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

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

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 vă n 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 N am 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 động 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 tiên 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 mẩu 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 nhiễu. 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 dụng, 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 hiện hơn phương pháp phân tích bảng.