Mục lục

Danh mục các hình ảnh

MỞ ĐẦU

I. Đặt vấn đề

Ngày nay việc sử dụng máy tính để lưu trữ tài liệu không còn là vấn đề mới mẻ và cần phải chứng minh tính an toàn, thuận tiện của nó. Tuy nhiên việc sử dụng giấy để lưu trữ tài liệu trong một số mục đích vẫn không thể thay thế được (như báo, sách, 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.

Chúng ta mong muốn có thể điện tử hóa hàng tỉ trang tài liệu đó và cất chúng chỉ trong một ổ cứng kích thước bằng một cuốn sách nhỏ, tìm kiếm thông tin mà chỉ cần tốn vài giây với một cái gõ phím Enter. Giải pháp là gì?

Thông thường người ta sẽ phải thuê người cùng với việc tốn hàng tháng, hàng năm mới có thể nhập vào máy tính được hết lượng tài liệu đó. Hiện nay chúng ta đã có các máy Scan với tốc độ cao, công nghệ xử lý của máy tính ngày càng siêu việt với tốc độ tính toán vượt cả tốc độ ánh sáng, vậy tại sao chúng ta không quét toàn bộ các trang tài liệu vào và chuyển chúng thành văn bản một cách tự động?

Bằng cách đó tốc độ và tính chính xác sẽ tăng hàng trăm lần trong khi chi phí lại là cực tiểu. Vấn đề là khi quét vào máy tính chúng ta không thu được ngay các dòng văn bản từ các trang tài liệu kia, để có thể soạn thảo, sửa chữa và tìm kiếm như làm trên Office. Tất cả những gì thu được chỉ là các tấm ảnh của các trang văn bản, máy tính lại đối xử công bằng như nhau với mọi điểm ảnh, máy tính không có "mắt" như chúng ta để biết đâu là điểm ảnh của chữ, đâu là điểm ảnh của đối tượng đồ họa.

Một giải pháp được nghĩ đến ngay đó là đó là xây dựng các hệ thống nhận dạng chữ, trong tấm ảnh chứa cả chữ và đối tượng đồ họa cần tách và chuyển thành dạng trang văn bản, từ đó có thể mở và soạn thảo được trên các trình soạn thảo văn bản.

Một cách tổng quát thì cách thức làm việc của một hệ thống nhận dạng chữ như sau[5]:

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

2. Sử dụng một chương trình xử lý ảnh để phân tích hình ảnh sau khi quét, đọc được ký tự trên hình ảnh đó và ghi lại vào máy tính theo cách mà máy tính quản lý được thông tin dữ liệu đó.
    a. Bước 1 là phân tích cấu trúc của ảnh tài liệu, từ đó xác định đâu là phần chứa chữ, 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.
    b. 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 (phi chữ).

II. Nội dung nghiên cứu

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,…)
    - Cài đặt thử nghiệm một giải pháp phân tích có hiệu quả cao so với các phương pháp truyền thống như top-down hay bottom-up trên **ảnh vào là ảnh đa cấp xám có cấu trúc phức tạp.**
    - 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. 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: Lưu trữ thư viện, điện tử hóa văn phòng, nhận dạng và xử lý ảnh, …

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.

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.

5. Phạm vi nghiên cứu

Bài toán nhận dạng và xử lý ảnh 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 và 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.

III. 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 các khái niệm và mô hình tổng quát của hệ thống nhận dạng chữ viết, cùng với một số phần mềm nhận dạng tiêu biểu hiện nay.

Chương 2: Trình bày một số phương pháp phân tích trang tài liệu, từ đó đánh giá ưu nhược điểm để lựa chọn phương pháp Fractal Signature cho chương trình thử nghiệm. Trình bày về thiết kế cho chương trình demo.

Chương 3: Trình bày chi tiết về việc cài đặt chương trình cũng như các thủ tục sử dụng trong chương trình với phương pháp phân tích Fractal Signature và ảnh đầu vào là ảnh đa cấp xám có độ phức tạp cao.

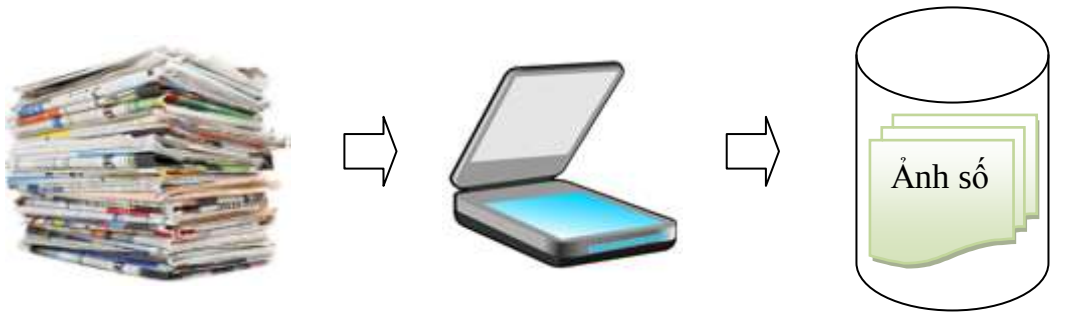
Chương I. TỔNG QUAN VỀ NHẬN DẠNG CHỮ VIẾT VÀ PHÂN TÍCH TRANG TÀI LIỆU

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.

I.1. Ảnh tài liệu và nhận dạng ảnh tài liệu

I.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, 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 màu, 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**.



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