行政院國家科學委員會專題研究計畫成果報告

表格公文處理系統之設計
Design of a Form Document Processing System

計畫編號：NSC 88-2213-E-009-060
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

本計畫乃"公文處理"第二年計畫，主要目的乃是針對校園公文的處理，提出一套自動化的處理系統。在這套系統中，我們將使用已發展多年的文字識別模當作本系統的核心模。發展及研究公文自動化中所需其他的模及理論。

在本計畫中，為了加速處理時效及處理的正確性與實用性，本系統首先需判斷公文的種類。我們在學習階段中，分析入的表格以獲得實體及邏輯資訊來建構文件的樣本。在辨識階段中，系統從待辨識的表格文件中抽取出欄位，並用以辨識文件的類別，然後從該類別的樣本中獲取已知的正確資訊。此外，我們也提出一個合併了相連元件法(Connected-Component-Based Method)與投影法(Projection-Profile-Based Method)的文字切割方法(Character Segmentation method)以獲得更好的文字切割結果。為了提升文字的辨識率，我們從測試的文件中立了常用詞字典，並以之來更正部份文字辨識所產生的錯誤。

在我們的實驗中，文件類別的辨識率為 100%，文字的切割率為 96.3%，而文字的辨識率為 96.6%。

關鍵詞：校園公文、資訊處理、文字切割、文字識別、表格分析、表格辨識

Abstract

This project is the secondary year of the "Document Recognition and Processing System" We propose a system to process campus documents automatically. This campus-document processing system will use character recognition modules and other document-processing modules developed in our laboratory. These modules will be combined in this research.

In this project, to increase the efficiency, correctness and the usability, our system need to recognize the form type of the document first. In the learning phase, the input document is analyzed to obtain physical and logical information to construct the document template. In the recognition phase, the system uses the extracted fields of an unknown document to determine the document type and then obtains field knowledge from the document template. Besides, we also present a character segmentation method which combines the connected-component-based and the projection-profile-based methods to segment those characters in form documents more efficiently. To increase the recognition rate of our system, we correct some wrong character recognition results by using the

frequently-used word we collected from test documents.

In our experiments, the accurate rate of document type determination is 100%, the character segmentation rate is 96.3% and the character recognition rate is 96.6%.

Keywords: campus document, automation, character segmentation, character recognition, form document analysis, form recognition

二、緣由與目的

近年來，文件分析以及文字辨識的技術已達到實用性的階段，而且陸續有多項成果發表，我們企盼能整合這些技術使其能成為一個能夠實用性的系統。在校園中，每天收發各單位的公文不計其數，需要眾多人力處理，造成人力資源的浪費。有鑑於此，我們希望能發展一校園公文的自動處理系統，利用已發展成熟的技術加以改進，並針對此項需求，發展各項新的技術，為日後其他可能的應用，奠定銀好的基礎。

本計畫的最終目的在擴充中文文字辨識系統的功能於實用的文件分析系統中，以促進校園公文自動化，使堆積如山的學校公文能夠電腦化。在校園公文的處理上，我們使用光學掃描器，或數位照相機，將公文影像擷取下來。首先，將影像作適當的前處理:如二值化、去雜點、去除文件中的印章以及簽名。接下來抽取公文中的表格線，利用這些抽出的表格線做文件類別的區分。如果我們在抽取公文格線時，有部份格線未標定出來，則我們如何還是能夠比對出公文的類別，是我們研究的一個重點項目。區分類別之後，再從該類文件的實體與邏輯描述，確定所有格線及欄位的實際位置與相互關係。例如:欄位的位置; 欄位的名稱，然後再從影像中將文字切割出來，並加以辨識。然後我們將所辨識後的資料，利用語言模組，來作適當的矯正。最後，將資料存入資料庫中，以利以後的使用者使用。

三、結果與討論

在我們的實驗中，利用欄位位置的資訊來做文件類別的判斷，其辨識率可達100%，即使文件有些傾斜，甚至有些欄位沒有抽取出來，還是可以正確的辨識出來。在表格中文字切割部分，提出一個合併相連元件法(Connected-Component-Based Method)與投影法(Projection-Profile-Based Method)的文字切割方法，使得表格中文字切割的成功率獲得改進。此外，計算元件(component)的大小及距離，將相近的元件合併(merge)起來，也使切割率提升。在文字辨識部分，先利用英文的辨識核心，將英數字去除，並利用字詞更正的方法，將辨識錯誤的字更正，以提升文字的辨識率。

四、計劃結果自評

確實改進文字辨識核心及表格的辨識核心，且在表格中文字的切割部分，提出的新方法確實改進切割的成功率，在提高字詞的辨識率、判斷公文之結構、欄位資料的配合...等經過適當的修改，成為適合處理校園公文之所需。因為使用 object oriented 的方式來開發，我們所完成的程式能夠很容易地為後續的計劃、研究使用。對於其他子計劃所需的表單辨識部分有很大的助益。

五、參考文獻

[1]Y.W Shen, Design of a *Campus Document Processing System*, Master thesis, Institute of Computer Science and Information Engineering, National Chiao Tung University, Taiwan, R.O.C, 1996.

[2]J.L Chen and H.J Lee, "An efficient algorithm for form structure extraction using strip projection," to appear in *Pattern Recognition*, 1998.

[3]Wayne Niblack, "*An Introduction to Digital Image Processing*,"pp.115-116, Prentice Hall, 1986.

[4]Friedrich M. Whal, Kwan Y. Wong, and Richard G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents, " COMPUTER GRAPHICS AND IMAGE PROCESSING, Vol.20,pp.375-390,1982.

[5]O.D Trier and T. Taxt, "Evaluation of binarization methods for document images, "IEEE *Trans. On Pattern Analysis and Machine Intelligence,* Vol.17,no.3, pp.312-315, 1995.

[6]R.G Casey, D.R Ferguson, K. Mohiuddin and E. Walach, "Intelligent forms processing system, " *Machine Vision and Applications,* Vol.5, pp.143-155,1992.

[7]T. Watanabe, Q.Luo and N. Sugie,"Layout Recognition of Multi-Kinds of Table-form Documents, " IEEE *Trans. Pattern Analysis and Machine Intelligence*, Vol.17, no.4, pp.432-445, 1995.

[8]S.L Taylor, R. Fritzson and J.A.Paster, "Extraction of data from preprinted forms, "*Machine Vision and Applications*, no.5, pp.211-222, 1992.

[9]H. Fujisawa, T. Nakano and K. Kurino, "Segmentation method for character recognition: from segmentation to document structure analysis, "*Proc. Of the IEEE*, Vol.80, no.7, pp.1079-1091, 1992.

[10]Y.Y Tang, C. De Yan and C.Y. Suen, :Document processing for automatic kowledge cquisition, "IEEE *trans. On knowledge and data engineering*, Vol.6, no.1.pp.3-21,1994.

[11]K.C Fan, J.M Lu, L.S Wang and H.Y Liao, "Extraction of characters from documents by feature point clustering, "*Pattern Recognition letters*, Vol.16, pp.963-970, 1995.

[12]S.W Kam, L. Javanbakht, and S.N Srihari, "Anatomy of a form reader, "*Proc. 2nd Intern. Conf. On Document Analysis and Recognition*, pp. 506-509, 1993.

[13]X.N Chen and D.C Tseng,"Form-structure extraction for table-form recognition, "*Porc. Of 8th IPPR Conf. On Computer Vision, Graphics and Image processing*, Taiwan, pp. 496-503, 1995.

[14]C.T Ho and L.H Chen, "A high-speed algorithm for line detection, "*Pattern Recognition Letters*, Vol. 17, pp. 467-473, 1996.

[15]J. Wang and J. Jean, "Segmentation of merged characters by neural network and shortest path, "*Pattern Recognition,* Vol. 27, no. 6, pp. 825-840, 1994.

[16]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.
[17]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.
[18]R. Safari, N. Narasimhanurthi and M. Ahmadi, " Document registration using projective geometry, " *ICDAR*, 1995.
[19]S.S You, P. C Chang, N. J Cheng and Y. J Tsay, "Automatic Knowledge Acquisition for Chinese Archive Document, " *CVGIP* 1995.
[20]Y. I Lu, " Machine printed character segmentation－An overview, " *Pattern Recognition*, Vol. 28, no. 1, pp. 67-80, 1995.
[21]G. Seni and E. Cohan, " External word segmentation of off-line handwritten text lines, " *Pattern recognition,* Vol. 27, no. 1, pp. 41-52, 1994.
[22]S. Liang, M. Shridhar and M. Ahmadi, "Segmentation of touching characters in printed document recognition, " *Pattern Recognition*, Vol. 27, no. 6, pp. 825-840, 1994.
[23]W. Lee, C. F Lin and Y. T Juang, " A new line extraction algorithm for form documents, "*CVGIP* 1995.
[24]Shuichi Tsujimoto and Haruo Asada, " Major components of a complete text reading system, " *Proceeding of the IEEE*, Vol. 80, no.7. pp.1133-1149.