

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN
----- ✧ -----

Nguyễn Thị Hiếu

**TÌM HIỂU PHƯƠNG PHÁP PHÂN TÍCH BẰNG
BÊN TRONG TÀI LIỆU ẢNH**

Luận văn Thạc sỹ Công nghệ thông tin

Thái Nguyên, tháng 11 năm 2009

ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN
----- ✧ -----

Nguyễn Thị Hiếu

**TÌM HIỂU PHƯƠNG PHÁP PHÂN TÍCH BẰNG
BÊN TRONG TÀI LIỆU ẢNH**

Luận văn Thạc sỹ: Công nghệ thông tin

Chuyên ngành: Khoa học máy tính

Mã số: 604801

Người hướng dẫn Khoa học:

PGS.TS Ngô Quốc Tạo

Thái Nguyên, tháng 11 năm 2009

MỤC LỤC

Trang phụ bìa	
Lời cảm ơn	
MỤC LỤC	i
THUẬT NGỮ TIẾNG ANH:	iii
DANH MỤC CÁC HÌNH VẼ	iv
CHƯƠNG I: MỞ ĐẦU	1
1.1. Cơ sở nghiên cứu và mục đích của luận văn	3
1.2. Tổ chức của luận văn:	4
CHƯƠNG II: TỔNG QUAN VỀ HỆ PHÂN TÍCH TÀI LIỆU ẢNH	5
2.1. Tài liệu ảnh	5
2.2. Hệ phân tích trang tài liệu	5
2.3. Thu thập dữ liệu ảnh	6
2.4. Tiền xử lý điểm ảnh	9
2.4.1. Xử lý nhị phân	10
2.4.2. Giảm nhiễu	11
2.4.3. Phân đoạn ảnh	12
2.4.4. Làm mảnh và xác định vùng	12
2.4.5. Mã hóa CC và vectơ hóa	13
2.5. Phân tích đặc trưng của tài liệu ảnh	15
2.6. Phân tích đối tượng văn bản trong tài liệu	15
2.6.1. Xác định góc nghiêng của văn bản	16
2.6.2. Phân tích bố cục của trang tài liệu ảnh	18
2.7. Nhận dạng ký tự quang học (OCR)	19
2.7.1. Thuật toán OCR	20
2.7.1.1. Trích chọn đặc trưng	20
2.7.1.2. Phân loại	21
2.7.2. Nhận dạng ký tự dựa trên ngữ cảnh	21
2.8. Phân tích các đối tượng ảnh trong tài liệu	22

CHƯƠNG 3: THUẬT TOÁN TÁCH VĂN BẢN - ẢNH TỪ TRANG TÀI	
LIỆU ẢNH -----	24
3.1. Tổng quan về phân tách văn bản – ảnh -----	24
3.2. Những đặc trưng chung của một tệp tài liệu ảnh -----	27
3.3. Thuật toán phân tách văn bản - ảnh-----	30
3.3.1. Xoá bỏ các đối tượng tuyến tính -----	31
3.3.2. Phân tích các thành phần liên thông của nét bút -----	32
3.3.3. Kết hợp các nét ký tự tạo thành các chuỗi văn bản-----	34
3.3.4. Thực hiện các phép toán hình thái -----	35
3.3.5. Phân tích các thành phần liên thông mới -----	35
3.3.6. Biểu diễn cấu trúc thông tin của các chuỗi văn bản -----	36
CHƯƠNG IV: PHƯƠNG PHÁP PHÂN TÍCH BẢNG T-RECS TRONG	
TRANG TÀI LIỆU ẢNH-----	39
4.1. Giới thiệu-----	39
4.2. Thuật toán phân đoạn khởi tạo -----	41
4.2.1. Trường hợp thuật toán nhận dạng sai cột -----	42
4.2.2. Cải tiến các bước của thuật toán phân đoạn khởi tạo T - Recs++	44
4.2.3. Những ưu điểm của thuật toán -----	46
4.2.4. Những mặt hạn chế của thuật toán khởi tạo -----	47
4.3. Các bước xử lý khối sau khi phân đoạn-----	48
4.3.1. Trộn các khối phân đoạn sai -----	48
4.3.2. Phân tách các cột bị trộn vào một khối-----	49
4.3.3. Nhóm các từ bị phân tách -----	52
4.4. Phân tích khối-----	53
4.5. Xác định cấu trúc các cột, hàng-----	54
CHƯƠNG 5 CHƯƠNG TRÌNH THỬ NGHIỆM VÀ MINH HỌA THUẬT	
TOÁN T-RECS++ -----	56
5.1. Mô tả chương trình -----	56
5.2. Một số kết quả -----	58
KẾT LUẬN VÀ ĐỀ XUẤT -----	61

THUẬT NGỮ TIẾNG ANH

3 – D	3 Dimensions
CAD	Computer Aided Design
CAM	Computer Aided Manufacturing
CC	Chain Code
CCs	Connected Components
CPU	Control Processing Unit
DP	Dynamic Programming
HWRatio	Heigh Width Ratio
K – NNR	K – Nearest Neighbor Rule
LC	Linear Component
LSD	Local Stroke Density
NCCs	New Connected Components
NNR	Nearest Neighbour Rule
OCR	Optical Character Recognition
T-Recs	Table Recognition System
WBRatio	White Black Ratio
WDG	White-space Density Graphs

DANH MỤC CÁC HÌNH VẼ

Hình 2.1	Sơ đồ quá trình xử lý tài liệu
Hình 2.2	Sơ đồ quá trình phân tích trang tài liệu
Hình 2.3	Phương pháp nhị phân ảnh. (a) Histogram của ảnh đa cấp xám nguyên bản, (b) chọn ngưỡng thấp, (c) chọn ngưỡng hợp lý, (d) chọn ngưỡng quá cao.
Hình 2.4	Ảnh nguyên bản (Văn tay) bên trái và ảnh sau khi làm mảnh bên phải.
Hình 2.5	Tài liệu ảnh trước và sau các bước tiền xử lý. Ảnh (a) gốc, ảnh (b) ảnh sau khi chuyển về ảnh nhị phân, ảnh (c) ảnh sau khi chỉnh nghiêng, ảnh (d) ảnh sau khi lọc nhiễu.
Hình 2.6	văn bản bị nghiêng sau khi được quét qua máy quét.
Hình 2.7	Ví dụ minh họa kết quả phân tích bố cục của trang tài liệu ảnh
Hình 2.8	Chữ viết tay có thể gây nhầm lẫn
Hình 3.1	Ví dụ về các đối tượng văn bản và đối tượng ảnh
Hình 3.2	Biểu diễn các điểm ảnh giao nhau
Hình 3.3	Một số trường hợp ngoại lệ
Hình 3.4	Sơ đồ thuật toán phân tách văn bản
Hình 3.5	Hình 3.5 Nhận dạng đường kẻ nghiêng với phép toán kéo dẫn
Hình 4.1	Ví dụ minh họa tư tưởng của thuật toán khởi tạo
Hình 4.2	thuật khởi tạo đối với một đoạn văn bản
Hình 4.3	Trường hợp thuật toán nhận dạng sai cột
Hình 4.4	Trường hợp giữa các dòng của một cột trong bảng có ô trống
Hình 4.5	Mô phỏng việc thực hiện các bước đã cải tiến của thuật toán
Hình 4.7	Quá trình phân đoạn các cột của bảng
Hình 4.8	Trường hợp một ô của bảng chiếm nhiều dòng
Hình 4.9	Những mặt hạn chế của thuật toán
Hình 4.10	Trộn hai khối bị phân tách
Hình 4.11	Tách các cột bị trộn
Hình 4.12	Trộn lại các khối con bị tách
Hình 4.14	Tách các khối loại 1 thành các ô của bảng
Hình 4.15	Tách các khối loại 2 thành các hàng trong bảng
Hình 5.1	Giao diện chương trình T-Recs
Hình 5.2	Nhận dạng khối văn bản với T-Recc++
Hình 5.3	Nhận dạng tài liệu ảnh là bảng quy chế với T-Recs++

LỜI CẢM ƠN

Trong quá trình 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 Việt Nam, luận văn của tôi đã được hoàn thành. Mặc dù đã cố gắng không ngừng cùng với sự tận tâm của thầy hướng dẫn song do thời gian và khả năng vẫn còn nhiều hạn chế nên luận văn khó tránh khỏi những thiếu sót trong quá trình làm luận văn.

Để hoàn thành được luận văn này. Em xin bày tỏ lòng biết ơn sâu sắc tới PGS. TS Ngô Quốc Tạo - người thầy đã tận tình giúp đỡ em trong suốt quá trình tìm hiểu, xây dựng và phát triển luận văn này.

Em xin chân thành cảm ơn các thầy, cô giáo trong Viện Công nghệ Thông tin – Viện khoa học Việt Nam đã giảng dạy và hướng dẫn em trong suốt 2 năm học qua. Em cũng xin cảm ơn ban lãnh đạo khoa và toàn thể thầy cô giáo trong khoa Công Nghệ thông tin – Đại Học Thái Nguyên đã tạo điều kiện tốt nhất giúp em học tập và hoàn thành luận văn này. Và cuối cùng tôi cũng xin cảm ơn gia đình, các bạn trong nhóm luận văn và toàn thể các học viên lớp Cao học K 6 đã đồng viên, quan tâm và giúp đỡ tôi trong thời gian qua.

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

CHƯƠNG 1: MỞ ĐẦU

Nhận dạng và xử lý ảnh là một lĩnh vực mang tính khoa học và công nghệ. Ở Việt Nam Nhận dạng và xử lý ảnh là một ngành khoa học mới mẻ so với nhiều ngành khoa học khác nhưng tốc độ phát triển của nó rất nhanh. Sự ra đời của nó đã tạo ra các kỹ thuật quan trọng ảnh hưởng trực tiếp đến các lĩnh vực như: Tivi, truyền thông, kỹ xảo đồ hoạ...

Cùng với sự phát triển đó có những nhu cầu thực tế đặt ra thách thức các nhà khoa học máy tính càng nhiều. Những công việc, những bài toán được xử lý theo lối cổ truyền không theo kịp tốc độ phát triển của công nghệ ngày nay. Một trong những bài toán đó chính là các tài liệu được lưu trữ trên các chất liệu cổ truyền như giấy, gỗ, vải với khối lượng khổng lồ, chứa đựng rất nhiều tri thức của nhân loại nhưng lại không có độ bền vĩnh cửu, khó xử lý và lưu trữ. Một bài toán khác là ngày nay công việc văn phòng liên quan nhiều đến các tài liệu không đơn thuần là tài liệu chữ mà tài liệu có nhiều thành phần như bảng biểu, ảnh. Tất cả các tài liệu đó nếu tiếp tục lưu trữ theo phương pháp cổ truyền thì rất phức tạp và khó xử lý. Vậy làm thế nào để chuyển đổi những kho tàng tri thức trên vào máy tính để lưu trữ, xử lý dễ dàng, thuận tiện và nhanh gọn. Một lĩnh vực của khoa học nhận dạng là Phân tích tài liệu ảnh đã ra đời nhằm biểu diễn thông tin trong các tài liệu ảnh dưới dạng có cấu trúc.

Hệ phân tích và nhận dạng tài liệu ảnh có mục đích là chuyển đổi tự động những thông tin lưu trữ trong tài liệu giấy thành biểu diễn dưới dạng những cấu trúc mà có thể truy xuất, thay đổi được bằng máy tính. Quy trình xử lý của một hệ phân tích tài liệu bắt đầu bằng việc lấy dữ liệu, các tài liệu từ giấy in sẽ được quét qua máy quét để lưu trữ trong máy tính dưới dạng các tệp dữ liệu ảnh.

Một tài liệu ảnh là một cách biểu diễn trực quan của các trang tài liệu được in như một bài tạp chí, một lá thư, một trang báo, một mẫu thư hay một bản vẽ kỹ thuật, .v.v.. Một tài liệu ảnh có thể bao gồm các chuỗi ký tự, các hình vẽ, các bức ảnh, .v.v.. Bên cạnh việc chuyển toàn bộ nội dung của tài liệu sang tài liệu điện tử cũng cần phải bảo toàn cấu trúc và định dạng của tài liệu. Mục tiêu cơ bản của một hệ phân tích tài liệu ảnh hoàn chỉnh đó là chuyển một tài liệu lưu trữ bằng giấy sang dạng biểu diễn có thứ tự cấu trúc và nội dung của nó. Tài liệu được chuyển sang phải có khả năng thay đổi, soạn thảo và lưu trữ bởi vì nội dung của tài liệu có thể truy cập bởi cấu trúc của nó thay vì phải truy cập dưới dạng những mẫu ảnh. Có một số lượng lớn ứng dụng của hệ phân tích tài liệu ảnh được ứng dụng trong các lĩnh vực như: dịch vụ bưu chính, Chính phủ, chăm sóc y tế, thư viện, ...v.v.

Mục đích của luận văn là nghiên cứu kỹ thuật nhận dạng bảng và trích chọn ra đối tượng của tài liệu ảnh. Kỹ thuật “Phân tích bảng – T-Recs” là nghiên cứu chính.

Với tư tưởng chính của “Phương pháp phân tích bảng” đó là không xem xét đến bất cứ một loại đường phân cách nào để xác định cấu trúc bảng. Thay vào đó phương pháp sẽ tập trung vào việc nhận biết các từ trong cùng một khối logic (chẳng hạn các từ trong cùng một cột dữ liệu sẽ được cho vào trong cùng một khối). Phương pháp sẽ không đi tìm những đặc trưng để phân biệt hai vùng dữ liệu (hai cột) khác nhau mà tìm những đặc trưng để tìm ra các từ trong cùng một khối logic và từ đó xây dựng cấu trúc riêng theo phương pháp tiếp cận *bottom - up*.

1.1. Cơ sở nghiên cứu và mục đích của luận văn

Ảnh là một đối tượng khá phức tạp về đường nét, dung lượng điểm ảnh, độ sáng tối, môi trường để thu nhận ảnh phong phú kéo theo nhiều. Trong nhiều khâu phân tích ảnh ngoài việc đơn giản hoá các phương pháp toán học đảm bảo tiện lợi cho xử lý, người ta mong muốn bắt chước quy trình tiếp nhận và xử ảnh theo cách của con người. Trong các bước xử lý đó nhiều khâu hiện nay đã xử lý theo các phương pháp trí tuệ của con người. Những hệ thống nhận dạng cấu trúc không chỉ đơn giản là chuyển một tài liệu in thành một tài liệu điện tử mà hơn thế nữa còn là xây dựng những quá trình xử lý kết hợp chẳng hạn như: tự động chép nội dung, đánh chỉ mục và phân loại. Do đó việc quan trọng là kèm theo nội dung của tài liệu cũng phải trích chọn ra những cấu trúc đi kèm với từng nội dung đó.

Nhận dạng bảng là bài toán nhận dạng ra cấu trúc bảng có trong trang tài liệu ảnh, bao gồm việc nhận dạng các cột, các dòng và các ô có chứa dữ liệu trong bảng. Nhận dạng đối tượng ảnh là bài toán nhằm phân tách các đối tượng ảnh trong những trang tài liệu ảnh có chứa hỗn hợp các đối tượng là chuỗi ký tự và các đối tượng ảnh như: các sơ đồ, hình vẽ, bức ảnh ... v.v.

Mặc dù đã có nhiều kỹ thuật trong hệ thống nhận dạng cấu trúc. Tuy nhiên những nghiên cứu trên những vấn đề đó vẫn còn tiếp tục phát triển bởi vì chất lượng, độ chính xác, tính hiệu quả của những phương pháp được công bố trước đây vẫn còn chưa hoàn chỉnh và cần phải cải tiến chúng. Luận văn này trình bày kỹ thuật nhận dạng cấu trúc bảng bên trong tài liệu ảnh T-Recs và đề xuất một số phương pháp khắc phục hạn chế thuật toán T-Recs để hoàn thiện hơn phương pháp phân tích bảng.