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

BÙI THỊ THI

PHÁT HIỆN CẤU TRÚC BẢNG TRONG NHẬN DẠNG VĂN BẢN

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

LỜI CẢM ƠN 6
DANH SÁCH CÁC HÌNH ẢNH.....7
MỞ ĐẦU.....8
CHƯƠNG 1 10
TỔNG QUAN VỀ XỬ LÝ ẢNH VÀ HỆ PHÂN TÍCH TÀI LIỆU ẢNH....10
1.1. Tổng quan về xử lý ảnh.....10
1.1.1. Xử lý ảnh 10
1.1.2. Các bước cơ bản trong xử lý ảnh.....10
1.1.2.1. Thu nhận ảnh.....11
1.1.2.2. Tiền xử lý 11
1.1.2.3. Phân đoạn ảnh 12
1.1.2.4. Biểu diễn và mô tả 13
1.1.2.5. Nhận dạng và nội suy ảnh.....14
1.1.2.6. Cơ sở tri thức 14
1.1.2.7. Trích chọn đặc điểm.....15
1.2. Hệ phân tích tài liệu ảnh.....15
1.2.1. Tài liệu ảnh 15
1.2.2. Hệ phân tích trang tài liệu ảnh.....16
1.2.3. Các bước xử lý của một hệ phân tích tài liệu ảnh 17
1.2.3.1. Thu nhận dữ liệu ảnh 18
1.2.3.2. Tiền xử lý điểm ảnh 18
1.2.3.2.1. Xử lý nhị phân 18
1.2.3.2.2. Khử nhiễu.....19
1.2.3.3. Phân đoạn ảnh 20
1.2.3.4. Làm mảnh và xác định vùng.....20
1.2.3.5. Mã hóa Chain Code và vector hóa.....21
1.2.4. Phân tích đặc trưng của tài liệu ảnh.....22
1.2.5. Phân tích đối tượng văn bản trong tài liệu ảnh.....23
1.2.5.1. Xác định góc nghiêng của văn bản 23
1.2.5.2. Phân tích bố cục của trang tài liệu ảnh 25
CHƯƠNG 2 27

PHƯƠNG PHÁP PHÁT HIỆN BẢNG TESSERACT, PHÂN TÍCH BẢNG T-RECS TRONG TRANG ẢNH TÀI LIỆU 27
2.1. Phương pháp phát hiện bảng T-Recs trong trang ảnh tài liệu 27
2.1.2. Phân tích cấu trúc văn bản thông qua phát hiện TAB-STOP.....29
2.1.3. Phương pháp phát hiện bảng Tesseract 31
Thuật toán phát hiện bảng được xây dựng với hai thành phần của mô đun phân tích cấu trúc sau: 31
2.1.3.1. Xác định các phần của bảng 32
2.1.3.2. Xác định các trang cột phân tách.....34
2.1.3.3 Xác định các cột bảng.....34
2.1.3.4. Đánh dấu các vùng bảng.....34
2.1.3.5. Loại bỏ các lỗi 35
2.2. Phương pháp phân tích bảng T-Recs 35
2.2.1. Giới thiệu 35
2.2.2. Thuật toán phân đoạn khởi tạo 37
CHƯƠNG 3 39
CÀI ĐẶT THỬ NGHIỆM VÀ ĐÁNH GIÁ.....39
3.1. Môi trường cài đặt và dữ liệu kiểm thử.....39
3.2. Trình tự thực hiện của thuật toán 39
3.3. Kết quả thực nghiệm 39
3.4. Đánh giá 45
KẾT LUẬN.....51
TÀI LIỆU THAM KHẢO.....52

LỜI CẢM ƠN

Trước hết em muốn được gửi lời cảm ơn đến các thầy, cô giáo ở Viện Công nghệ thông tin, trường ĐH Công nghệ Thông tin và Truyền thông… đã quan tâm tổ chức chỉ đạo, quản lý lớp, trực tiếp giảng dạy khóa học của chúng em.

Em xin được bày tỏ lòng biết ơn sâu sắc tới thầy giáo TS. Nguyễn Đức Dũng – Viện Công nghệ Thông tin – Viện Khoa học Việt Nam, người thầy đã tận tình giúp đỡ, chỉ bảo em trong suốt quá trình tìm hiểu, viết đề cương và phát triển luận văn. Em xin được cảm ơn thầy giáo Lê Đức Hiếu – Viện Công nghệ Thông tin – Viện Khoa học Việt Nam người thầy đã tận tình giúp đỡ, chỉ bảo em trong suốt quá trình phát triển phần cài đặt, thử nghiệm.

Mặc dù đã có cố gắng song khả năng của bản thân em còn rất nhiều hạn chế nên luận văn không tránh khỏi những thiếu sót. Em rất mong chỉ bảo, góp ý của các thầy cô giáo và các bạn để luận văn của em được hoàn thiện hơn.

DANH SÁCH CÁC HÌNH ẢNH

Hình 1.1 *Quá trình xử lý ảnh*

Hình 1.2 *Các bước cơ bản trong quá trình xử lý ảnh*

Hình 1.3 *Lân cận các điểm ảnh của tọa độ (x, y)*

Hình 1.4 *Văn bản bị nghiêng sau khi được quét qua máy quét*

Hình 2.1 *Kết quả đầu ra của các bước khác nhau của các mô-đun phân tích trí trong tài liệu ảnh*

Hình 2.2 *Kết qủa của các bước khác nhau trong việc phân tích bố trí của c vùng bảng của Tesseract's*

Hình 2.3 *Kết quả của các bước khác nhau trong thuật toán phát hiện bảng*

Hình 2.4 *Ví dụ minh họa tư tưởng của thuật toán khởi tạo*

Hình 2.5 *Thuật toán khởi tạo đối với một đoạn văn bản*

Hình 3.1 *Phát hiện một phần*

Hình 3.2 *Chia nhỏ bảng*

Hình 3.3 *Gộp bảng với vùng văn bản*

Hình 3.4 *Phát hiện sai*

Hình 3.5 *Kết quả thực nghiệm 1*

Hình 3.6 *Kết quả thực nghiệm 2*

Hình 3.7 *Kết quả thực nghiệm 3*

Hình 3.8 *Kết quả thực nghiệm 4*

Hình 3.9 *Kết quả thực nghiệm 5*

MỞ ĐẦU

Xử lý ảnh là một trong những chuyên ngành quan trọng và lâu đời của Công nghệ thông tin. Xử lý ảnh được áp dụng trong nhiều lĩnh khác nhau như y học, vật lý, hoá học, tìm kiếm tội phạm, trong quân sự và trong một số lĩnh vực khác....

Phần lớn con người thu nhận thông tin bằng thị giác, cụ thể đó là các hình ảnh. Vì vậy xử lý ảnh là vấn đề không thể thiếu và hết sức quan trọng để thu được hình ảnh tốt hơn, đẹp hơn, nhằm đáp ứng yêu cầu thông tin khác nhau của người nhận.

Một trong những lĩnh vực của xử lý ảnh đó là xử lý, nhận dạng thông tin chứa đựng trong các tài liệu ảnh, tài liệu ảnh đa dạng, phức tạp không đơn thuần là các ký tự văn bản, hình vẽ, hình ảnh, bảng biểu…Trong đó phát hiện các bảng trong các tài liệu hình ảnh là một khâu rất quan trọng vì không những chúng ta phải xác định các thông tin chứa trong các bảng mà hầu hết các phương pháp hiện nay đều gặp khó khăn trong việc nhận diện các bảng. Các phương pháp phát hiện các bảng hiện nay tập trung chủ yếu vào các bảng chỉ có một cột mà nó không làm việc tốt với các bảng có nhiều dạng khác nhau.

Xuất phát từ thực tế đó, luận văn lựa chọn đề tài "Phát hiện cấu trúc bảng trong nhận dạng văn bản". Mục đích chính của đề tài là tìm hiểu các phương pháp phát hiện cấu trúc bảng, trình bày, cài đặt một thuật toán phát hiện các bảng với độ chính xác cao áp dụng cho các dạng tài liệu phức tạp như: các báo cáo của các công ty, các bài báo, các trang tạp chí,…

Ngoài phần mở đầu, kết luận luận văn được chia làm 3 chương cụ thể như sau:

Chương 1: Tổng quan về xử lý ảnh và hệ phân tích tài liệu ảnh

Trong chương này trình bày sơ lược về xử lý ảnh, giới thiệu các bước xử lý trong một hệ thống xử lý ảnh, tổng quan về hệ phân tích tài liệu ảnh và các thành phần chính trong hệ phân tích tài liệu ảnh: lấy dữ liệu, xử lý ảnh, trích chọn đặc trưng, nhận dạng đối tượng ảnh và nhận dạng văn bản.

Chương 2: Phương pháp phát hiện bảng Tesseract, phân tích bảng T-Recs trong trang tài liệu ảnh

Trình bày các phương pháp phát hiện bảng, thuật toán phát hiện cấu trúc bảng. Minh họa phát hiện cấu trúc bảng trong trang ảnh tài liệu.

Tìm hiểu về thuật toán T-Recs do Thomas G.Kieninger [7] đề xuất.

Chương 3: Cài đặt thử nghiệm và đánh giá

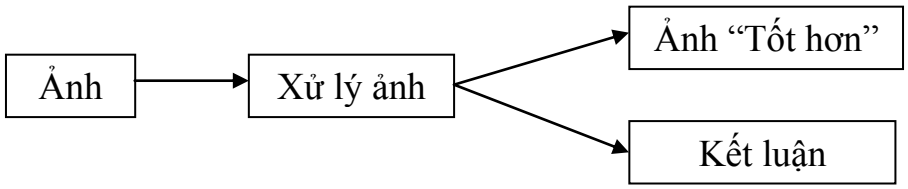
Mô tả chi tiết quá trình cài đặt thử nghiệm thuật toán, cũng như đánh giá các kết quả đạt được trên bộ dữ liệu thu thập được.

CHƯƠNG 1
TỔNG QUAN VỀ XỬ LÝ ẢNH VÀ HỆ PHÂN TÍCH TÀI LIỆU ẢNH

1.1. Tổng quan về xử lý ảnh

1.1.1. Xử lý ảnh

Quá trình xử lý nhận dạng ảnh là một quá trình thao tác nhằm biến đổi một ảnh đầu vào để cho ra một kết quả mong muốn. Kết quả đầu ra của một quá trình xử lý ảnh có thể là một ảnh "tốt hơn" hoặc một kết luận[1].



Hình 1.1: Quá trình xử lý ảnh

Như vậy mục tiêu của xử lý ảnh có thể chia làm ba hướng như sau:

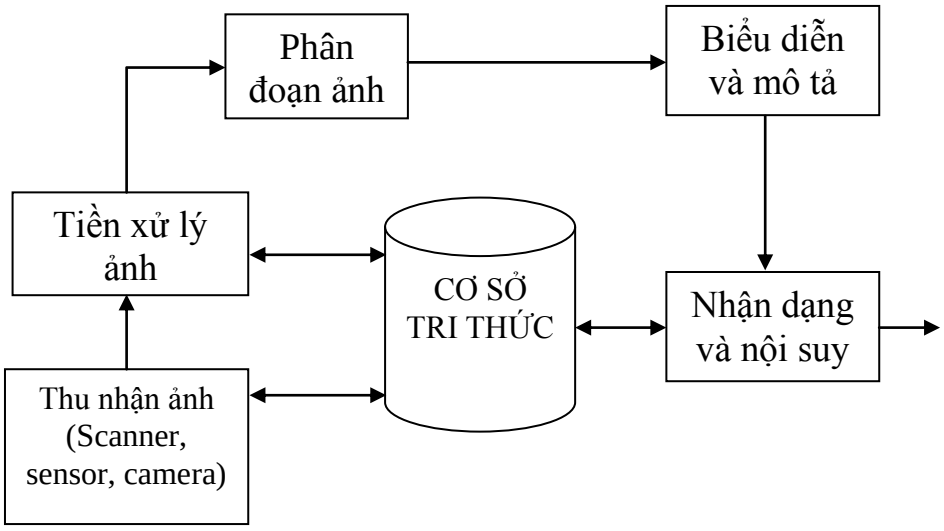
- Xử lý ảnh ban đầu để cho ra một ảnh mới tốt hơn theo một mong muốn của người dùng (ví dụ: ảnh mờ cần xử lý để được rõ hơn).

- Phân tích ảnh để thu được thông tin nào đó giúp cho việc phân loại và nhận biết ảnh (ví dụ: phân tích ảnh vân tay để trích chọn các đặc trưng vân tay).

- Từ ảnh đầu vào mà có những nhận xét, kết luận ở mức cao hơn, sâu hơn (ví dụ: ảnh một tai nạn giao thông phác họa hiện trường tai nạn).

1.1.2. Các bước cơ bản trong xử lý ảnh

Quá trình xử lý một ảnh đầu vào nhằm thu được một ảnh đầu ra mong muốn thường phải trải qua rất nhiều bước khác nhau [2]. Các bước cơ bản của một quá trình xử lý ảnh được thể hiện thông qua hình sau:



Hình 1.2: Các bước cơ bản trong quá trình xử lý ảnh

1.1.2.1. Thu nhận ảnh

Đây là bước đầu tiên trong quá trình xử lý ảnh. Để thực hiện điều này, ta cần có bộ thu ảnh và khả năng số hoá những tín hiệu liên tục được sinh ra bởi bộ thu ảnh đó. Bộ thu ảnh ở đây có thể là máy chụp ảnh đơn sắc hay màu, máy quét ảnh, máy quay... Trong trường hợp bộ thu ảnh cung cấp chưa phải là dạng số hoá ta còn phải chuyển đổi hay số hoá ảnh. Quá trình chuyển đổi ADC (Analog to Digital Converter) để thu nhận dạng số hoá của ảnh. Mặc dù đây chỉ là công đoạn đầu tiên song kết quả của nó có ảnh hưởng rất nhiều đến công đoạn kế tiếp.

1.1.2.2. Tiền xử lý

Ở bước này, ảnh sẽ được cải thiện về độ tương phản, khử nhiễu, khôi phục ảnh, nắn chỉnh hình học... Với mục đích làm cho chất lượng ảnh trở lên tốt hơn nữa, chuẩn bị cho các bước xử lý phức tạp kế tiếp sau đó.

*** Khử nhiễu:** Đặc trưng của nhiễu hệ thống là tính tuần hoàn. Do vậy, có thể khử nhiễu này bằng việc sử dụng phép biến đổi Fourier và loại bỏ các đỉnh điểm. Đối với nhiễu ngẫu nhiên, trường hợp đơn giản là các vết bẩn tương ứng với các điểm sáng hay tối, có thể khử bằng phương pháp nội suy,