# 行政院國家科學委員會專題研究計畫成果報告

## 病歷資料處理系統的設計-子計劃四：病歷表單的判定與文字切割系統(III)

## Determination of Clinical Documents and Character Segmentation (III)

計畫編號：NSC 88-2213-E-009-051
執行期限：87 年 8 月 1 日至 88 年 7 月 31 日

主持人：李錫堅　　國立交通大學資訊工程研究所

一、中文摘要

在已知表格文字辨識中，文字和表格重疊是一個嚴重的問題就是，這些受干擾的文文字不能正確的被抽取出來，而且對於傳統的光學文字辨識核心會導致辨識錯誤，因此我們提出兩種方法來處理有關於在已知表格中與線重疊之文字辨識。

第一種方法是移除表格線然後重建文字：當線移除後，文字就會破碎而且筆劃會分成兩群筆劃端（Stroke-ends），我們利用筆劃端的共線性（colinearity）和位置來找出正確的連接對應，同時，破碎筆劃之間的縫隙填補能盡量重建成原來的文字。第二種方法是修改 OCR 來辨識受干擾的文字：我們依據投影資訊來均勻分割含有表格線的印刷字體，我們找出表格線在受干擾字的位置並且計算表格線的兩種特徵值，：CDFs（contour-direction features）和 CCFs（crossing-count features），所以我們根據表格線特徵值修正 OCR 的特徵值來辨識這些受干擾的文字。

在第一個實驗中，我們先用 938 個受干擾的手寫字作測試，辨識率為 23.7%，經過使用第一種方法之後，辨識率提升到 78.3%。再第二個實驗中，我們測試了 695 個與線重疊的印刷字，辨識率為 64.3%，當使用第二種方法之後，辨識率增加到 77.3%。

關鍵詞：表單資料庫、格線去除、文字辨識

**Abstract**

It is a very important problem, characters be overlapped with lines in forms, to recognition character in known forms. The interfered-characters can't be extracted from the text lines exactly and the traditional OCR engines will fail to recognize characters with interference. The first method is to remove form lines and reconstruct characters. Characters are broken with line removal and strokes are separated into two sets of stroke-ends. The colinearity and position of the stroke-ends are used to find out correct connecting correspondence. Gaps of the broken strokes are filled to reconstruct the original characters. The second method is to modify the OCR model to fit interfered-characters. Printed characters with form lines are uniformly segmented according to projection profiles. The locations of form lines in the interfered-characters are extracted and both CDFs (contour-direction features) and CCFs (crossing-count features) of form lines are calculated. Trained features of the OCR engine are modified by the features of form lines to match interfered-characters.

In the first experiment, 938 handwritten characters with form lines are tested, and the recognition rate is 23.7%. After using the first method, the accuracy is raised to 78.3%. In the second experiment, 695 printed characters with form lines are tested, and the recognition rate is 64.3%. After using the second method, the accuracy is increased to 77.3%.

**Keywords**: form database, form line removal, character recognition

## 二、緣由與目的

病例資料無論對醫院或是對病人都是很重要的資料，但是面對龐大的資料，醫院如何管理，才能在需要的時候快速的找到一份病例資料，是一個很傷腦筋的問題，隨著科技的日新月異，如果能將病例完全輸入電腦，利用電腦快速的索引和網路的傳輸，將可在最短的時間之內調閱出所要的一份病例，但是如何將龐大的病例輸入電腦，卻又是另一個難題，因此我們希望能在本計劃中，利用影像處理的技術將病例影像轉換成電腦能夠方便處理的文字資料。

一般病例資料多半是以表單的形式存在，以一般文章的辨識方式來處理，將會受到表格線的影響，而無法辨識出正確的結果，由於在一間醫院中表單種類是有限的，我們可以先利用程式分辨表單的形式，然後再將有用的資料欄位擷取下來，加以辨識。

但是病例資料很多都是先有表格才寫上去或印製上去，常常會有文字和表格線重疊的情形，這種情形將嚴重影響辨識正確率，而單純去除表格線或是將表格線保留再加以辨識，效果都不是很好。我們嘗試從修改影像和修改辨識核心兩種方法著手，希望能使辨識率達到明顯的提昇。

## 三、結果與討論

我們測試的環境是 Pentium II 233 PC，128MB RAM 的機器上，作業平台是 Microsoft Windows NT 4.0，使用程式語言和模組是 Microsoft Visual C++ 5.0 的 MFC (Microsoft Foundation Class)。

我們拿了兩種測試資料來測試，一種是手寫字的表格資料，一種是印刷字的表格資料，我們將手寫字中有將線段除去再補字的資料和沒有去線的資料作比較，發現無去線的辨識率只有 23.7%，有去線補字的辨識率高達 78.3%，由這個結果可以得知，我們所採用去線補字的方法是有辦法明顯改善辨識率的。印刷字的測試資料方面，我們和手寫字做同樣的處理，另外加上測試調整過辨識核心後的辨識系統，發現沒有去線的辨識率為 64.3%，去線補字的辨識率為 79.6%，修改辨識核心的辨識率為 77.3%，由這個結果看來，修改辨識核心似乎沒有多大的作用，但是由於去線補字必須對影像作處理，而修改辨識核心，不需要動到影像，相對來講，修改辨識核心在速度上有明顯的改善，而辨識率只下降了一些而已，因此就某些用途來講，也算是一個值得考慮的用法。

## 四、計劃結果自評

在本年度的計劃中，我們針對字和表格線的重疊現象加以改善，這個現象之前的計劃雖然有考慮進去，但是由於並沒有詳細的去研究解決之道，形成辨識率的瓶頸，因此這份研究有助於使我們辨識系統更加完善。

本年度的計劃雖然沒有大幅度變動整個系統，不過卻可以讓辨識率大幅度的提昇，雖然離希望達到的目標還有一段距離，但是只要再加上後續的研發，整個系統的完成將指日可待。

## 五、參考文獻

[1]. R. G. Casey and D. R. Ferguson, "Intelligent Forms Processing," *IBM Systems Journal*, Vol. 29, no. 3, pp. 435-450, 1990.

[2]. Y. W. Shen, "Design of a Campus Document Processing System," *Master Thesis*, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1997.

[3]. K. C. Fan, J. M. Lu, and G. D. Chen, "A feature point clustering approach to the recognition of form documents," *Pattern Recognition*, Vol. 31, no. 9, pp. 1191-1406, 1998.

[4]. C. Z. lai, "Design of a Form Document Processing System," *Master Thesis*, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1998.

[5]. X. N. Chen and D. C. Tseng, "Form-structure Extraction for Table-form Recognition," *Proc. of 8th IPPR Conf. on Computer Vision, Graphics and Image Processing*, Taiwan, pp. 496-503, 1995.

[6]. S. S. You, P. C. Chang, N. J. Cheng and Y. J. Tsay, "Automatic Knowledge Acquisition for Chinese Archive Document," *CVGIP*1995.

[7]. R. G. Casey and D. R. Ferguson, K. Mohiuddin and E. Walach, "Intelligent Forms Processing," *Machine Vision and Applications*, vol. 5, pp. 143-155, 1993.

[8]. J. L. Chen and H. J. Lee, "An efficient algorithm for form structure extraction using strip projection," *Pattern Recognition*, Vol. 31, no. 9, pp. 1191-1406, 1998

[9]. C. T. Ho and L. H. Chen, "A High-speed Algorithm for Line Detection," *Pattern Recognition Letters*, Vol. 17, no. 6, pp. 467-473, 1996.

[10].W. Lee, C. F. Lin and Y. T. Juang, "A new line extraction algorothm for form documents," *CVGIP*1995.

[11].Y. H. Tseng and H. J. Lee, "Interfered-character Recognition by Removing Interfering-lines and Adjusting Feature Weights," *Proc. Int. Conf. on Pattern Recognition*, pp. 1865-1867, 1998.

[12].V. Govindaraju and S. N. Srihari, "Separating Handwritten Text from Interfering Strokes," *From Pixels to Features III: Frontiers in Handwriting Recognition*, Elsevier Science Publisher, pp. 17-28, 1992.

[13].S. Liang, M. Ahmadi, M. Shridhar, "Segmentation of Interference Marks Using Morphological Approach," *Proc. Third Int. Conf. Document Analysis and Recognition*, pp. 1042-1046, Montreal, Canada, 1995.

[14].Bin Yu and Anil K. Jain, "A Generic System for Form Dropout," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, no. 11, pp. 1127-1134, 1996.

[15].W. Huang, G. Rong and Z. Bian, "Strokes Removing from Static Handwriting," *Proc. Third Int. Conf. Document Analysis and Recognition*, pp. 861-864, Montreal, Canada, 1995.

[16].J. Y. Yoo et. al. "Line Removal and Restoration of Handwritten Characters on the Form Documents," *Proc. Fourth Int. Conf. Document Analysis and Recognition*, pp. 128-131, Ulm, Germany, 1997.

[17].M. D. Garris, "Method and evaluation of character stroke preservation on handprint recognition," NIST Internal Report 5687, July 1995.

[18]. V. K. Govindan and A. P. Shivaprasad, "Chinese Recognition - a Review," *Pattern Recognition*, vol. 23, no. 7, pp. 671-683, 1990.

[19]. F. H. Cheng and W. H. Hsu, "Research on Chinese OCR in Taiwan," *International Hournal of Pattern Recognition and Artificial Intelligence*, vol. 5, nos. 1&2, pp. 139-164, June 1991.

[20]. C. C. Kuo, "Evaluation and Speeding Up of Statistical Chinese Character Recognition," *Master Thesis*, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1996.

[21]. Y. H. Tseng, K. C. Kuo and H. J. Lee, "Speeding Up Chinese Character Recognition in an Automatic Document Reading System," *Proc. Fourth Int. Conf. Document Analysis and Recognition*, pp. 629-632, Ulm, Germany, 1997.

[22]. Y. O. Graham Leedham, "Segmentation and recognition of handwritten pitman shorthand outlines using an interactive heuristic search," *Pattern Recognition*, Vol. 26, no. 3, pp. 277-294, 1994.

[23]. C.Y. Suen, C. Nadal, R. Legault, T.A Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proceedings of the IEEE*, Vol. 80, no. 7, July 1992.

[24]. Richard G. Casey and Eric Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 7, pp. 690-706, 1996.

[25]. G. Seni and E. Cohan, "External word segmentation of off-line handwritten text lines," *Pattern Recognition*, Vol.27, no.1, pp.41-52, 1994.

[26]. J. Wang and J. Jean, "Segmentation of Merged Characters by Neural Networks and Shortest Path," *Pattern Recognition*, vol. 27, No. 5, pp. 649-658, 1994.

[27]. Y. Lu and M. Shridhar, "Character Segmentation in Handwritten Words - an Overview", *Pattern Recognition*, vol. 29, no. 1, pp. 77-96, 1996.

[28]. Y. Lu, "Machine Printed Character Segmentation - an Overview", *Pattern Recognition*, vol. 28, no. 1, pp. 67-80, 1996

[29]. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[30]. S. H. Lee, "Design of a Chinese Business Card Understanding System," *Master Thesis*, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1998

[31]. K. C. Kao, "Recognition of Chinese Characters with Overlapped Lines in Known Forms," *Master Thesis*, Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C., 1999