ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÙI PHƯƠNG THẢO

PHƯƠNG PHÁP PHÂN TÍCH TRANG VĂN BẢN DỰA TRÊN TAB-STOP

Chuyên ngành : Khoa học máy tính
Mã số : 60.48.01

Luận văn thạc sĩ khoa học máy tính

Người hướng dẫn khoa học:
TS. Nguyễn Đức Dũng

Thái Nguyên, 2012

MỞ ĐẦU

1. Đặt vấn đề

Hiện nay, hầu hết tài liệu của con người đều đã được số hóa và được lưu trữ trên máy tính, việc số hóa đảm bảo tính an toàn và thuận tiện hơn hẳn so với sử dụng tài liệu giấy. Tuy nhiên việc sử dụng giấy để lưu trữ tài liệu trong một số mục đích là không thể thay thế hoàn toàn được (như sách, báo, tạp chí, công văn,…). Hơn nữa, lượng tài liệu được tạo ra từ nhiều năm trước vẫn còn rất nhiều mà không thể bỏ đi được vì tính quan trọng của chúng.

Việc chuyển đổi tài liệu điện tử sang tài liệu giấy có thể thực hiện được dễ dàng bằng cách in hay fax, nhưng công việc ngược lại là chuyển từ tài liệu giấy sang tài liệu điện tử lại là một vấn đề không hề đơn giản. Chúng ta mong muốn có thể số hóa tất cả các tài liệu, sách, báo đó và lưu trữ chúng trên máy tính, việc tổ chức và sử dụng chúng sẽ thuận tiện hơn rất nhiều. Vậy nhưng giải pháp sẽ là gì?

Công nghệ đang phát triển một cách chóng mặt, các máy scan với tốc độ hàng nghìn trang một giờ, các máy tính với công nghệ xử lí nhanh chóng và chính xác một cách siêu việt. Vậy tại sao chúng ta không quét các trang tài liệu vào và xử lý, chuyển chúng thành các văn bản một cách tự động? Nhưng vấn đề là khi quét chúng ta chỉ thu được các trang tài liệu đó dưới dạng ảnh nên không thể thao tác, sửa chữa, tìm kiếm như trên các bản Office được, khi đó máy tính không phân biệt được đâu là điểm ảnh của chữ và đâu là điểm ảnh của đối tượng đồ họa.

Một giải pháp được đưa ra đó là xây dựng các hệ thống nhận dạng chữ trong các tấm ảnh chứa cả chữ và đối tượng đồ họa, sau đó chuyển thành dạng trang văn bản và có thể mở, soạn thảo được trên các trình soạn thảo văn bản. Một cách tổng quát thì cách thức hoạt động của một hệ thống nhận dạng chữ đó như sau [5]:

1. Chụp ảnh hoặc scan các trang tài liệu và lưu lại trên máy tính dưới dạng hình ảnh.

2. Phân tích hình ảnh sau khi quét, đọc được ký tự trên hình ảnh và ghi lại vào máy tính theo cách mà máy tính quản lý được thông tin dữ liệu đó.

- Bước 1: phân tích cấu trúc của ảnh tài liệu, từ đó xác định đâu là phần chứa ký tự, đâu là phần chứa cả ảnh lẫn ký tự và đâu chỉ chứa hình ảnh. Bước này thực sự quan trọng cho bước nhận dạng. Bởi nó định vị chính xác cho việc áp dụng các thuật toán nhận dạng lên vùng đã xác định tính chất, nếu bước này chính xác trước tiên nó hạn chế thời gian cho việc nhận dạng, sau là tăng ngữ nghĩa bổ sung cho việc nhận dạng.
- Bước 2: nhận dạng ký tự dựa vào các tính chất của ký tự, ví dụ như sắp xếp theo dòng, khoảng cách giữa 2 từ lớn hơn khoảng cách giữa 2 ký tự, dùng trí tuệ nhân tạo để dự đoán các ký tự kề nhau phải như thế nào, các từ trong câu phải như thế nào để câu có nghĩa. Từ đó có nội dung đúng để lưu trữ, quản lý….

Trong thực tế không phải quá trình nhận dạng nào cũng chỉ trải qua hai bước như trên, bởi vì có rất nhiều tham số ảnh hưởng đến kết quả của các chương trình nhận dạng như nhiễu, Font chữ, kích thước chữ, kiểu chữ nghiêng, đậm, gạch dưới. Ngoài ra các dòng chữ cũng có thể trộn lẫn với các đối tượng đồ họa, vì thế trước khi nhận dạng chữ, một số thao tác tiền xử lý sẽ được tác động lên ảnh như, lọc nhiễu, chỉnh góc nghiêng và đặc biệt quan trọng là phân tích trang tài liệu để xác định cấu trúc của trang văn bản đồng thời tách biệt hai thành phần là chữ và các đối tượng đồ họa.

2. Nội dung nghiên cứu

2.1.Mục tiêu nghiên cứu chính của đề tài

- Tìm hiểu cấu trúc trang tài liệu (cấu trúc vật lý, logic)
- Tìm hiểu một số kỹ thuật phân tích trang tài liệu (phân vùng, phân đoạn, top-down hay bottom-up, …)
- Trình bày kỹ thuật phân tích trang văn bản Tab-Stop
- Cài đặt thử nghiệm một giải pháp phân tích trang văn bản trên kỹ thuật Tab-Stop.

- Từ kết quả nghiên cứu có một sự chuẩn bị kiến thức đầy đủ cho bước nghiên cứu tiếp theo là nhận dạng ký tự quang.

2.2.Ý nghĩa khoa học của đề tài

- Giải quyết được vấn đề về học thuật: đề tài sẽ mang ý nghĩa cung cấp về mặt lý thuyết để làm rõ về các phương pháp phân tích trang tài liệu.
- Đáp ứng được yêu cầu của thực tiễn: từ các lý thuyết đã được nghiên cứu, từ đó liên hệ và gắn vào thực tiễn để có thể áp dụng vào các lĩnh vực như: Số hóa tài liệu, lưu trữ thư viện, điện tử hóa văn phòng, nhận dạng và xử lý ảnh, …

2.3.Nhiệm vụ nghiên cứu

Mục đích của luận văn đề cập được đến hai phần:

- Phần lý thuyết: Nắm rõ và trình bày những cơ sở lý thuyết liên quan đến cấu trúc trang tài liệu, một số kỹ thuật phân tích trang tài liệu, từ đó có để có thể xác định tính quan trọng của bước này trong nhận dạng ký tự, đồng thời hiểu các công việc kế tiếp cần làm trong bước nhận dạng ký tự.
- Phần phát triển ứng dụng: Áp dụng các thuật toán đã trình bày ở phần lý thuyết từ đó lựa chọn một giải pháp tối ưu và cài đặt thử nghiệm chương trình phân tích trang tài liệu.

2.4. Phương pháp nghiên cứu

- Tìm kiếm, tham khảo, tổng hợp tài liệu từ các nguồn khác nhau để xây dựng phần lý thuyết cho luận văn.
- Sử dụng các kỹ thuật được áp dụng phân tích trang tài liệu để làm rõ bản chất của các vấn đề được đưa ra trong phần lý thuyết.
- Xây dựng chương trình Demo.

2.5. Phạm vi nghiên cứu

Bài toán phân tích trang tài liệu đã được phát triển với nhiều thành tựu trong thực tế, có rất nhiều thuật toán tối ưu đã được các nhà khoa học đề nghị. Tuy nhiên có thể nói chưa có một chương trình nào có thể "đọc" một ảnh văn bản như con người, vì thực tế có rất nhiều kiểu trang văn bản khác nhau, khác nhau về cấu trúc

trình bày, ngôn ngữ, kiểu font, chữ viết tay,… Đây thực sự là một bài toán lớn, chính vì thế trong phạm vi của luận văn chỉ tìm hiểu một số kỹ thuật phân tích trang văn bản tiêu biểu với mục đích để so sánh với một thuật toán mới chưa được đưa ra ở các đề tài trước. Cuối cùng, dựa vào đó để xây dựng Demo cho một ứng dụng.

Các kết quả nghiên cứu dự kiến cần đạt được:

- Tìm hiểu tài liệu liên quan đến lĩnh vực quan tâm để nắm bắt được bản chất vấn đề đặt ra.
- Báo cáo lý thuyết
- Chương trình Demo.

3. Bố cục của luận văn

Nội dung của luận văn được trình bày trong ba chương với nội dung chính sau:

Chương 1: Trình bày nội dung trang văn bản và các phương pháp tiền xử lý trang văn bản, cấu trúc trang tài liệu và quá trình phân tích trang tài liệu

Chương 2: Trình bày một số phương pháp phân tích trang tài liệu, từ đó đánh giá ưu nhược điểm để lựa chọn kỹ thuật Tab-Stop cho chương trình thử nghiệm.

Chương 3: Cài đặt chương trình Demo và đánh giá kết quả chương trình

CHƯƠNG 1. NỘI DUNG TRANG VĂN BẢN VÀ CÁC PHƯƠNG PHÁP TIỀN XỬ LÝ

Chương này đưa ra các khái niệm về đối tượng làm việc của đề tài là ảnh tài liệu, khái niệm về cấu trúc vật lý và cấu trúc logic. Giới thiệu các khâu trong một hệ thống nhận dạng chữ viết hoàn chỉnh. Đồng thời đưa ra một số phần mềm nhận dạng của Việt Nam và Thế giới cùng với các mẫu kết quả phân tích của nó nhằm mục đích so sánh và xác định phạm vi cho đề tài.

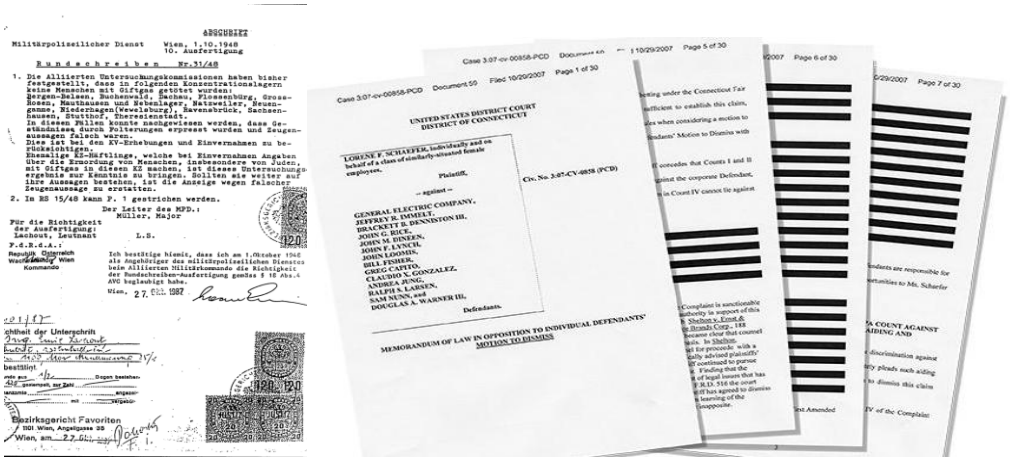
1.1. Ảnh tài liệu và nhận dạng ảnh tài liệu

1.1.1. Tổng quan về ảnh tài liệu

Trang ảnh tài liệu được đề cập ở đây là các file ảnh số hoá thu được bằng cách quét các trang tài liệu dùng máy scanner, hoặc chụp từ các máy ảnh số, hay nhận từ một máy fax (Hình 1), file ảnh này được lưu giữ trong máy tính. Ảnh tài liệu có nhiều loại: ảnh đen trắng, ảnh đa cấp xám, ảnh đa cấp xám với các phần mở rộng như TIF, BMP, PCX, …(Hình 2) và ảnh tài liệu được đưa ra trong luận văn này là **ảnh đa cấp xám**.

Tài liệu | Thiết bị thu nhận ảnh | Ảnh số tài liệu

Hình 1: Sơ đồ tổng quan quá trình tạo ảnh tài liệu



Hình 2: Ví dụ ảnh tài liệu

1.1.2. Nhận dạng tài liệu và vai trò của phân tích ảnh tài liệu

Ngày nay, máy tính đang phát triển mạnh mẽ, tốc độ xử lý không ngừng được nâng lên. Cùng với nó là sự ra đời của các phần mềm thông minh đã khiến máy tính ngày một gần gũi với con người hơn. Một trong các khả năng tuyệt vời của con người mà các nhà khoa học máy tính muốn đạt được đó là khả năng nhận dạng và lĩnh vực nhận dạng thu được nhiều thành công nhất là nhận dạng ký tự quang OCR–Optical Character Recognition. OCR có thể được hiểu là quá trình chuyển đổi tài liệu dưới dạng file ảnh số hoá (là dạng chỉ có người đọc được) thành tài liệu dưới dạng file văn bản (là tài liệu mà cả người và máy đều có thể đọc được).

OCR có rất nhiều ứng dụng hữu ích trong cuộc sống như:

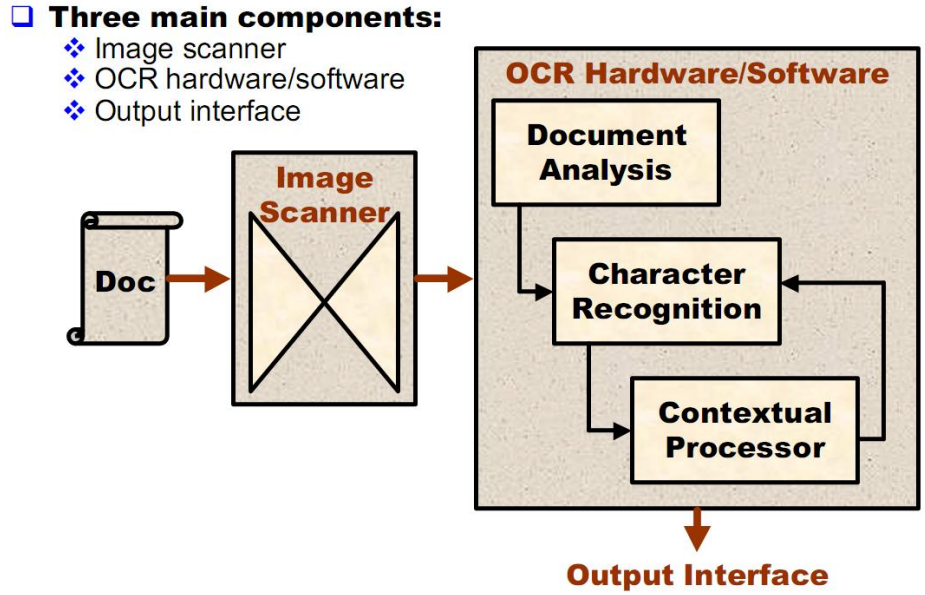
- Sắp xếp thư tín, dựa vào việc nhận dạng mã bưu chính (Zipcode) hay địa chỉ gửi tới.
- Tự động thu thập dữ liệu từ các mẫu đơn/báo biểu hay từ các hồ sơ lao động.
- Hệ thống tự động kiểm tra trong ngân hàng (tự động xác nhận chữ ký)
- Tự động xử lý các hóa đơn hay các yêu cầu thanh toán
- Hệ thống tự động đọc và kiểm tra passport
- Tự động phục hồi và copy tài liệu từ các ảnh quét.
- Máy đọc cho những người khiếm thính
- Các ứng dụng Datamining
- …

Sơ đồ một hệ thống OCR cơ bản ở Hình 3.

Trong đó:

- Scanner: Thiết bị quét ảnh
- OCR hardware/software:
  - Document analysis: Phân tích tài liệu
  - Character recognition: Nhận dạng ký tự
  - Contexttual processor: Xử lý văn cảnh
- Output interface: Đầu ra

❖ Như vậy vai trò chính của khâu phân tích ảnh tài liệu là việc phân đoạn trang, tách vùng văn bản ra khỏi nền và đồ họa tạo mẫu chuẩn cho khâu nhận dạng. Rõ ràng là kết quả của khâu phân tích này ảnh hưởng rất lớn đến hiệu qủa của khâu nhận dạng nếu sử dụng mẫu hay các chuỗi văn bản đầu ra của nó.



Hình 3: Sơ đồ OCR cơ bản

1.2. Cấu trúc của ảnh tài liệu

Một khái niệm mấu chốt trong xử lý tài liệu đó là cấu trúc của tài liệu. Cấu trúc tài liệu thu được từ việc liên tiếp chia nhỏ nội dung của tài liệu thành các phần nhỏ đơn vị (tức không thể phân chia được nữa) và chúng được gọi là các đối tượng cơ sở (basic objects). Còn tất cả các đối tượng khác được gọi là các đối tượng hỗn hợp.

Có hai loại cấu trúc của tài liệu được quan tâm ở đây đó là cấu trúc vật lý

(hay bố cục vật lý) và cấu trúc logic mô tả mối quan hệ logic giữa các vùng đối tượng trong tài liệu.

1.2.1. Cấu trúc vật lý

Bố cục vật lý của một tài liệu mô tả vị trí và các đường danh giới giữa các vùng có nội dung khác nhau trong một trang tài liệu[6]. Quá trình phân tích bố cục tài liệu là thực hiện việc tách từ một trang tài liệu ban đầu thành các vùng có nội dung cơ sở như hình ảnh nền, vùng văn bản,…

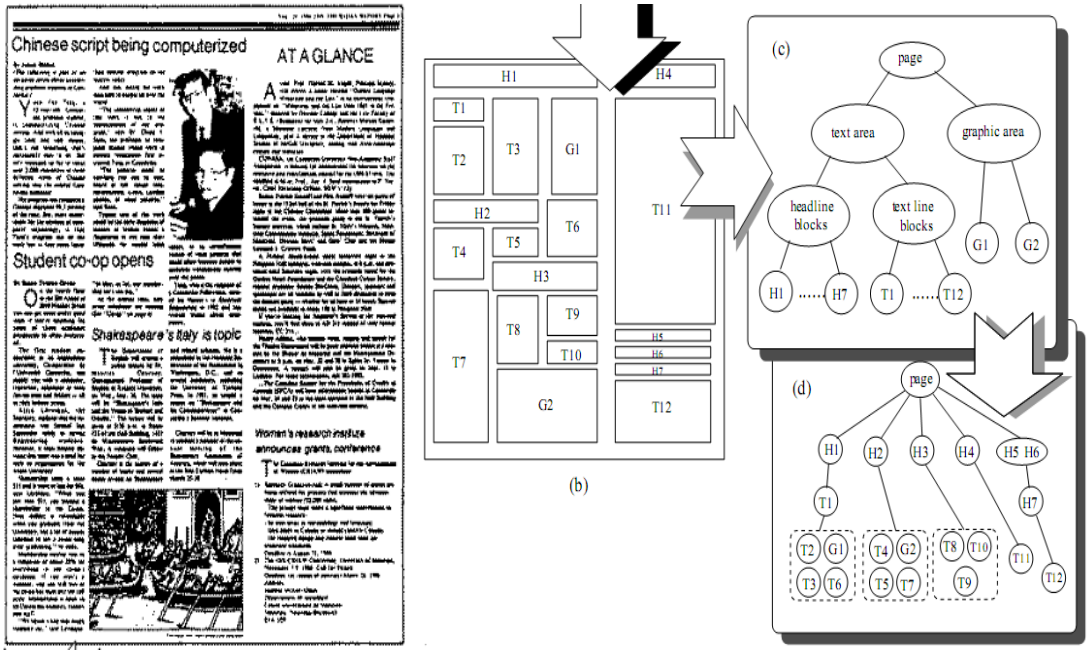
Để mô tả bố cục vật lý của tài liệu người ta sử dụng một cấu trúc hình học với mỗi đối tượng trong cấu trúc là một phần tử chỉ chứa nội dung đồng nhất. Các kiểu đối tượng hình học được định nghĩa như sau[4]:

- Block là đối tượng cơ sở tương ứng với một vùng hình chữ nhật chứa một phần nội dụng của tài liệu.
- Frame một đối tượng hỗn hợp tương ứng với một hình chữ nhật bao gồm một hoặc nhiều block hoặc bao gồm các frame.
- Page là đối tượng hình học hoặc hỗn hợp các thành phần cơ sở tương ứng với một vùng hình chữ nhật, nếu là đối tượng hỗn hợp nó chứa một hoặc nhiều block, một hoặc nhiều frame.
- Page set (tập trang) là một tập của một hoặc nhiều page.
- Điểm gốc của cấu trúc (hay nút gốc) là một đối tượng ở mức cao nhất trong sơ đồ phân cấp của cấu trúc hình học tài liệu. Hình 4(b) cho ví dụ một cấu trúc hình học mô tả bố cục vật lý của trang tài liệu tương ứng.

Các thuật toán phân tích bố cục tài liệu có thể được chia làm ba loại chính dựa theo phương pháp thực hiện của nó.

- Bottom-up: Ý tưởng chính của các thuật toán loại này là bắt đầu từ những phần tử nhỏ nhất (như từ các pixel hay các phần tử liên thông) sau đó liên tục nhóm chúng lại thành các vùng lớn hơn.
- Top-down: Thuật toán này bắt đầu từ vùng lớn nhất chứa cả trang tài liệu sau đó liên tục phân chia thành các vùng nhỏ hơn.

- Các thuật toán không theo thứ bậc: như Fractal Signature, Adaptive split-and-merge …



Hình 4: b-Cấu trúc vật lý: c,d-Cấu trúc logic của một tài liệu[4]

1.2.2. Cấu trúc logic

Ngoài bố cục vật lý, các trang tài liệu còn chứa đựng nhiều thông tin về ngữ cảnh và nội dung như các tiêu đề, đoạn văn, đề mục, …và mỗi vùng nội dung này lại được gán các nhãn logic hay nhãn theo chức năng tương ứng, khác biệt hoàn toàn với các nhãn trong bố cục vật lý. Hầu hết các tài liệu đều có một quy tắc đọc để có thể hiểu hết nội dung của tài liệu. Với một số ngôn ngữ đặc biệt như tiếng Trung, tiếng Ả rập lại có quy cách đọc khác biệt (như đọc từ phải qua trái, trên xuống). Tập hợp tất cả các yếu tố logic và chức năng trong một tài liệu và mối quan hệ giữa chúng được gọi là cấu trúc logic của tài liệu[6]. Thông thường pha phân tích cấu trúc logic của tài liệu được thực hiện trên kết quả của bước phân tích bố cục vật lý. Tuy nhiên với một số loại tài liệu phức tạp, thì pha phân tích bố cục vật lý lại cần thêm một số thông tin logic liên quan đến các vùng để có thể phân đoạn một cách chính xác. Hình 4(c,d) mô tả một ví dụ cấu trúc logic của tài liệu.