病歷表上英數字資訊處理(II)

Processing of Alphanumerical Information on Clinic Data(II)

計畫編號:NSC-87-2218-E009-033

執行期限:民國86國8國1日至民國87國7國31日

主持人:蔡文祥 交通大學資訊科學系教授

一、中文摘要

病歷表是文件的一種，但是對於病歷表格方面的分析和研究在以往並不常見，主要原因是因爲在於病歷表上的資訊非常的龐雜，中英文數字、圖形、符號都有，文字還有分成印刷字和手寫字兩種，圖形亦有先印好的及手繪的繪分，符號更是五花八門，打勾、蓋章、下線(underlines)等皆是。而文字輸入部分，亦有離線及線上兩種情形。最難的是文字辨認部份，主要困難在於醫師填寫的病歷內容常是非常潦草的筆跡，而字母幾乎無從認起。在第二年的計畫中，我們進一步對離線輸入的英文字及線上輸入的文字步辨認，以方便病歷表資訊之整體自動分析。整體而言，各種病歷內容之自動分析皆很不容易，整合亦是一大挑戰。本計畫在今年的研究成果主要包含下列三大類：

1、表格資訊的強化與重現

2、表格資訊的壓縮與重建

3、表格的建整

4、離線英文字辨認(一)

5、線上手寫數字及特殊符號的辨認

以上五個研究成果包含在此一報告之中。

英文摘要

Clinic data forms are a kind of document. But there is little research for clinic data forms. The major reason is the information on the clinic data forms is very complicated. There are Chinese, English, numerals, graphics, and symbols. In the category of word, there are still printed and handwritten words. There also exists difference between printed and painted one in graphics. Symbols are more variant, which include ticks, seals, underlines, and so on. There are two conditions ,on-line and off-line, in the part of character input. The most difficult is that the contents of clinic data forms written by doctors are very cursive writings, and there is no way to recognize the characters. In the second year of the project, we finished the off-line English handwriting recognition and On-line handwriting recognition to make the automatic clinic data forms analyzing easy. As a whole, automatic analysis of each clinic data is not very easy, and to integrate it is also a great challenge. The result of this year include the following five items:

1.Enhancement and display of form

information
2.Compression and reconstruction of form information
3.Form correction
4.off-line English handwriting recognition (Part 1)
5.On-line numeral and special-symbol handwriting recognition

二、計畫緣由與目的

本計畫擬集合「文件分析與文字識別」及「資料庫系統」領域內的一些教授來共同完成一個實用的病歷資料處理系統，病歷可以說是醫院裡面最重要的的資訊之一，這些資料具有量多、種類多的特性。本計畫擬針對舊有的病歷資料和新輸入的資料建立一個完整的處理系統，所謂舊的病歷資料指的是目前一般醫院中印刷和手寫的病歷，這些資料必須經過數位化處理，新的資料指的是電腦化後的資料，包括藥品種類和劑量等等。新舊資料必須整合成一個資料庫系統，能夠辨識的便加以辨識，不能辨識的存成數位檔，影像加以壓縮，圖形則加以向量化。再利用查詢語言，醫生或獲得授權的人員便可以抽取想要的某一筆資料。由於我們必須處理的文件型態很多，所以必須整合多個教授共同爲同一目標努力。

在第二年的本計畫中，我們力對表格資訊因掃瞄或切割而不清楚的結果及表格資訊過於龐大的情形再步處理，這必須進行三種工步，也就是表格資訊的強化與重現，表格資訊的壓縮與重建和表格的建整。此就，我們更進一步的的對離線輸入的英文字及線上輸入的文字就辨認，以方便病歷表的資訊之整體自動分析。

三、研究方就與成果

1.Enhancement and display of form information

In most filled-in data form images, the filled-in components might touch or break form frames. Sometimes the high density text components are very similar to lines. This is a big problem for form recognition. By using the proposed modified Hough transform and the line detection algorithm to detect significant lines, these high density text components can be ignored. Also, the transform takes less time. With significant lines, the proposed form recognition algorithm can recognize all kinds of data forms even though a significant line is broken into several significant lines. Image skewing may influence the result of the line detection. So, other methods for adjusting image skewing are needed. In our experiment results, if there is a skew angle in an input image, a false result of form recognition will occur due to the false detection of significant lines. When processing blank form images, fields to be filled in are extracted and relationships among fields are detected.

2.Compression and reconstruction of form information

A modified Hough transform method is proposed, which can detect significant lines from noisy forms quickly. If the result of recognition says that there exists no blank data form corresponding to an input filled-in form in the form database, the filled-in form is converted into a blank form. The resulting blank data form is analyzed and understood next. Otherwise, a form registration method is performed in order to find out all the handwritten characters in filled-in fields. An interface also is provided to type in the handwritten characters. Finally, the filled-in form can be reconstructed and stored in database. At this moment, digitization, layout enhancement, and compression of filled-in forms are achieved.

3.Form correction

When processing blank data forms, frames, printed characters, and fields to fill in are extracted and the relationships among fields are detected. The results are saved in a form database. A form filling module is also developed for filling form by computers conveniently. For filled-in forms, the first thing is form recognition. In this study, form recognition is based on the significant vertical and horizontal lines, instead of the layout structures of forms. Moreover, the width and height of forms are also used for form recognition. A modified Hough transform method is proposed, which can detect significant lines from noisy forms quickly. If the result of recognition says that there exists no blank data form corresponding to an input filled-in form in the form database, the filled-in form is converted into a blank form. The resulting blank data form is analyzed and understood next. Otherwise, a form registration method is performed in order to find out all the handwritten characters in filled-in fields. An interface also is provided to type in the handwritten characters. Finally, the filled-in form can be reconstructed and stored in database. At this moment, digitization, layout enhancement, and compression of filled-in forms are achieved. Experimental results show the feasibility and practicability of the proposed approaches.

4.Off-line English handwriting recogni-tion (Part 1)

An integration scheme of three matching methods for recognition of small sets of handwritten English words is proposed. The first method is matching by the distance-weighted correlation, which is a measure of the similarity of the relative spatial positions of the stroke skeletons of two words. The second method is matching by the 1-D Fourier descriptors, which represent the vertical projection histogram of the word. The last method is matching by mesh features, which represent the local spatial information of the word. After applying these methods, three matching measures are obtained and we combine them together to get a final measure, and an handwritten English word can be recognized according to the smallest final similarity measure between the input word and all model words. Experimental results are shown to prove the feasibility of the proposed approach.

5.On-line numeral and special-symbol handwriting recognition

A stroke-based scheme for recognition of on-line numeral and special-symbol handwriting is proposed. The directions of strokes are used as the major features of the numerals and special-symbols. The directions of strokes are first encoded. Then the codes are matched with the ones of all the models. The problem is transformed into a string matching problem, and dynamic programming is used to solve the problem with high complexity. After the matching process, some ambiguity still exists. Some post-process is applied and the automatic learning is used to dynamically adjust the personal models. Experimental results are shown that the system is feasible.

四、結論與討論

A form document processing system has been successfully implemented. It contains four major parts, including preprocessing, processing of blank form images, processing of filled-in form images, and form filling. Several major achievements in different phases are summarized as follows.

1. In the phase of preprocessing, minimum outer frame detection, significant line detection, and form recognition have been proposed. These algorithms are used to recognize an input form image with noise and are the kernel of the system.

2. In the phase of processing of blank form images, printed characters, and fields to fill in are extracted. Also, an interface is provided in order to type in printed characters, reconstruct and compress blank form images. This phase is the basis of the following phases.

3. In the phase of processing of filled-in form images, a form registration approach has been proposed. Also, an interface is provided in order to type in handwritten characters, reconstruct and compress filled-in form images. Besides, recovery of blank data forms from filled-in data forms is achieved.

4. In the phase of form filling, a form filling module is developed. Users can fill forms by computers and print or save the result of form filling.

The experimental results have revealed the feasibility of the above proposed algorithms.

A recognition system for small sets of handwritten English words has been successfully implemented. It is based on the use of some image processing operations, three individual matching methods, and an integration of these three matching methods. They are summarized as follows.

In the phase of image processing, an approach to automatic threshold selection using the moment-preserving principle is used. Then, the operations, including word slant correction, horizontal position adjustment, normalization, and thinning, are used to increase the similarity of the relative spatial positions of the words.

In the phase of recognition, three matching methods have been proposed. The first is the matching by the DWC. By using the DWC, a tested word is matched with a model word by measuring the similarity of the relative spatial positions of them instead of counting the overlapping pixels.

The second method is the matching by 1-D Fourier descriptors. A word image is transformed to its vertical projection histogram and represented by 1-D Fourier descriptors. Then two words are matched by computing the distance of their 1-D Fourier descriptors.

The last method is the matching by mesh features. A word image is divided into several uniform grids and the proportion of the number of black pixels in each grid with respect to the number of all black pixels of the whole word image is computed as the grid value. Then two words are matched by computing the distance of the grid values.

In the phase of integration, these three matching methods above are integrated by combining three measures of them to form a final measure for decision making.

To sum up, our experimental results prove that these three matching methods have been successfully integrated and the proposed system is feasible and effective.

The on-line input device for handwritten numerals and special symbols has been successfully implemented. Directions of strokes are the major features. They are simple. So the system is fast and the establishment of the model is easy. Farther, this property makes the automatic learning feasible. At the same time, this system still has high recognition rate, so it is also useful for the users. In the matching phase, elastic matching is used so writing difference is tolerant. Our experimental results prove this.

、參考文獻

[1] R. M. Bozinovic and S. N. Srihari, “Off-line cursive script word recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 1, pp. 68-83, Jan. 1989.

[2] J, H, Chiang and P. D. Gader, "Hybrid fuzzy-neural systems in handwritten word recognition," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 4, pp. 497-510, Nov. 1997.

[3] J. Cai and Z. Q. Liu, "Off-line unconstrained handwritten word recognition," *1996 Australian New Zealand Conference on Intelligent Information System. Proceedings. ANZIIS 96*, pp. 199-202, 1996.

[4] H. Bunke, M. Roth, and E. G. Schukat-Talamazzini, “Off-line cursive handwriting recognition using hidden Markov models,” *Pattern Recognition*, vol. 28, issue 9, pp. 1399, Sep. 1995.

[5] R. Buse, Z. Q. Liu, and T. Caelli, "A structural and relational

approach to handwritten word recognition," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 27, no 5, pp. 847-861, Oct. 1997.

[6] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690-706, Jul. 1996.

[7] S. Madhvanath and V. Govindaraju, "Contour-based image preprocessing for holistic handwritten word recognition," *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, pp. 536-539, vol. 2, 1997.

[8] T. J. Fan and W. H. Tsai, "Automatic Chinese seal identification," *Computer Vision, Graphics, and Image Processing*, vol. 25, pp. 311-330, 1984.

[9] B. Verma and M. Blumenstein, "An Intelligent neural system for a robot to recognise printed and handwritten postal addresses," *Proceedings of the Fourth IASTED International Conference Robotics and Manufacturing*, pp. 180-182, 1996.

[10] K. Hen and I. K. Sethi, "A off-line cursive handwritten word recognition system and its application to legal amount interpretation," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no, 5, pp. 757-770, Aug. 1997.

[11] W. H. Tsai, "Moment-preserving thresholding: a new approach," *Computer Vision, Graphics, and Image Processing*, vol. 29, pp. 377-393, 1985.

[12] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Company, U.S.A., 1993.

[13] [13]R. G. Casey, D.R. Ferguson. K. M. Mohiuddin and E. Walach, "Intelligent Forms Processing System," *Machine Vision and Application*, Vol. 5, pp. 143-155, 1992.

[14] K. C. Fan and M. L. Chang, "A Line Structure Based Approach to the Recognition of Form Documents," *proceeding of 7th Optical Character Recognition and Document Analysis Workshop*, pp. 4-9-4-16, 1997.

[15] K. C. Fan, J. M. Lu, L. S. Wang and H. Y. Liao, "Extraction of Characters from Form Documents by Feature point clustering," *Pattern Recognition Letters*, Vol. 16, pp. 936-970, September 1995.

[16] Y. W. Shen and H. J. Lee, "Design of a Campus Document Processing System," *Department of Computer Science and Information Engineering, National Chiao Tung University*, 1997.

[17] S. L. Taylor, R. Fritzson and J. A. Pastor, "Extraction of data from preprinted forms," *Machine Vision and Application*, Vol. 5, pp. 211-222, 1992.

[18] H. T. Fujisawa, Y. Nakano and K. Kurino, "Segmentation Methods for Character Recognition: from segmentation to document structure analysis," *Proc. of the IEEE*, vol 80, no 7, pp. 1079-1092, July 1992.

[19] Y. C. Tseng and W. H. Tsai, "Form Segmentation and Component

Classification for Clinic Data Image Analysis," *proceeding of 7th Optical Character Recognition and Document Analysis Workshop*, pp. 2-1-2-19, 1997.

[20] J. L. Chang and W. H. Tsai, "Automatic Recognition and Understanding of Image of Blank Data Forms," *Department of Computer and Information Science, National Chiao Tung University,* 1997.

[21] 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.

[22] R. Safari, N. Narasimhanurthi, M. Shridhar and M. Ahmadi, "Document Registration using Projective Geometry," *IEEE Trans. Image Process*, Vol. 6, no. 9, pp. 1337-1341, September 1997.