ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN THANH PHÚC

TÁCH VÀ NHẬN DẠNG SỐ VIẾT TAY TRONG PHIẾU NHẬP DỮ LIỆU

Ngành: Công nghệ Thông tin
Chuyên ngành: Công nghệ Phần mềm
Mã số: 60 48 10

LUẬN VĂN THẠC SĨ

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. Ngô Quốc Tạo

Hà Nội – 2008

Tóm tắt

Nhập dữ liệu tự động đang là bài toán ngày càng thu hút nhiều sự chú ý và đầu tư nghiên cứu bởi vì đây thật sự là một vấn đề quan trọng, cần thiết do khả năng áp dụng rộng rãi vào thực tế cũng như hiệu quả mà nó mang lại. Trong hệ thống này, khử nghiêng và phân vùng ảnh là một phần có vai trò đặc biệt quan trọng. Chức năng của nó là chính xác ảnh và tách ra các vùng được nhập thông tin để làm đầu vào cho module nhận dạng chữ. Dựa trên đặc điểm phân bố có hướng và đồng đều của form văn bản, chúng tôi đã sử dụng phương pháp phép chiếu để khử nghiêng ảnh do phương pháp này đạt được độ chính xác cao đối với những ảnh có đặc trưng trên. Cũng dựa trên đặc điểm của kiểu form văn bản là dữ liệu được nhập vào các ô trên form(nghĩa là nằm trong giới hạn giữa các đường thẳng), giải pháp đề ra cho phân vùng là thông qua việc xác định các đường thẳng kết hợp với sử dụng hệ tọa độ tương đối để xác định các vùng nhập dữ liệu. Chúng tôi đã tiến hành thực nghiệm trên nhiều kiểu form văn bản khác nhau và thu được những kết quả rất khả quan.

Từ khóa : detect skew angle, project profile method, form recognition

Abstract

Automatic form data reading is an attractive subject to many researchers because of its importance and widely applicability. Deskew and region extraction module plays an important role in this system. Its funtion is to correct skewed images and extract regions of input data. Its outputs are the inputs of the character recognition module. Basing on the directional and equilateral distribution of form document images, we used project profile method to detect skew angle - for this method is highly accurate when applied to this kind of image. On the other hand, in form documents, infomation is often entered into cells(surrounded by lines). Therefore, we extract entered information regions by detecting lines and using local co-ordination. Experimental results on variety of form documents show that our approach has achieved good and accurate results.

Keywords : detect skew angle, project profile method, form recognition

Chương 1 Giới Thiệu

1.1 Đặt vấn đề

Nhận dạng là bài toán đã xuất hiện khá lâu và đã đạt được nhiều thành tựu to lớn. Tuy nhiên nhận dạng một văn bản bất kì bao gồm cả các văn bản có lẫn chữ viết tay hay hình ảnh luôn là một bài toán khó và hiện nay vẫn chưa thật sự có giải pháp hoàn chỉnh. Để giải quyết bài toán nhận dạng hiện có nhiều xu hướng tiếp cận khác nhau tương ứng với những loại văn bản khác nhau. Trong đó, nhập dữ liệu tự động là phương pháp tiếp cận về nhận dạng đối với các văn bản kiểu form.

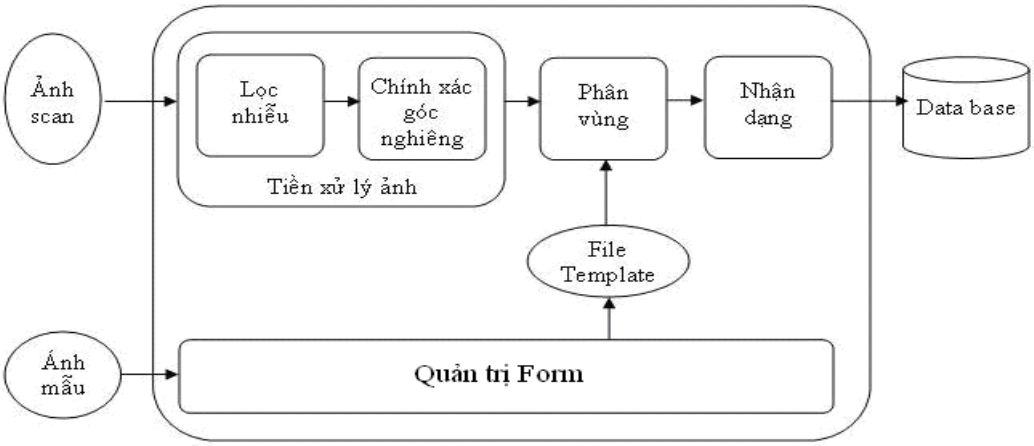
Càng ngày, nhu cầu xử lý dữ liệu của con người trên máy tính càng tăng lên. Chính vì lý do đó, bài toán nhập dự liệu tự động đang ngày càng thu hút được nhiều sự chú ý và đầu tư. Nó đã vượt ra ngoài khuôn khổ các lĩnh vực nghiên cứu và đang dần được áp dụng vào thực tế bởi khả năng áp dụng rộng rãi và những hiệu quả mà nó có thể mang lại.

Trên thế giới, hiện đã có nhiều ứng dụng liên quan đến vấn đề nhận dạng văn bản hay nhập dữ liệu tự động. Có thể kể đến như : sản phẩm FineReader, Scan To Office của hãng ABBYY, Smart scan Xpress của Pegasus Image, các ứng dụng chấm thi tự động …Ở Việt Nam cũng đã có các ứng dụng nhận dạng văn bản như VNDocR của Viện Công nghệ Thông tin hay ImageScan của CardPro. Đây là các ứng dụng nhận dạng chữ in. Việc nhận dạng chữ viết tay đang còn là một thách thức. Một số nghiên cứu về nhận dạng chữ viết tay đã được thực hiện tại Viện CNTT và Bộ môn Công nghệ Phần mềm. Tuy nhiên các ứng dụng này hiện vẫn còn rất nhiều hạn chế do khả năng nhận dạng chữ viết tay chưa đạt được độ chính xác cần thiết để có thể áp dụng rộng rãi trên thực tế.

Cùng với sự phát triển của công nghệ thông tin hiện nay, các thuật toán nhận dạng ngày càng chính xác và đưa ra được các kết quả đáng tin cậy. Ngay cả đối với chữ viết tay cũng có thể đạt được độ chính xác cao với điều kiện là chỉ nhận dạng từng chữ riêng biệt và chữ viết đẹp. Với các văn bản thông thường ta khó có thể đạt được điều này. Tuy nhiên, Các form nhập liệu là kiểu văn bản có cấu trúc và ta có thể đưa ra một số quy tắc ràng buộc để tăng độ chính xác cho việc nhận dạng - chẳng hạn như: các chữ được viết riêng rẽ trên các ô riêng biệt của các vùng nhập liệu. Mặt khác, việc nhận dạng không cần thiết phải tiến hành trên toàn bộ ảnh của tài liệu mà chỉ giới hạn

ở những vùng nhập dữ liệu. Đặc điểm này cũng cho phép ta tiếp cận bài toán một cách có hiệu quả hơn, chẳng hạn có thể sử dụng các thông tin sẵn có từ thiết kế form làm tham số nhận dạng. Một khía cạnh khác của nhận dạng form tài liệu là các dữ liệu nhận dạng được của mỗi vùng của form sẽ phải được tự động gắn vào một trường dữ liệu xác định của ứng dụng. Bài toán nhập liệu tự động từ form tài liệu sẽ gồm các vấn đề sau :

- Quản trị form bao gồm : thiết kế form nhập liệu ; quản lý và lưu trữ tự động các tham số của form để có thể cung cấp dữ liệu cho quá trình nhận dạng sau này nhanh và tin cậy ; tích hợp với cơ sở dữ liệu.
- Nhận dạng các vùng dữ liệu (bài toán phát hiện và phân vùng dữ liệu).
- Nhận dạng chữ viết tay trên các vùng dữ liệu ; xử lý từ vựng và ghi nhận vào cơ sở dữ liệu.
- Nhưng trước hết phải tiền xử lý ảnh để làm tốt ảnh, phục vụ cho quá trình nhận dạng, đảm bảo độ tin cây.



Hình 1: Sơ đồ hệ thống

Với số lượng công việc như vậy, đề tài chung được chia làm hai phần :

- Các giải pháp tối ưu cho tiền xử lý ảnh do Đinh Văn Phương thực hiện.
- Khử nghiêng văn bản bằng phương pháp phép chiếu và phân vùng ảnh do Nguyễn Thanh Phúc thực hiện.

Khóa luận này chỉ giới hạn tập trung trình bày về việc khử nghiêng văn bản bằng phép chiếu và phân vùng ảnh - các giải pháp và thực nghiệm. Bao gồm các công việc cụ thể như sau :

- Các thuật toán xử lý form văn bản :
  - Thuật toán xác định góc xoay dựa trên phép chiếu.
  - Thuật toán xác định các đường thẳng trong văn bản phục vụ cho việc xác định các vùng nhận dạng.
  - Phân vùng ảnh dựa trên các đường thẳng xác định được.
- Thực nghiệm
  - Thử nghiệm độ chính xác của các thuật toán.
  - Đánh giá kết quả, hiệu quả của thuật toán và nhận xét.
- Kết luận.

Cũng cần nói thêm rằng, đề tài này được đặt trong một dự án nghiên cứu phối hợp giữa Trung tâm Nghiên cứu và Phát triển Phần mềm và Bộ môn Công nghệ Phần mềm để đi đến một thương phẩm. Một phần mềm đóng gói đạt tính thương mại cần được xem là tiêu chuẩn cao nhất của các giải pháp pháp phần mềm.

1.2 Nội dung và cấu trúc của khóa luận

Bài toán con mà tôi thực hiện trong hệ thống chung là bài toán xác định góc nghiêng và phân vùng ảnh. Nắm bắt được khó khăn cũng như những đặc trưng của bài toán này, chúng tôi đã áp dụng một giải pháp có độ chính xác cao trong việc xác định góc nghiêm đó là sử dụng phương pháp phép chiếu, đồng thời sử dụng các đường thẳng có trong form để phân vùng, tách riêng ra các vùng cần xử lý.

Với nội dung chính là trình bày những lý thuyết cơ bản về xử lý ảnh, về các phương pháp xác định góc nghiêng, các phương pháp phân vùng và lựa chọn các giải pháp áp dụng vào bài toán, khóa luận được tổ chức như sau :

Chương 1: Giới thiệu

Phần đầu của chương giới thiệu về bài toán nhập dữ liệu tự động nói chung: tình hình Việt Nam và thế giới, các thành tựu đã đạt được trong lĩnh vực nhận dạng chữ viết, những khó khăn cũng như các đặc trưng của bài toán nhận dạng form nhập dữ liệu so với các bài toán nhận dạng khác. Phần tiếp theo giới thiệu về hệ thống chung mà nhóm chúng tôi đang tiến hành nghiên cứu và xây dựng : Nghiên cứu và xây dựng hệ thống nhập dữ liệu tự động bằng nhận dạng quang học, phạm vi giới hạn và quy trình giải quyết bài toán. Từ đó nêu lên nội dung mà tôi nghiên cứu và thực hiện trong bài toán chung thông qua việc trình bày nội dung và cấu trúc của khóa luận.

Chương 2: Tổng quan một số phương pháp khử nghiêng và phân vùng ảnh

Chương hai trình bày về các phương pháp khử nghiêng và phân vùng ảnh , các khái niệm và tầm quan trọng của khử nghiêng và phân vùng ảnh trong nhận dạng form. Xác định các ưu nhược điểm và miền áp dụng của mỗi phương pháp để từ đó lựa chọn giải pháp thích hợp.

Chương 3: Đề xuất giải pháp khử nghiêng và phân vùng ảnh

Chương này trình bày về phần việc chính mà tôi đã thực hiện trong đề tài chung đó là : Giải pháp cho việc khử nghiêng ảnh bằng phép chiếu và phân vùng ảnh. Nội dung của chương tập chung vào :

- Phân tích những đặc trưng của ảnh dạng form nhập liệu từ đó đưa ra giải pháp cho việc xác định góc nghiêng và phân vùng ảnh.
- Quy trình thực hiện các giải pháp này.
- Đánh giá ưu và nhược điểm của các phương pháp.

Chương 4: Thực nghiệm

Chương bốn mô tả chi tiết quá trình thực nghiệm với phương pháp khử nghiêng bằng phép chiếu và phân vùng ảnh cùng với thực nghiệm về hệ thống chung. Đồng thời chương cũng đề cập đến quá trình thu thập và xây dựng cơ sở dữ liệu ảnh dạng form sử dụng cho thực nghiệm.

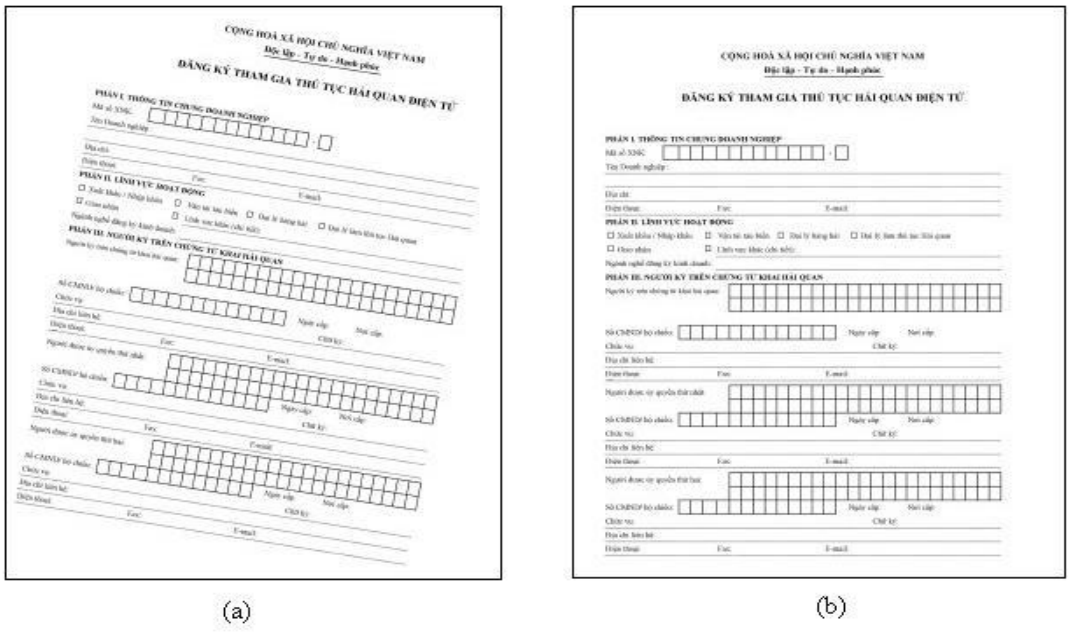
Chương 5: Kết luận

Chương năm tổng kết lại những kết quả đạt được và chưa đạt được trong quá trình chúng tôi nghiên cứu và thực hiện khóa luận. Từ đó nêu lên những kết quả cần hướng tới và hướng nghiên cứu, phát triển tiếp theo.

Chương 2 Tổng quan một số phương pháp khử nghiêng và phân vùng ảnh

2.1 Một số phương pháp khử nghiêng ảnh

Văn bản bị nghiêng xảy ra trong quá trình quét vào máy tính hay copy, điều này ảnh hưởng đến toàn bộ các đối tượng có trong văn bản nhất là các vùng mà ta cần phải nhận dạng. Văn bản bị nghiêng là một điều không thể tránh khỏi, và trong nhiều trường hợp gây ảnh hưởng không tốt đến độ chính xác đối với kết quả phân vùng và nhận dạng ký tự. Cũng có một số phương pháp về phân vùng ảnh không yêu cầu văn bản phải có góc nghiêng bằng không [4,13]. Tuy nhiên các phương pháp này vẫn đòi hỏi góc nghiêng của văn bản nằm trong một khoảng giới hạn cho phép. Bên cạnh đó, đơn giản hóa vấn đề này sẽ dẫn tới phức tạp hóa cũng như tốn thời gian xử lý đối với các nhiệm vụ khác. Do đó chính xác lại góc nghiêng của ảnh là một việc làm tất yếu và phải được thực hiện trước khi tiến hành phân vùng và nhận dạng ảnh.



Hình 2: (a) ảnh sau khi khử nhiễu và tách nền; (b) ảnh sau khi khử nghiêng

Một văn bản có rất nhiều các đặc trưng so với các loại hình ảnh khác như các đặc trưng về hướng, về cấu trúc phân bố các đối tượng ... Từ đó cũng có một số phương pháp

xác định góc nghiêng cho ảnh của văn bản. Dưới đây là chi tiết về một số phương pháp mà tôi cho là tiêu biểu.

Tài liệu tham khảo

Tài liệu tham khảo tiếng Việt

[1] N.T.M. Ánh, Đ.V. Cường, N.T. Hoài. Ứng dụng mạng Neural trong nhận dạng văn bản., Khoa Công Nghệ, ĐHQGHN. NCKH SV 2004

[2] Lương Mạnh Bá, Nguyễn Thanh Thủy. Nhập môn xử lý ảnh số. Nhà xuất bản khoa học và kỹ thuật 5/1999, tr143-144

[3] Phan Văn Thuận, Ứng dụng nhận dạng trong xử lý kết quả điều tra, Luận văn tốt nghiệp ngành công nghệ thông tin – Đại Học Quốc Gia Hà Nội, Khoa Công Nghệ, 2004, tr21-22

Tài liệu tham khảo tiếng Anh

[4] A. Antonacopoulos and R.T. Ritchings, Representation and Classification of Complex-shaped Printed Regions Using White Tiles, In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, 1995, tr.1132-1135.

[5] Bo Yuan, Leong Keong Kwoh, and Chew Lim Tan. Finding Best-Fit Bounding-Boxes. Center of Remote Imaging, Sensing and Processing National University of Singapore, Singapore 119260 ; Department of Computer Science, School of Computing National University of Singapore, Singapore 117543. 2001, tr2.

[6] Dipti Deodhare, NNR Ranga Suri, R.Amit. Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System. International Journal Of Computer Science & Application, Vol. II, No. II, pp. 131-144. 2005 tr131-135

[7] D. X. Le, G. Thoma. Document Skew Angle Detection Algorithm. Proc. 1993 SPIE Symposium on Aerospace and Remote Sensing - Visual Information Processing II, Orlando, FL April 14-16, 1993, Vol. 1961, tr. 251-262

[8] E.Kavallieratou, D.C.Balcan, M.F.Popa, N.Fakotakis IEEE member. Handwritten text localization in skewed documents. University of Bucharest 14, Academiei

St., 79543 Bucharest, RomaniaWire Communications Laboratory University of Patras, 26500 Patras, Greece. Int. Conference on Document Analysis and Recognition, ICDAR'99. 1999, tr. 705 – 708

[9] Fu Chang, Chien-Hsing Chou, and Shih-Yu Chu. A New Approach to Estimation of Document Skew Angles Based on Piecewise Linear Approximation of Line Objects. Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. NSC93-2422-H-001-0004. CVGIP(Graphical Models and Image Processing), 2004, tr1-3

[10] Hanchuan Peng, Member, IEEE, Fuhui Long, and Zheru Chi, Member, IEEE. Document Image Recognition Based on Template Matching of Component Block Projections. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL.25, NO.9, SEPTEMBER 2003. tr1188-1192

[11] Junichi Kanai, Andrew D. Bagdanov. Projection profile based skew estimation algorithm for JBIG compressed images. IJDAR(1998), tr43-51

[12] J. L. Chen and H. J. Lee. An efficient algorithm for form structure extraction using strip projection. Pattern Recognition, 31(9): 1353–1368, May 1998. tr1353–1368

[13] K. Etemad, D. Doermann, and R. Chellappa, Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration, IEEE Trans. on Pattern Recognition and Machine Intelligence, Vol. 19, No. 1, 1997, tr. 92-96.

[14] Oleg Okun, Matti Pietik¨ainen and Jaakko Sauvola. Robust Skew Estimation on Low-Resolution Document Images. Machine Vision and Media Processing Group, Infotech Oulu and Dept. of EE, University of Oulu. Tr1-4

[15] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena. Geometric layout analysis techniques for document image understanding: a review. Technical report, IRST, Trento, Italy, 1998. tr5

[16] Shamik Sural, CMC Ltd. 28 Camac Street, Calcutta 700 016, India ; P.K.Das, Dept. of CSE. Jadavpur University, Calcutta 700 032, India. A Document Image Analysis System on Parallel Processors. tr2-3

[17] Yue Lu, Chew Lim Tan. A nearest-neighbor chain based approach to skew estimation in document images. Pattern Recognition Letters 24 (2003)2315–2323, Department of Computer Science, School of Computing National University of Singapore, Kent Ridge, Singapore 117543, 2003, tr2315-2319

[18] Z.Shi, V.Govindaraju, Skew Detection for Complex Document Images Using Fuzzy Run length, Proc. Of the Seventh Int. Conf. on Document Analysis and Recognition, ICDAR'03. tr1-4