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

Nguyễn Cảnh Ân

PHƯƠNG PHÁP PHÁT HIỆN BẢNG TRONG TÀI LIỆU TỔNG HỢP

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

THÁI NGUYÊN- 2014

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

Nguyễn Cảnh Ân

PHƯƠNG PHÁP PHÁT HIỆN BẢNG TRONG TÀI LIỆU TỔNG HỢP

Chuyên ngành : Khoa học máy tính
Mã số: 60 48 01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS Ngô Quốc Tạo

THÁI NGUYÊN- 2014

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ i

LỜI CẢM ƠN iii

MỞ ĐẦU 1

CHƯƠNG 1: HỆ PHÂN TÍCH TRANG TÀI LIỆU ẢNH VÀ BÀI TOÁN PHÁT HIỆN BẢNG 4

1.1. Giới thiệu chung hệ phân tích trang tài liệu và bài toán phát hiện bảng 4

1.1.1. Quá trình thu nhận ảnh 7

1.1.2. Các bước xử lý điểm ảnh 8

1.1.2.1. Phương pháp nhị phân 8

1.1.2.2. Giảm nhiễu 10

1.1.2.3. Phân đoạn 11

1.1.2.4. Làm mảnh và xác định vùng 11

1.1.2.5. Mã hóa CC và véctơ hóa 13

1.1.3. Phân tích các đặc trưng của tài liệu ảnh 14

1.1.4. Phân tích các đối tượng văn bản trong tài liệu 15

1.1.4.1. Ước lượng độ nghiêng của văn bản 15

1.1.4.2. Phân tích sơ đồ trình bày của trang tài liệu 17

1.1.5. Nhận dạng ký tự quang học (OCR) 19

1.1.5.1. Trích chọn đặc trưng 21

1.1.5.2. Phân loại 22

1.1.5.3. Nhận dạng ký tự dựa trên ngữ cảnh 25

1.2. Bài toán phát hiện bảng 26

1.2.1. Mô tả bài toán 27

1.2.2. Một số hướng tiếp cận 29
1.3. Kết luận chương 30
CHƯƠNG 2: PHÂN TÍCH BẢNG DỰA TRÊN T-RECS 32
2.1. Phương pháp phát hiện bảng trong tài liệu ảnh 32
2.2. Giới thiệu thuật toán T-Recs 38
2.2.1. Các bước khởi tạo và phân đoạn của thuật toán 41
2.2.2. Trường hợp xác định sai cột của thuật toán 43
2.2.3. Cải tiến một số bước của thuật toán 44
2.2.4. Những ưu điểm của thuật toán 48
2.2.5. Những mặt hạn chế của thuật toán khởi tạo 49
2.3. Xử lý khối sau khi phân đoạn 51
2.3.1. Trộn các khối phân đoạn sai 51
2.3.2. Phân tách các cột bị trộn vào một khối 52
2.3.3. Nhóm các từ bị phân tách 55
2.4. Phân tích khối 56
2.5. Phát hiện cấu trúc các cột, hàng 57
2.6. Kết luận chương 58
CHƯƠNG 3: CHƯƠNG TRÌNH DEMO CỦA THUẬT TOÁN 59
3.1. Giới thiệu chung 59
3.2. Mô tả chương trình 60
3.3. Một số kết quả thử nghiệm 61
KẾT LUẬN 63
DANH MỤC CÁC TÀI LIỆU THAM KHẢO 66

DANH MỤC CÁC HÌNH VẼ

Hình 1.1 Sơ đồ khối của việc xử lý tài liệu

Hình 1.2 Các bước xử lý cho một hệ phân tích tài liệu, đi kèm sơ đồ là một thí dụ với các kết quả thu được từ từng bước

Hình 1.3 Phương pháp nhị phân ảnh. (a) Histogram của ảnh đa cấp xám nguyên bản. Trục ngang biểu diễn các giá trị ngưỡng được chọn. Ảnh sau khi được nhị phân: (b) sử dụng ngưỡng thấp, (c) ngưỡng hợp lý, (d) ngưỡng quá cao

Hình 1.4 Ảnh nguyên bản bên trái và ảnh sau khi làm mảnh bên phải. (a) Ký tự "m". (b) Một sơ đồ. (c) Vân tay……

Hình 1.5 Cửa sổ 3x3 điểm ảnh với điểm ảnh X nằm ở tâm. Các giá trị số biểu diễn cho hướng mà một điểm láng giềng của X thuộc: 0 (hướng tây), 1(tây - bắc), 2(bắc), 3(đông - bắc), 4(đông), 5(đông – nam), 6(nam), 7(tây – nam)

Hình 1.6 Văn bản bị nghiêng khi quét

Hình 1.7 Biểu đồ Histogram của phép chiếu ngang và dọc của ảnh (a) và (b)

Hình 1.8 Kết quả phân tích cấu trúc và chức năng các khối

Hình 1.9 Để phân tách và nhận dạng hai số 4,2 có các nét nối liền nhau như trên dễ gây nhầm lẫn

Hình 1.10 Các ký tự viết bằng tay sẽ rất dễ nhầm lẫn……..

Hình 1.11 Các cấu trúc đặc trưng nét, tính lõm, lỗ hổng, các điểm cắt ngang và kết thúc có thể được sử dụng làm các chiều của không gian đặc trưng để phân loại ký tự

Hình 1.12 Các đặc trưng của ảnh ký tự được trích ra

Hình 1.13 Một số nhầm lẫn giữa bảng và đối tượng khác

Hình 1.14 Khái niệm các thành phần trong bảng

Hình 2.1 Một số lỗi phổ biến của các thuật toán phát hiện cấu trúc bảng

Hình 2.2 Thuật toán phát hiện bảng dựa Tab-stop

Hình 2.3 Các từ láng giềng của từ "consist" theo chiều dọc

Hình 2.4 Thuật toán phân đoạn khởi tạo đối với một đoạn văn bản

Hình 2.5 Trường hợp thuật toán nhận dạng sai cột

Hình 2.6 Trường hợp giữa các dòng của một cột trong bảng có ô trắng

Hình 2.7 Mô tả kết quả thuật toán đã được điều chỉnh nhận dạng khối

Hình 2.8 Kết quả nhận dạng các cột từ hình 2.5

Hình 2.9 Mô tả quá trình phân khối của văn bản trong các cột có khoảng cách rất hẹp

Hình 2.10 Trường hợp một ô của bảng chiếm nhiều dòng dữ liệu

Hình 2.11 Những mặt hạn chế của thuật toán

Hình 2.12 Trộn hai khối bị phân tách

Hình 2.13 (a):Tách các cột nhỏ trong cột lớn;(b):Trộn các khối nhỏ vào khối lớn

Hình 2.14 Trộn các từ bị tách nhờ vào các đoạn thẳng canh lề

Hình 2.15 (a) Phân tích khối loại 1 thành cấu trúc các ô của bảng ; (b) Ô khối loại 2 được phân tich nhờ vào ô khối loại 1

Hình 2.16 Tách các khối loại 2 thành các hàng trong bảng

Hình 3.1 Giao diện chương trình thử nghiệm

Hình 3.2 Kết quả nhận dạng khối của chương trình

Hình 3.3 Trường hợp nhận dạng có môi trường bảng

Hình 3.4 Nhận dạng ra các cột, các khối văn bản

LỜI CẢM ƠN

Trong suốt thời gian làm luận văn vừa qua, dưới sự giúp đỡ và chỉ bảo nhiệt tình của PGS.TS Ngô Quốc Tạo – Viện Công nghệ Thông tin – Viện Khoa học và công nghệ Việt Nam, luận văn của em đã được hoàn thành. Mặc dù bản thân đã cố gắng không ngừng cùng với sự tận tâm của thầy hướng dẫn song do thời gian và khả năng cũng còn nhiều hạn chế nên luận văn cũng không tránh khỏi những thiếu sót trong quá trình làm.

Để hoàn thành xong luận văn này, em xin bày tỏ lòng biết ơn sâu sắc tới PGS.TS Ngô Quốc Tạo – người thầy đã tận tình hướng dẫn em trong quá trình tìm hiểu, xây dựng và phát triển luận văn này.

Em xin chân thành cảm ơn các thầy cô giáo trong Ban giám hiệu, phòng Đào tạo, các thầy cô giáo của trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên cùng các thầy cô giáo trong Viện Công nghệ Thông Tin – Viện Khoa học và Công nghệ Việt Nam đã quan tâm, tạo điều kiện thuận lợi, nhiệt tình giảng dạy và hướng dẫn em trong suốt hai năm học qua. Và cuối cùng tôi xin gửi lời cảm ơn đến gia đình, cơ quan và toàn thể học viên lớp K11I Ninh Bình đã quan tâm, động viên và giúp đỡ tôi trong suốt hai năm học vừa qua.

Cuối cùng em rất mong nhận được sự chỉ dẫn, góp ý của các thầy cô giáo để luận văn của em được hoàn thiện hơn.

Em xin trân trọng cảm ơn !

MỞ ĐẦU

Trong những năm gần đây, các thiết bị phần cứng máy tính phục vụ cho công việc lưu trữ và xử lý hình ảnh đã phát triển vượt bậc cả về dung lượng lẫn tốc độ xử lý. Đồng thời, giá cả của các thiết bị này cũng đã giảm đến mức con người trên toàn thế giới dễ dàng sở hữu những thiết bị liên quan đến việc phân tích và xử lý hình ảnh.

Cùng với sự phát triển đó có những thách thức đặt ra đối với các nhà khoa học máy tính. Các loại tài liệu lưu trữ trên giấy và xử lý theo các cách thức cũ không theo kịp tốc độ phát triển của công nghệ. Những công việc ngày nay liên quan đến các loại tài liệu không chỉ là các tài liệu chữ chỉ để lưu trữ mà tài liệu bao gồm nhiều thành phần như các bảng biểu, ảnh…với số lượng khổng lồ tài liệu và xử lý những nhiệm vụ phức tạp trên máy tính ngày càng nhiều. Những công việc văn phòng hàng ngày đều liên quan đến tài liệu, một tài liệu không chỉ đơn giản được lưu trữ mà nó cần phải được xử lý để có khả năng thay đổi, soạn thảo, chỉnh sửa và trích chọn các thông tin quan trọng. Vì thế các hệ phân tích tài liệu ra đời, mục đích của chúng là giúp biểu diễn thông tin trong các tài liệu ảnh, tài liệu giấy được đưa vào từ máy quét dưới dạng có cấu trúc.

Lĩnh vực xử lý ảnh là một công việc có nhiều ứng dụng trong cuộc sống, theo đó, một số nước phát triển trên thế giới như Nhật Bản, Trung Quốc, Pháp, Mỹ, Canada đã không ngừng nghiên cứu phát triển công nghệ phần mềm liên quan đến ngành nhận dạng và xử lý hình ảnh để khai thác triệt để lợi thế của sức mạnh phần cứng hiện có. Cùng với sự phát triển công nghệ tri thức và nhận dạng trên thế giới, Việt Nam ta cũng đang từng bước đầu tư và phát triển ngành nhận dạng và xử lý ảnh. Điển hình là sự phát triển và ứng dụng mạnh mẽ của Viện Khoa học công nghệ Việt Nam – Viện Công nghệ Thông tin Việt Nam. Tại Viện

đã có nhiều tác giả nghiên cứu và cải tiến một số thuật toán quan trọng liên quan đến việc nhận dạng và phân tách các đối tượng khác nhau trong ảnh tài liệu. Từ đó đưa ra được một số phần mềm ứng dụng thiết thực trong cuộc sống. Điển hình là sản phẩm phần mềm Hệ nhận dạng quang học OCR, hay hệ nhận dạng các chuỗi văn bản, bảng biểu VnDOCR.

Nhiều thuật toán ra đời và từng bước phát triển đã phục vụ đắc lực cho việc đưa ra các ứng dụng khả thi vào cuộc sống cũng như góp phần xây dựng và bổ sung kho tri thức khoa học công nghệ của thế giới. Điển hình về thuật toán nhận dạng đối tượng trong ảnh tài liệu là thuật toán nhận dạng bảng theo phương pháp tiếp cận dưới lên (bottom-up) được đề xuất bởi tác giả Thomas G.Kieninger được đặt tên là T-Recs.

Phát hiện bảng và ảnh trong tài liệu ảnh là những bài toán khó và phức tạp. Trước đây các hệ phân tích tài liệu ảnh chỉ tập trung vào nhận dạng các chuỗi ký tự, phân đoạn các khối văn bản. Ngày nay tài liệu không chỉ đơn thuần là văn bản mà nó còn bao gồm hỗn hợp những đối tượng các chuỗi ký tự, ảnh, các hình vẽ, sơ đồ, các bảng biểu .v.v..

Một số yếu tố cấu thành nên bảng biểu (*structure of table*) đó là các ô (*cells*), các dòng (*rows*) và các cột (*columns*). Phát hiện bảng là bài toán phát hiện ra các cột, các dòng, các ô của bảng biểu. Việc phân tích cấu trúc của ảnh tài liệu có vai trò quan trọng rằng khi máy tính định hình được cấu trúc của ảnh thì sẽ giúp ích cho việc phục vụ mang tính chất đầu cuối cho những công đoạn xử lý khác, cũng như kết hợp xử lý tự động các dữ liệu thu thập được. Do đó, khi đã phát hiện được một đối tượng (*văn bản hay hình ảnh*) thì việc phát hiện luôn cả cấu trúc chứa đựng và liên quan với đối tượng đó là thật sự cần thiết. Một trong những cấu trúc quan trọng phổ biến thường được sử dụng mà trong luận văn quan tâm đề cập đến đó là việc phát hiện bảng biểu (*detect table*) trong ảnh tài liệu

Trong phạm vi một đề tài luận văn thạc sĩ với chủ đề "**Phương pháp phát hiện bảng trong tài liệu tổng hợp"** tôi sẽ tìm hiểu một số phương pháp, kỹ thuật phát hiện bảng trong tài liệu tổng hợp, đưa ra giải pháp cải tiến thuật toán, hướng phát triển của thuật toán, xây dựng chương trình thử nghiệm.

Bố cục của luận văn ngoài phần mở đầu và phần kết luận bao gồm 3 chương. Chương 1 trình bày ngắn gọn cấu trúc chung của một hệ phân tích tài liệu ảnh, bao gồm các thành phần chính như: lấy dữ liệu, xử lý điểm ảnh, trích chọn đặc trưng... và giới thiệu bài toán phát hiện bảng

Chương 2 đưa ra một thuật toán phát hiện bảng theo phương pháp tiếp cận dưới – lên (bottom – up). Thuật toán được đề xuất bởi Thomas G .Kieninger (1998) được đặt tên là T-Recs. Tuy nhiên để phát hiện được chính xác các cấu trúc bảng thì thuật toán còn nhiều hạn chế. Luận văn sẽ chỉ ra trường hợp thuật toán phát hiện sai và đưa ra giải pháp nâng cao độ chính xác khi phát hiện.

Cuối cùng chương 3 trình bày chương trình thử nghiệm: ***Nhận dạng bảng theo cấu trúc*** dùng để nhận dạng bảng trong trang tài liệu tổng hợp.

Phần kết luận nêu tóm tắt lại các vấn đề được đưa ra trong luận văn và đưa ra những vấn đề còn tồn tại để nâng cao tính hiệu quả của những thuật toán. Các hướng giải quyết và nghiên cứu trong tương lai đối với những phương pháp này cũng sẽ được đưa ra.