資工系實務專題研究計畫成果報告

Optical Character Recognition (OCR)

專題編號：098CSIE-S002
執行期限：97 年 1 學期至 98 年 1 學期
指導教授：杭學鳴
專題計劃參與人員：95590315 翁明陽

一、 中文摘要

本系統目的是辨識輸入之數字、英文大寫的圖片檔(JPG、BMP)，使得電腦也懂得圖片中的文字代表為何字。專題進行過程中，利用大學所習得的程式語言並且深入了解影像處裡，進行辨識系統比對。作為日後的中文字辨識等相關研究的基礎，已達到學以致用的目的。

關鍵詞：水平投影、行切割、字切割、歸一化、文字辨識

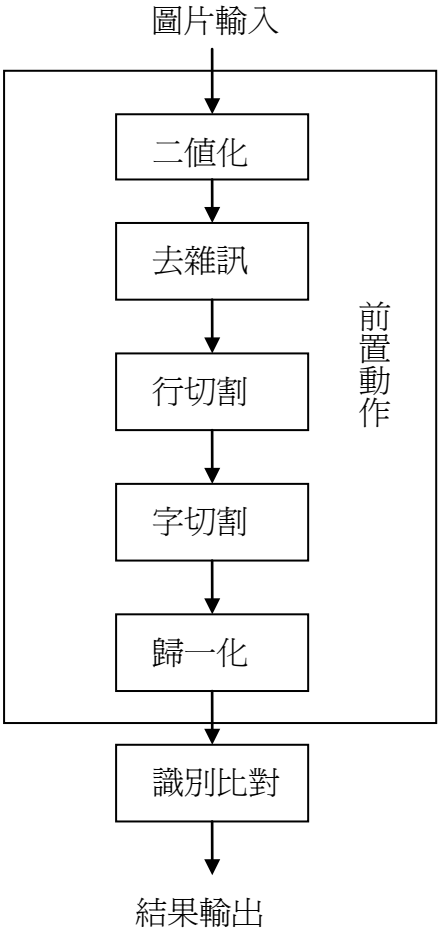
二、緣由與目的

市面上，OCR 技術的產品已經算是很完整，功能相當穩定了。對經常使用掃描機的使用者來說，OCR 可是一大利器。一大篇福掃瞄的文章，在短短的時間內就可以翻譯成文字檔，相當便利，取代鍵盤，節省輸入文字的時間。

本系統的主要工具是利用 Matlab 強大的影像處理能力及指令，節省開發影像處理程式碼的時間。

三、技術簡介

1. 執行流程圖：

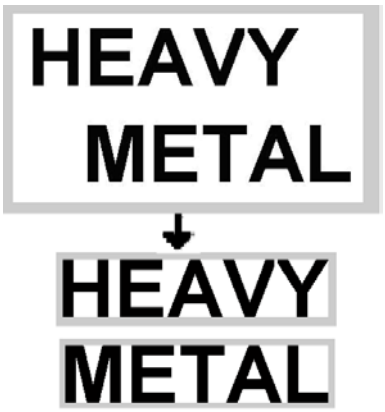


2. 流程圖說明：

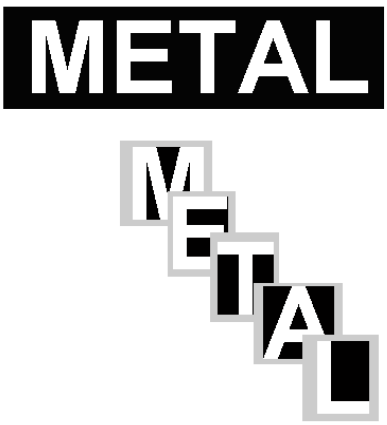
二值化：二值化又稱為灰度分劃(Threshold)， 一般影像的灰度分劃成只有兩種灰度值，亦即設定一個灰度值，凡是影像本身灰度大於它的便令其為亮點而灰度值低於設定值的，便令其為暗點，如此可得到一個二元的影像。

去雜訊：經由黑白反轉把像素面積低於 30 的清除，可以清除多數雜訊。

行切割：對二值圖像從上到下逐行掃描，同時計算每掃描行的像素，以獲取圖像的水平投影。利用文字行間空白間隔造成的水平投影空隙，即可以將各行文字分割開來。



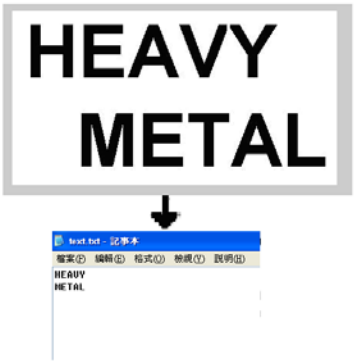
字切割：從行切割後得到的文字圖像行中將單字的圖像分割出來。字切割的正確與否直接影響識別結果，是單字識別系統中較重要也較困難的環節。



歸一化：此動作主要是把一單字位置、大小調整，伸縮待識字大小，使其與字庫中的標準字大小一致。字庫單字圖大小為 24×42。

識別比對：把待識單字與字庫中的每一標準字一一相減，而誤差值最低者，由電腦判斷為識別之結果。

結果輸出：把識別結果用 test.txt 檔輸出，並且儲存。



四、結論

如果使用其他字體，會發生辨識結果不是正確結果，例如輸入數字 1 但辨識結果可能會是 T 或者 I。經由去雜訊後，有些字，會因為彼此之相似，而使輸出結果錯誤。這一問題有待再進一步的研究，使用較先進的瘦化、骨架化等等技術。但可以目前測試結果，如用與字庫相同字體的待測字，其正確性可達 100%。雖然這是初階的專題研究，但是學習的經驗，希望有助於日後其他課題的研討。也希望有機會改進文字辨識的能力。

五、參考文獻

[1] 中國 OCR 信息網：**http://chinaocr.net/**

[2] **http://www.eqbyte.com/**

[3] **http://www.eqbyte.com/**

[4] Matlab 影像處理程式：**http://cslin.auto.fcu.edu.tw/fcu/matlabdsp/dspapp.htm**