組範圍的中間值 4100，也就是說，進行預先縮放處理的影像檔，皆透過以下換算式獲得縮放倍率再進行縮放：

$$\text{縮放倍率} = \frac{\text{最佳字級與掃瞄解析度乘積值}}{(\text{字級} \times \text{掃描解析度})}$$

至於驗證方式，則與假設一、假設二相同，採用 SPSS11.5 以 0.05 為顯著水準分別進行變異量分析與獨立樣本 T 考驗。從檢測結果可發現，兩項檢測的顯著值皆小於顯著水準，且預先縮放處理的樣本群組辨識率平均值高於未預先縮放處理的樣本群組，字級與掃瞄解析度乘積值分組百分位排行與辨識率百分位排行有顯著關係，預先縮放至最佳組合的影像整體辨識率也優於未縮放處理的影像。

	Between Groups	Within Groups	Total
Sum of	41116.600	306394.143	347510.743



Squares			
自由度	6	413	419
平均值	6852.767	741.874	
Square			
F	9.237		
Sig.	.000		

真實印刷樣本驗證的變異量分析結果表（應變項：辨識率百分位排行（%） 因子：字級與掃瞄解析度乘積值分組百分位排行）

	預先縮放的 全樣本辨識率 百分位排行 (%)	未縮放的 全樣本辨識率 百分位排行 (%)
樣本數	210	210
平均值	40.5159	59.7222
標準差	28.39159	25.90992
Std. Error	1.95921	1.78795
平均值		

真實印刷樣本驗證的獨立樣本 T 檢定結果一

未/預先縮放的 全樣本辨識率 百分位排行 (%)

Levene's Test for Equality of Variances	Equal variances assumed		Equal variances not assumed
	F	.248.071	
Sig.		.000	
t		-12.849	-12.849
自由度	4198	3985.098	
Sig. (2-tailed)	.000	.000	
平均差		-11.1841	-11.1841
Std. Error Difference	.87044	.87044	
95% Confidence Interval of the Difference	Lower	-12.8907	-12.8917
	Upper	-9.4776	-9.4776

真實印刷樣本驗證的獨立樣本 T 檢定結果二



		未/預先縮放的 全樣本辨識率 百分位排行 (%)	
		Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F	248.071	
	Sig.	.000	
	t	-12.849	-12.849
	自由度	4198	3985.098
	Sig. (2-tailed)	.000	.000
	平均差	-11.1841	-11.1841
	Std. Error Difference	.87044	.87044
	95% Confidence Interval of the Difference	Lower	-12.8907 -12.8917
		Upper	-9.4776 -9.4776

真實印刷樣本驗證的獨立樣本 T 檢定結果三

十、其他發現

預先縮放處理使整體辨識穩定度提高

當假設檢定過後，本研究再以預先縮放處理後的全樣本進行描述性統計，發現進行縮放處理後的樣本，辨識率的偏斜度與標準差明顯降低，顯示辨識率更趨於常態分佈，且辨識結果更加穩定。

由於蒙恬認識王專業版本身已具有偵測字體大小的能力，所以能對各種不同大小的文字影像進行辨識，但本研究透過預先縮放處理後，辨識穩定度仍可再提高，因此本研究推測，蒙恬認識王專業版對字體大小偵測能力尚有改進空間。

肆、結論與建議

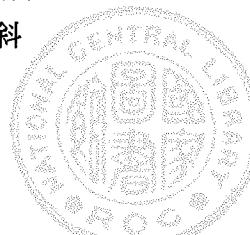
一、結果

在本組實驗中可以得知在於母群體樣本之辨識率上，大部分的辨識率都有到達平均值則在 97.7948，且偏斜度大於 -1，為明顯負偏斜分佈，由以上兩點可知，大部分樣本的辨識率都較接近最大值，接進最大直顯示其蒙恬認識王辨識軟體，對於本組之全體樣本有良好辨識率。

而在於本組實驗對於各字型、各字級、各掃描解析度上，可以發現。標楷體及新細明體的最大值、平均值皆相同，且偏斜情形也都是明顯的負偏斜分佈，代表兩字型的樣本辨識率大部分皆較接近 100%，即可敘述為中文光學辨識系統，對於這兩種字型有著不錯的辨識效果。但新細明體最小值為 82.5%，異動範圍較標楷體大 7.5%，且標準差也較高，整體來看，新細明體的辨識率較不穩定，換句話說，本實驗所探討的中文光學文字辨識系統對於辨識標楷體的能力及穩定度大於新細明體。

在各字級實驗中，各字級的最大值、平均值以及其大小不等的負偏斜情形可知，各字型辨識率整體上也是較接近最大值，但 10 號字偏斜度較低，較接近常態分佈，因此可以了解到字級提高或降低會影響辨識率，在字級方面本次的實驗以 10 號字為最高辨識程度之字級單位，再將字級提高或是降低，則分佈會較為負偏斜傾向。而離散程度最低的則是標準差為 0.89083 的 9 號字，字級增加或減少，離散度都有增加的趨勢。

在各解析度實驗中，各解析度的辨識率情形，相對低於 300DPI 或高於 600DPI，低於解析度 300DPI 或高解析度 600DPI 的最大值皆未達 100%，且最小值、平均值較低，標準差也較大，顯示解析度若高於 600DPI 或低於 300DPI 解析度的辨識率都會較 300~600DPI 低。至於偏斜情形，各解析度皆為明顯負偏斜，但 400DPI、500DPI 的偏斜



程度更大。

在字級與掃瞄解析度乘積值的實驗
中，本主將其字級與掃瞄解析度乘積值劃分成從 1600 至 9600 的單位量，加以分析，再將此分析歸類成 8 組，由各字級與掃瞄解析度乘積值分組的樣本來看，各組的平均值差異較從字型、字級、解析度探討時所觀察到的變化來得大，以第二組與第八組為例，兩者平均辨識率相差了約 6.74%（小數點下第二位四捨五入），由標準差所描述的離散度也差異得十分明顯。由此可猜測，字級與掃瞄解析度乘積值分組變項比起字型、字級、解析度等三個變項對常用內文中文字辨識率的關聯性更高，實際關聯性的強度，尚待後續統計分析。就本組實驗的數據來探討，在字級與掃瞄解析度乘積值 3600 至 4600 實有最佳的辨識效果，也可以敘述成在這個字級與掃瞄解析度乘積值標準中的辨識文字有最高的辨識率，在得知此發現後本組再將預先縮放全樣本影像至最佳字級與掃瞄解析度乘積值組合後，試驗辨識率是否顯著高於未作縮放處理的全樣本。本實驗同樣透過對縮放為最佳組合後的 2100 份全樣本辨識率百分位排行及未作縮放處理的全樣本辨識率百分位排行進行差異性檢測。由實驗結果得知，Sig. 值小於顯著水準 0.05，兩獨立樣本有顯著差異，事先縮放其影像檔至字級與掃瞄解析度乘積值辨識率最高的標準範圍中，可以增加其辨識率及減少辨識誤差，舉例來說，如果有一份文件，需要經過中文光學辨識系統處理，在得知此文件的文字字級時，可以從在掃描解析度上調整，使其成為高辨識率文件。

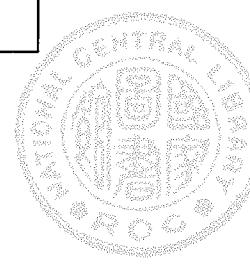
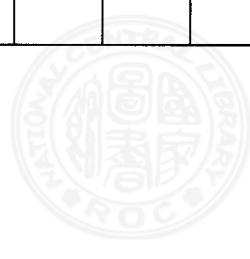
在經過三、四章的數據量畫統計及實驗後，本組的實驗數據總體來說，可以論述出在標楷體及新細明體，這兩種常用內文字型中，有其辨識的最佳操控位置，本主也把最佳位置整理分組成，可供光學辨識系統使用

者對照的數據，單單對於這兩種內文文字的實驗，本組也將各掃描解析度、各字級、兩種字型單獨討論出其最高辨識率的條件，希望使用者在選擇辨識條件時，有各種數據可以參考。

	預先縮放至最佳 字級與掃瞄解析 度乘積值	未縮放至最佳字 級與掃瞄解析度 乘積值
樣本數（份）	2100	2100
最大值（%）	100.00%	100.00%
最小值（%）	82.50%	94.00%
平均值（%）	97.7948%	2.51668
標準差	98.7625%	0.67646

是否預先縮放處理至最佳字級與掃瞄解析度乘積值的全樣本辨識率描述性統計表

	8pt	9pt	10pt	11pt	12pt
200DPI					
300DPI					3600
400DPI		3600	4000	4400	
500DPI	4000	4500			
600DPI					
700DPI					
800DPI					



最佳字級與掃描解析度乘積值對照表

	8pt	9pt	10pt	11pt	12pt
200DPI					
300DPI					99.09
400DPI		99.69	99.48	99.17	
500DPI	99.17	99.87			
600DPI					
700DPI					
800DPI					

新細明體最佳組合辨識率對照表

	8pt	9pt	10pt	11pt	12pt
200DPI					
300DPI					99.42
400DPI		98.50	98.97	99.19	
500DPI	99.07	98.97			
600DPI					
700DPI					
800DPI					

標楷體最佳組合辨識率對照表

二、研究建議與後續展望

本組實驗中所做的標楷體與新細明體兩字級，在辨識率比較下，標楷體辨識率

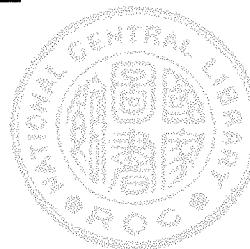
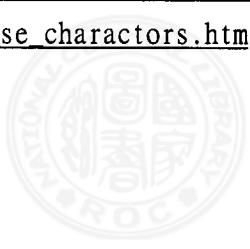
相較高於新細明體辨識率，影響新細明體辨識率的原因可能是新細明體字型的粗細結構，從本組研究發現新細明體在愈小字級與愈低的解析度下會造成新細明體的字型產生斷字及不清晰的情況發生，進而造成中文光學辨識系統在解讀上增加辨識錯誤率；另外，標楷體在辨識率較高也可能是它的字型結構趨於書法字型，並且相較於新細明體字型在做愈小字級與愈小解析度更容易讓中文光學辨識系統來辨識。

本研究對新細明體與標楷體的 8、9、10、11、12 字級進行深入探討，並藉由不同的掃描機解析度 (200DPI、300DPI、400DPI、500DPI、600DPI、700DPI、800DPI)，來嘗試找出對中文光學文字辨識系統最佳的辨識方法，透過一連串的辛苦研究，本研究找出一個相當簡易的換算公式：字縮放倍率 = 最佳字級與掃描解析度乘積值 / (字級 × 掃描解析度)，透過此公式，在使用中文光學文字辨識系統時，就無須考慮到影像畫素與掃描解析度 DPI 之間的繁雜換算

伍、參考文獻

一、中文部分

- 方御凡，(1997)，中文字型辨識、常用字型辨識、與手寫或印刷字體之辨識，國立中央大學/電機工程研究所，p f。
- 中國科普博覽，(2005)，線上資料：
http://www.kepu.com.cn/gb/technology/telecom/multimedia/mlt206_02.html
- 中國文字的演進，(2004)，線上資料：
<http://www.kepu.com.cn/gbcivilization/printingevolvev1132.html>
- 方塊字的結構，(2004)，線上資料：
http://tds.ic.polyu.edu.hk/vct1/comm_thinking_with_visual_imagesstructure_of_chinese_characters.htm



5. 行政院主計處電子處理資料中心，
(2004)，線上資料：
[#1](http://www.cns11643.gov.twwebword.jsp)
6. 吳偉賢，(2000)，筆劃特徵用於中文離線辨識，國立中央大學/資訊工程研究所，博士論文，p4，p 5-18。
7. 甚麼是內碼，(2004)，線上資料：
<http://www.cmi.hku.hkRefNotesChCmpchisys.html>
8. 陳子昂，(1994)，"智慧型中文輸----中文字辨識、語音辨識"，工研院電通所，新竹。
9. 陳紀安，(1982)，中文光學辨識系統應用於印刷上的發展與未來，華岡印刷傳播學報 25 期，p 47-48。
10. 陳宏興，(1989)，印刷文件和電子文件的轉換，華岡印刷傳播學報 24 期，p 73。
11. 陳俊龍，2000，彩色影像掃描系統之設計與實現，國立海洋大學/電機工程系，p11
12. 曾元顯，(2003)，中文 OCR 文件錯誤詞彙之自動偵測與更正及其在資訊檢索上的應用，台北。
13. 屠樂挺，(1992)，文字辨識之研究及系統製做，電腦與通訊月刊，p 28。
14. 黃澤祥，(1986)，"雷射印表機及影印機簡介"，光學工程第 22 期，p 4-9。
15. 黃士晉，(1999)，在文中 OCR 系統偵測並且調整斜體文字，國立台灣交通大學/資訊工程系，碩士論文，p2-20。
16. 數位內容新世紀，(2004)，線上資料：
http://www.digitalcontent.org.tw/site_map.html
17. 傑通華文，(2004)，線上資料：
<http://www.sinopen.com.cn/c-1.htm>
18. 蒙恬科技，(2004)，線上資料：
<http://www.penpower.com.tw/product/ocr.htm>

二、西文部分

19. Chang, Shou Shan (1977), Reliability of Reading Rate Measures As a Function of Warm-up Passage and of Time Limit, Doctoral Dissertation, Kansas State University.
20. G.G. Field, (1999), Color and Its Reproduction, GATF Press, Pittsburgh, U.S.A
21. R.M. Romano, (1998), The GATF Encyclopedia of Graphic Communications, GATF Press, Pittsburgh, U.S.A , p556-747.
22. Y.H. Chiou, (1996), Recognition of Chinese Business Cards, Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C.
- 86 年 9 月出版／師大書苑發行

