MỤC LỤC

MỞ ĐẦU

Ngày nay khi máy tính phát triển, cùng với tốc độ và không gian lưu trữ trong máy tính đã được nâng cấp lên rất nhiều. Việc lưu trữ 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.

Một hệ phân tích và nhận dạng tài liệu 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. Rõ ràng rằng khi máy tính ra đời và phát triển đã giải quyết được nhiều vấn đề trong việc lưu trữ thông tin. Theo ước tính trên thế giới, chỉ có một số lượng nhỏ tài liệu từ những thư viện giấy khổng lồ được đưa lên mạng và vì vậy vẫn còn số lượng lớn những nguồn tri thức của nhân loại đang được lưu trữ theo cách thức cổ điển trong những thư viện mà việc bỏ ra chi phí duy trì (chủ yếu trả lương cho nhân viên) cho những nguồn tài liệu này là rất lớn. Thông tin bây giờ không nhất thiết phải lưu trữ bằng giấy, một cách lưu trữ không an toàn, không bền vững theo thời gian, thay vì đó nó được lưu trữ một cách ổn định và an toàn trong máy tính. Do đó bằng cách này hay cách khác tài liệu giấy được quét thành các tệp dữ liệu ảnh và được lưu trữ trong máy tính. Không chỉ đơn giản là vấn đề lưu trữ, các tài liệu từ giấy in được đưa vào máy tính còn cần được xử lý và trích chọn ra những thông tin quan trọng. Một tài liệu giấy in được đưa vào máy tính còn yêu cầu có khả năng soạn thảo, hiệu chỉnh và khôi phục lại. Một tệp dữ liệu cần phải chuyển được sang những định dạng khác để có khả năng soạn thảo, khi đó phải đảm bảo các thông tin được chuyển sang từ tệp dữ liệu phải không bị mất đi, không bị thiếu thông tin và cấu trúc vị trí của dữ liệu vẫn được giữ nguyên. Chẳng hạn vị trí các đoạn văn bản, tiêu đề, các bảng dữ liệu, .v.v.. phải được chuyển sang đúng theo cấu trúc thể

hiện trên tệp dữ liệu. Vì thế ngành nhận dạng hay các hệ phân tích tài liệu ảnh ra đời và phát triển để giải quyết những vấn đề trên.

Một vài sản phẩm thương mại đã có chẳng hạn như các hệ nhận dạng quang học OCR để nhận dạng các ký tự in, ký tự viết tay, các bảng biểu tuy nhiên vẫn còn cần nhiều nghiên cứu để cải thiện độ chính xác của các hệ thống này. Một số sản phẩm chẳng hạn như VnDOCR (của Việt Nam) cho phép nhận dạng các chuỗi văn bản, các bảng biểu hay Omnipage, Find Reader .v.v.. là những sản phẩm nhận dạng nổi tiếng.

Bài toán nhận dạng bảng 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.. 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. Đã có rất nhiều phương pháp, thuật toán tách bảng, tách ảnh được công bố trước đây. 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 án bao gồm 4 phần chủ yếu tập trung vào trình bày thuật toán nhận dạng bảng.

Chương 1 trình bày ngắn gọn cấu trúc chung của một hệ phân tích tài liệu ảnh, sơ lược về nhận dạng ký tự quang học (OCR).

Chương 2 đưa ra một thuật toán nhận dạng 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 nhận dạng được chính xác các cấu trúc bảng thì thuật toán còn nhiều hạn chế. Luận án sẽ chỉ ra trường hợp thuật toán nhận dạng sai và đề xuất thuật toán cải tiến T-Recs++.

Cuối cùng chương 3 trình bày chương trình thử nghiệm: ***T-Recs++*** dùng để nhận dạng bảng.

Phần kết luận nêu tóm tắt lại các vấn đề được đưa ra trong luận án 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.

CHƯƠNG 1
TỔNG QUAN HỆ PHÂN TÍCH TÀI LIỆU

1.1. Giới thiệu chung một hệ phân tích trang tài liệu

Một hệ phân tích tài liệu ảnh nói đến một hệ thống bao gồm những thuật toán và các kỹ thuật mà có thể áp dụng cho các tài liệu ảnh để lấy ra được các thông tin mà máy tính có thể đọc được và hiểu được từ các điểm dữ liệu ảnh. Một sản phẩm phân tích tài liệu ảnh mà nhiều người biết đến đó là phần mềm Nhận dạng Ký tự Quang học (OCR), phần mềm có khả năng nhận dạng các ký tự từ các loại tài liệu dưới dạng ảnh. OCR giúp người dùng có khả năng soạn thảo và tìm kiếm nội dung của tài liệu. Chương này sẽ mô tả tóm tắt các thành phần chính có trong một hệ phân tích tài liệu.

Mục đích của một hệ phân tích tài liệu là có khả năng nhận dạng ra các đối tượng văn bản, đối tượng ảnh trong tài liệu ảnh và có khả năng trích chọn ra được các thông tin mà người dùng mong muốn. Chúng ta có thể chia một hệ phân tích tài liệu thành hai phần (Hình 1). Phần thứ nhất là xử lý văn bản, liên quan đến việc xử lý các đối tượng văn bản: ký tự, chuỗi ký tự, các từ. Xử lý văn bản bao gồm các công việc sau: xác định độ nghiêng của tài liệu (độ nghiêng hay độ xiên của tài liệu ảnh do tài liệu được đặt không đúng khi thực hiện quét vào từ máy quét), tìm các cột, các đoạn văn bản, các dòng văn bản, các từ và cuối cùng là nhận dạng văn bản (có thể thêm các thuộc tính như loại phông chữ, kích thước của phông chữ) bởi phương pháp nhận dạng ký tự quang học (OCR). Phần thứ hai là xử lý các đối tượng ảnh là các đối tượng tạo ra từ các đường kẻ trong sơ đồ, các đường kẻ phân tách giữa các đoạn văn bản, các hình vẽ, các lôgô của công ty… Sau khi áp dụng các kỹ thuật phân tích ảnh và văn bản, các đối tượng cần nhận dạng trong tài liệu ảnh được trích ra và được biểu diễn dưới dạng một tài liệu định dạng khác, chẳng hạn như word, html…

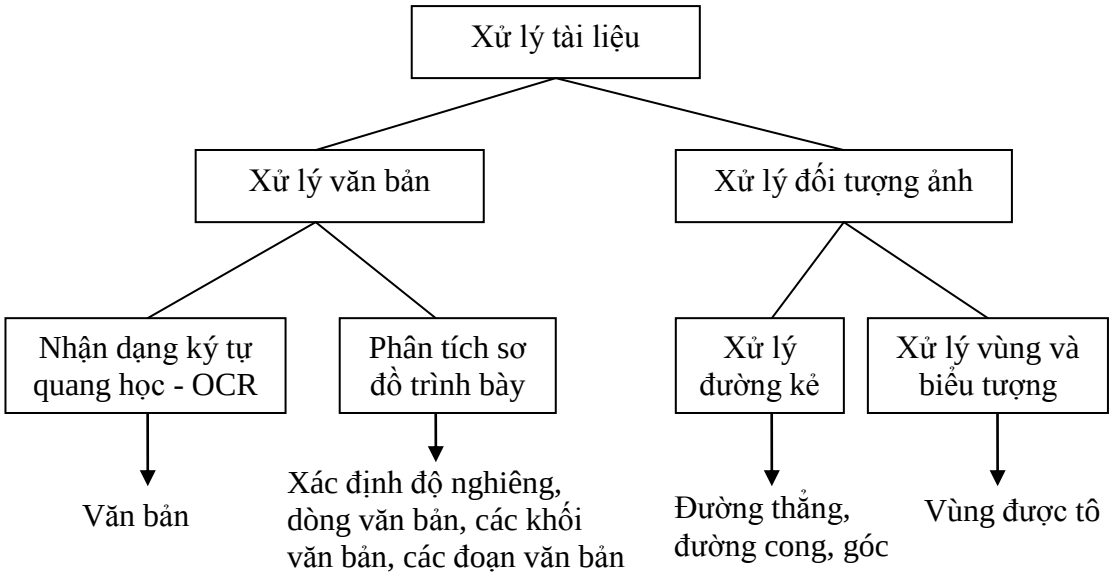
Chúng ta có thể xem xét 3 ví dụ cụ thể được chỉ ra dưới đây để thấy được sự cần thiết của việc phân tích tài liệu:

1) Phần lớn các tài liệu văn bản trong văn phòng làm việc đều được tạo ra từ máy tính, và thậm chí chúng được tạo ra bởi các máy tính, phần mềm khác

nhau, và do đó có thể định dạng của chúng là không tương thích với nhau. Chúng có thể bao gồm các định dạng văn bản, các bảng dữ liệu và cũng có thể là các văn bản viết bằng tay. Chúng có kích thước khác nhau, có thể từ một tấm thẻ kinh doanh nghiệp đến một ảnh vẽ kỹ thuật lớn. Một hệ phân tích tài liệu sẽ giúp nhận dạng các loại tài liệu, có khả năng trích chọn ra được các phần chức năng và có khả năng chuyển từ một định dạng máy tính này sang một định dạng khác.

2) Một thí dụ khác là các máy phân loại thư tự động dùng để phân loại, sắp xếp thư và nhận dạng địa chỉ thư. Những máy này đã có từ những thập kỷ trước, nhưng ngày nay yêu cầu cao hơn đó là xử lý nhiều thư hơn, nhanh hơn và yêu cầu chính độ xác cao hơn.

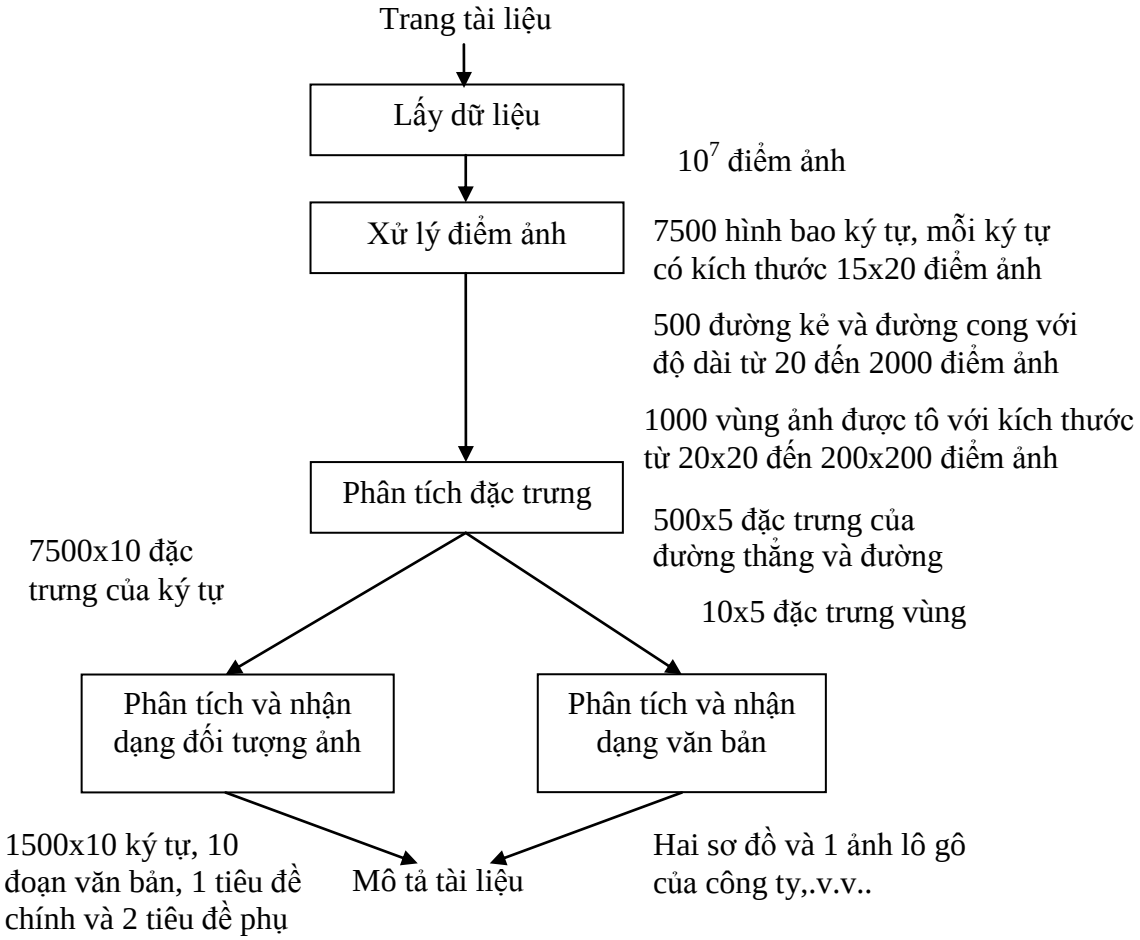
3) Hơn thế nữa trong những thư viện cổ điển trước đây (thư viện lưu trữ các loại sách báo dưới dạng giấy tờ, vẫn còn tồn tại nhiều), việc các tài liệu bị mất các thông tin, thiếu thông tin, số lượng bản sao hạn chế hay thậm chí các tài liệu bị thoái hoá theo thời gian là những vấn đề phổ biến. Vì vậy chúng cần phải được khôi phục và chỉnh sửa bằmg những kỹ thuật phân tích tài liệu. Những ví dụ trên đã tạo ra thách thức và động lực cho sự phát triển những giải pháp trong tương lai của một hệ phân tích tài liệu.



Hình 1 Sơ đồ khối của việc xử lý tài liệu

Các hệ phân tích tài liệu sẽ ngày càng phát triển và hiển nhiên là chúng sẽ có trong các hệ xử lý tài liệu. Chẳng hạn như, hệ thống OCR sẽ được sử dụng rộng rãi để lưu trữ, tìm kiếm và trích dẫn từ các tài liệu lưu trữ trên giấy. Các kỹ thuật phân tích cách bố trí trong một trang tài liệu giúp nhận dạng những biểu mẫu (form) riêng biệt, hay

định dạng của một trang tài liệu và cho phép sao lưu tài liệu đó. Các sơ đồ có thể được đưa vào từ các bức ảnh hay vẽ bằng tay và có thể thay đổi, soạn thảo lại chúng. Sử dụng máy tính có thể chuyển các tài liệu viết bằng tay thành các tài liệu điện tử được lưu trữ trong máy tính. Các tài liệu được lưu trữ trong các thư viện, các tài liệu kỹ thuật trong các công ty sẽ được chuyển đổi sang thành tài liệu điện tử nhằm nâng cao hiệu quả, thuận tiện trong việc lưu trữ và dễ dàng mang đến cơ quan hay mang về nhà. Mặc dù tài liệu sẽ ngày càng được xử lý và lưu trữ nhiều trong máy tính nhưng trên thực tế có rất nhiều các hệ thống khác nhau mà tài liệu giấy là phương tiện làm việc hiệu quả và chắc chắn rằng tài liệu giấy vẫn sẽ là phương tiện làm việc với chúng ta trong một vài thập kỷ nữa. Vấn đề ở đây là làm sao chúng ta tích hợp những tài liệu giấy vào trong máy tính xử lý.



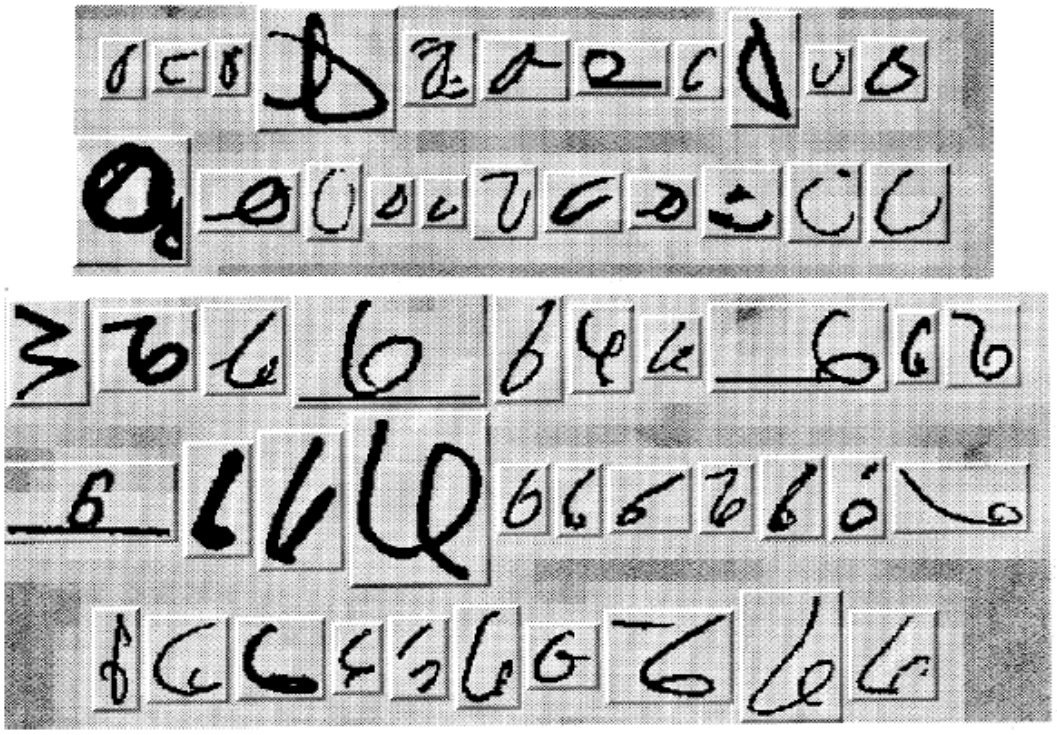
Hình 2 Các bước xử lý cho một hệ phân tích tài liệu, đi kèm sơ đồ là một thí dụ với các kết quả thu được từ từng bước.

Hình 2 minh hoạ cho các bước xử lý chung của một hệ phân tích tài liệu [3]. Sau khi dữ liệu được tạo ra, tài liệu ảnh phải trải qua các bước xử lý điểm ảnh và phân tích đặc trưng và sau đó tách ra từng phần nhận dạng văn bản và ảnh riêng rẽ.

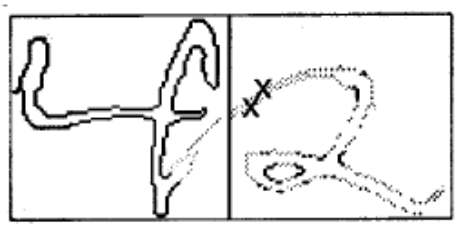
1.2. Sơ lược về nhận dạng ký tự quang học (OCR)

Nhận dạng ký tự quang học (OCR) là phần cốt lõi của ngành nhận dạng, trong đó mục đích của OCR là nhận biết được các chuỗi ký tự từ bảng chữ cái. Các ký tự trong bảng chữ cái thường có rất nhiều kiểu viết khác nhau. Trên thực tế các ký tự thường được viết bằng nhiều kiểu khác nhau tuỳ thuộc vào kích cỡ, loại phông chữ và nét bút viết tay của từng người. Mặc dù các ký tự có thể viết theo nhiều cách nhưng có lẽ rằng vẫn có những quy tắc xác định để nhận biết từng ký tự. Phát triển những thuật toán trên máy tính để nhận biết các ký tự trong bảng chữ cái là một nhiệm vụ trọng tâm của OCR. Nhưng thách thức đối với vấn đề này đó là – trong khi con người có thể nhận dạng gần như chính xác 100% các ký tự viết tay thì OCR vẫn chưa thể đạt tới được điều này.

Khó khăn đối với OCR thể hiện qua một số đặc điểm. Sự gia tăng số lượng và kích cỡ của phông chữ trong bảng chữ cái, không ràng buộc các kiểu chữ viết tay, các ký tự nối liền nhau, các nét bị đứt, các điểm nhiễu .v.v.. tất cả chúng làm cho quá trình nhận dạng gặp khó khăn. Hình 3 chỉ ra một thí dụ giữa số ‘0’ và số ‘6’ rất dễ nhầm lẫn khi chúng được viết bằng tay. Một từ cũng có thể hoàn toàn là các con số, chẳng hạn các số điện thoại, hay hoàn toàn là các ký tự trong bảng chữ cái hoặc có thể trộn lẫn giữa chữ cái và số.



Hình 3 Các ký tự viết bằng tay sẽ rất dễ nhầm lẫn



Hình 4 Sẽ không dễ dàng gì để phân tách và nhận dạng hai số 4,2 có các nét nối liền nhau như trên

Do đó quá trình nhận dạng sẽ càng trở nên khó khăn hơn khi các ký tự liền kề trong một chuỗi nối liền nét (Hình 4). Các ký tự nối liền nét là điều rất bình thường và mang ý nghĩa gắn kết (như ký tự gạch nối), khi nối một ký tự số với một ký tự chữ cái viết hoa trong một từ viết tắt thì sẽ rất khó nhận dạng.

1.3. Kết luận chương

Chương này đã mô tả ngắn gọn các thành phần chung của một hệ phân tích tài liệu ảnh và nêu sơ lược về nhận dạng ký tự quang học (ORC). Các chương tiếp theo sẽ mô tả chi tiết phương pháp nhận dạng bảng bằng thuật toán T-Recs.

CHƯƠNG 2
THUẬT TOÁN TÁCH BẢNG T-RECS

2.1. Giới thiệu

Ngày nay mục tiêu của một hệ thống nhận dạng quang học (OCR) đã tiến xa hơn rất nhiều, không chỉ là những phép chuyển đổi đơn giản một tài liệu ảnh sang một tài liệu văn bản bao gồm các từ mà hơn thế nữa nó còn tập trung vào việc xác định đúng những cấu trúc đặc trưng trong tài liệu. Trong khi một số hệ phân tích cấu trúc tập trung vào xác định tính logíc của các đối tượng trong một số miền giới hạn như nhận dạng mẫu viết thư [19], một số khác lại đi vào tập trung nhận biết một số cấu trúc phổ biến như đoạn văn bản, dòng tiêu đề hay danh sách. Hu [17] và Condit [18] đều miêu tả các hệ thống biểu diễn những cấu trúc trên.

Mục đích của 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 **Error! Reference source not found.**. 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.

Khi đề cập đến vấn đề nhận dạng cấu trúc trong các tài liệu có chứa dữ liệu bảng biểu sẽ có hai hướng tiếp cận khác nhau: cách tiếp cận thứ nhất đó là xác định chính xác cấu trúc của bảng, bao gồm các ô trong bảng, cách này thường được gọi là *phân đoạn hay nhận dạng cấu trúc.* Cách thứ hai là dựa vào hình dạng bất kỳ của các khối đã được sắp xếp và đưa tập các đối tượng trong các khối về một cấu trúc bậc cao hơn. Quá trình này được gọi tên là *gán nhãn lôgíc, phân tích cấu trúc hay phân tích sơ đồ trình bày.*

Tìm hiểu những phướng pháp nhận dạng cấu trúc bảng đã có trước đây đều cho thấy một điểm giống nhau, đó là các phương pháp này đều nhận dạng ra cấu trúc bảng bằng xác định ra các dấu hiệu phân cách, có thể là các khoảng trắng, các đường kẻ. Chẳng hạn như Rus và Summers **Error! Reference source not found.** mô tả một hệ nhận dạng cấu trúc bảng có khả năng xác định được bảng mà các cột cách nhau một khoảng hẹp sử dụng WDG. Trong khi đó một số phương pháp khác lại dựa vào độ

rộng thích hợp của khoảng trắng giữa hai cột để nhận dạng **Error! Reference source not found.**.

Một số phương pháp khác xác định cấu trúc của bảng bằng quy tắc các đường kẻ. Một trong số đó là mô tả của Green và Krishnamoorthy **Error! Reference source not found.**, các ông đã áp dụng phân tích vị trí của các đường kẻ để đưa ra cấu trúc của bảng, hay Itonori **Error! Reference source not found.** chỉ quan tâm đến khía cạnh các nhãn và các khối sau khi phân đoạn làm dữ liệu đầu vào, hay Hirayama **Error! Reference source not found.** sử dụng phương pháp *DP matching.* Còn Chandran và Kasturi thì xem xét cả hai (quy tắc các đường kẻ và các khoảng trắng) để xác định cấu trúc của bảng.

Tư tưởng cốt lõi trong phương pháp sẽ trình bày dưới đây đó là không xem xét đến bất cứ một loại đường phân cách nào để xác định bảng mà sẽ đi vào 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). Chúng ta 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 *dưới lên (bottom - up).*

Một điều dễ nhận thấy ngay từ phương pháp này đó là chúng ta sẽ không phụ thuộc vào kiểu của đường thẳng được vẽ trong bảng nếu có hay là các khoảng trắng đủ rộng giữa các khối để nhận dạng cấu trúc của bảng.

Đầu vào của thuật toán là tập hợp các hình bao chữ nhật của các từ trong một đoạn văn bản. Đầu ra là các cột, các dòng, các ô của bảng nếu tồn tại môi trường bảng trong đoạn văn bản. Thuật toán sẽ cần các bước tiền xử lý như nhận dạng các dòng văn bản của trang tài liệu, hình bao chữ nhật các từ trên từng dòng văn bản và nhận dạng các đoạn văn bản khác nhau. Từ đó có nhận dạng môi trường bảng trên từng đoạn văn bản của trang tài liệu.

Chương này sẽ mô tả toàn bộ chức năng của thuật toán T-Recs, phần đầu mô tả thuật toán phân đoạn khởi tạo - phần cốt yếu. Đầu tiên luận án sẽ trình bày thuật toán phân đoạn khởi tạo do Thomas G. Kieninger [15] đề xuất và sau đó chỉ ra những trường hợp mà thuật toán phân đoạn do G. Kieninger sẽ nhận dạng sai. Tiếp theo luận án sẽ trình bày thuật toán phân đoạn cải tiến (T-Recs++) để có thể nhận dạng chính xác các cột dữ liệu tồn tại trong một bảng.

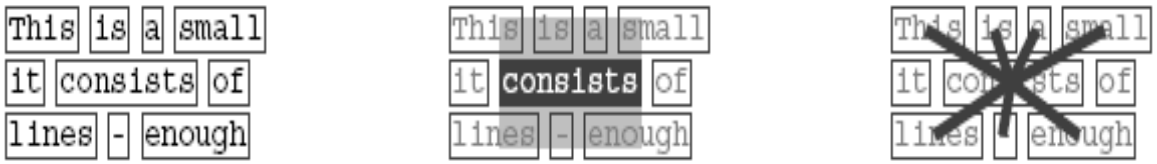
Những ưu điểm và hạn chế của thuật toán cũng được chỉ ra trong phần đầu của chương. Phần tiếp theo trong chương này luận án sẽ chỉ ra một số bước xử lý sau khi phân đoạn (postprocessing) để khắc phục những hạn chế của thuật toán phân đoạn khởi tạo.

Phần cuối của chương luận án mô tả việc phân tích các cột được nhận dạng thành các dòng và các ô trong bảng để đưa ra được cấu trúc chính xác của bảng.

2.2. Thuật toán phân đoạn khởi tạo

Tư tưởng cốt lõi của toàn bộ hệ thống chính là phần phân đoạn khởi tạo, có thể coi như là phân cụm các từ. Trong khi các phương pháp tiếp cận *dưới-lên* khác thường xác định các đường kẻ từ các từ liền kề theo chiều ngang và các khối từ các đường liền kề theo chiều dọc (chẳng hạn như Gorman [10] sử dụng những từ láng giềng gần nhất để phân cụm các từ), hệ thống sẽ trực tiếp đánh giá các cấu trúc khối văn bản từ việc phân đoạn các từ.

Vì vậy chúng ta sẽ lấy một từ bất kỳ làm nhân để xây dựng một khối mới. Nhìn trên Hình 5 (ở giữa), ta vẽ một vùng mờ ảo bao quoanh hình chữ nhật bao của từ (consist). Vùng mờ ảo này có độ rộng bằng với độ rộng của hình bao của từ và chiều dọc mở rộng đến các dòng liền kề với từ đó. Tất cả các từ mà có hình bao gối lên vùng mờ ảo của từ làm nhân sẽ nằm trong cùng một khối với từ đó. Do đó một khối bao gồm tất cả các từ được liên kết với nhau (hình bên phải của Hình 5).



Hình 5 Các từ láng giềng của từ *"consist"* theo chiều dọc

Thủ tục trên sẽ được mở rộng bằng cách thực hiện đệ quy cho tất cả các từ cho đến khi không tìm thấy có từ nào mới mà không nằm trong một khối nào đó. Đầu vào của thủ tục là hình bao chữ nhật của các từ, đầu ra là các khối lôgíc và các từ thuộc từng khối lôgíc.

Các bước thực hiện của thủ tục như sau:

1) Tìm một từ bất kỳ nào đó W_x mà chưa được đánh dấu là mở rộng (expanded).

2) Tạo một khối mới B_i

3) *Đánh dấu Wx là đã mở rộng và thêm W_x vào B_i*

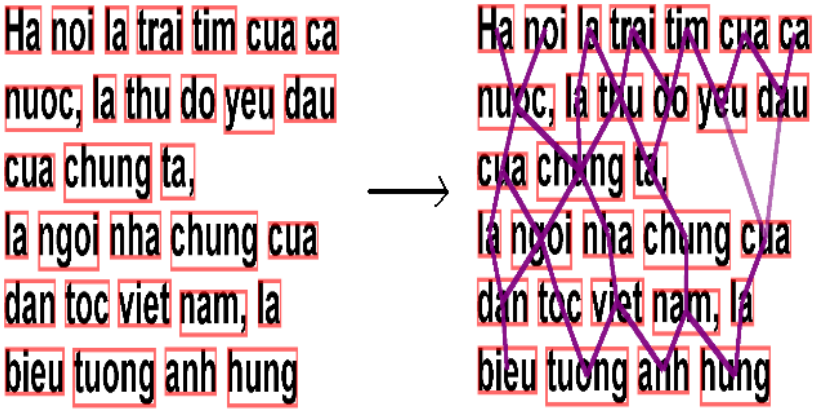
4) *Tìm tất cả các từ W_j theo chiều ngang ở dòng trước và dòng kế tiếp, sao cho W_j nằm chồng lên W_x (có nghĩa là W_j gối lên vùng mờ ảo của W_x).*

5) *Thực hiện đệ quy các bước 3, 4, và 5 cho các từ W_j vừa tìm được.*

6) *Nếu không tìm được từ nào mà chưa đánh dấu và không nằm chồng lên nhau (theo ý nghĩa của bước 4) thì tăng i lên một và quay trở lại bước 1.*

7) *Dừng thủ tục lại nếu không tìm thấy từ nào chưa được đánh dấu trong tài liệu.*

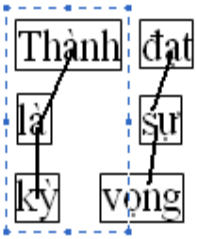
Hình 6 mô tả kết quả của thuật toán sau khi mở rộng tất cả các từ trong khối.



Hình 6 Thuật toán phân đoạn khởi tạo đối với một đoạn văn bản

2.2.1. Trường hợp thuật toán nhận dạng sai cột

Bảy bước trong thuật toán phân đoạn khối phía trên về cơ bản nhận dạng được các khối riêng rẽ nhưng cũng chưa đủ tốt để nhận dạng được tất cả các loại khối phân tách. Hình 7 mô phỏng một thí dụ về trường hợp thuật toán phân tách thành hai khối khác nhau nhưng về logíc hai khối trên thực chất là một khối.



Hình 7 Trường hợp thuật toán nhận dạng sai cột

Khi phân tích các bước của thuật toán trên ta thấy có một hạn chế, đó là khi một từ W_j mới được xem xét có thêm vào khối đang duyệt B_i hay không thì thuật toán chỉ

quan tâm xem W_j có nằm chồng lên từ W_x (là từ ở dòng trước hay dòng sau của W_j) mà không xem xét W_j có nằm chồng lên bất kỳ từ nào thuộc khối B_i hay không.

Nhìn trên Hình 7, nếu thực hiện lần lượt các bước từ 1 đến 7 thì ta thấy các từ trên được chia thành hai khối riêng rẽ, nhưng ta thấy hai từ ***Thành*** và ***vọng*** tuy nằm chồng lên nhau nhưng lại thuộc hai khối khác nhau bởi vì khi thuật toán đi đến từ ***là*** nó sẽ xem xét hai từ là ***kỳ*** và ***vọng*** trong đó chỉ có mỗi từ ***kỳ*** là nằm chồng lên nó còn từ ***vọng*** không nằm chồng lên từ ***là***

Hình 8 Trường hợp giữa các dòng của một cột trong bảng có ô trắng

Hình 88 chỉ ra một thí dụ mà thuật toán do G. Kieninger có thể nhận dạng được các cột trong bảng. Trong 7 bước mà G. Kieninger đề xuất, khi thực hiện xuất phát từ một hình bao chữ nhật của một từ thuật toán chỉ tìm các từ có nằm chồng lên nó trong dòng trước và dòng kế tiếp. Vì vậy trong trường hợp một cột trong bảng mà có nhiều dòng để trống (chẳng hạn khi một ô của bảng kéo dài trên nhiều dòng) thì khi thực hiện tìm các từ ở dòng kế tiếp và dòng trước sẽ không tìm được từ nào thuộc cột đó. Do đó để tìm được chính xác các từ thuộc một cột của bảng thì xuất phát từ một từ phải tìm trên tất cả các dòng của đoạn văn bản.

Dưới đây sẽ trình bày những cải tiến các bước của thuật toán phân đoạn trên.

2.2.2. Cải tiến các bước của thuật toán phân đoạn khởi tạo - T-Recs++

Do các cột của một bảng đều nằm ở các vị trí là những khoảng khác nhau theo chiều ngang, vì vậy để cải tiến thuật toán ta sẽ đi xác định toạ độ *nhỏ nhất* - X_{min} và *lớn nhất* - X_{max} theo chiều ngang của một khối. Khi duyệt qua các từ cần thêm vào khối nếu như toạ độ nhỏ nhất và lớn nhất theo chiều ngang của khối có giao với khoảng (X_{min}, X_{max}) thì ta sẽ thêm từ đó vào khối và cập nhật lại toạ độ X_{min}, X_{max} của khối đó.

Đầu vào của thủ tục là hình bao chữ nhật của các từ, đầu ra là các khối lôgíc và các từ thuộc từng khối lôgic.

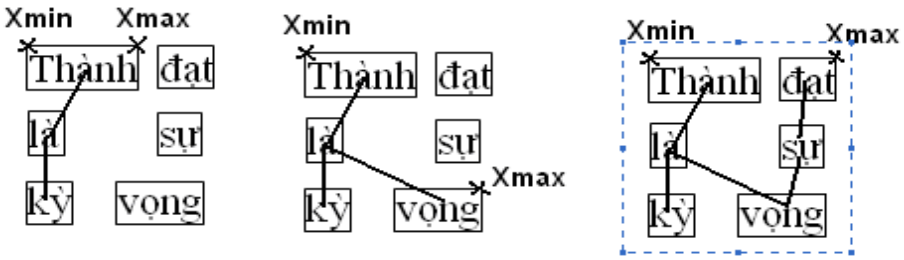
Các bước cải tiến của thuật toán phân đoạn khởi tạo sẽ gồm 8 bước như sau:

1) *Gán X_{min}= -1 và X_{max} = 0.*
2) *Tìm một từ bất kỳ nào đó W_x mà chưa được đánh dấu là mở rộng (expanded). Tính các toạ độ X_{Xmin}, X_{Xmax} lần lượt là 2 toạ độ nhỏ nhất và lớn nhất theo chiều ngang của hình bao của từ W_x.*
3) *Tạo một khối mới B_i*
4) *Đánh dấu Wx là đã mở rộng và thêm W_x vào B_i. Xét:*
   - *Nếu X_{min} = -1 thì gán X_{min}= X_{Xmin}.*
   - *Nếu X_{min} > X_{Xmin} thì gán X_{min}= X_{Xmin}.*
   - *Nếu X_{max} < X_{Xmax} thì gán X_{max} = X_{Xmax}.*
5) *Tìm tất cả các từ W_j nằm theo chiều ngang ở các dòng trước và những dòng kế tiếp (thuộc đoạn văn bản), sao cho:*

   $$(X_{min}, X_{max}) \cap (X_{Jmin}, X_{Jmax}) \neq \Phi$$

   Trong đó các toạ độ X_{Jmin}, X_{Jmax} lần lượt là 2 toạ độ nhỏ nhất và lớn nhất theo chiều ngang của hình bao của từ W_j.
6) *Thực hiện đệ quy các bước 4, 5, và 6 cho các từ W_j vừa tìm được.*
7) *Nếu không tìm được từ nào mà chưa đánh dấu và không thoả mãn điều kiện 5 thì tăng i lên một và quay trở lại bước 1.*
8) *Dừng thuật toán lại nếu không tìm thấy từ nào mà chưa được đánh dấu là mở rộng trong tài liệu.*

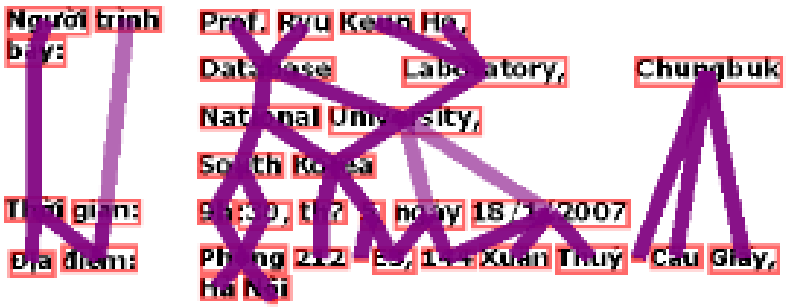
Hình 9 dưới đây mô tả các bước thuật toán phân đoạn đã cải tiến. Nếu như trên Hình 7, thuật toán trước có thể phân tách các từ vào hai khối riêng rẽ thì với các bước đã cải tiến trên thuật toán sẽ nhóm các từ trong Hình 7 vào thành một khối duy nhất (hình cuối bên phải của Hình 9).



Hình 9 Mô phỏng việc thực hiện các bước đã cải tiến của thuật toán

Trong bước thứ 5 của thuật toán, khi thực hiện tìm những từ thoả mãn để đưa vào một khối, thuật toán sẽ tìm tất cả các từ ở các dòng trước và các dòng kế tiếp chứ

không phải chỉ tìm ở dòng trước và dòng kế tiếp của dòng đang xét. Do đó việc nhận dạng đúng các cột của bảng từ Hình 8 được minh hoạ trên Hình 10.



Hình 10 Kết quả nhận dạng các cột từ Hình 8

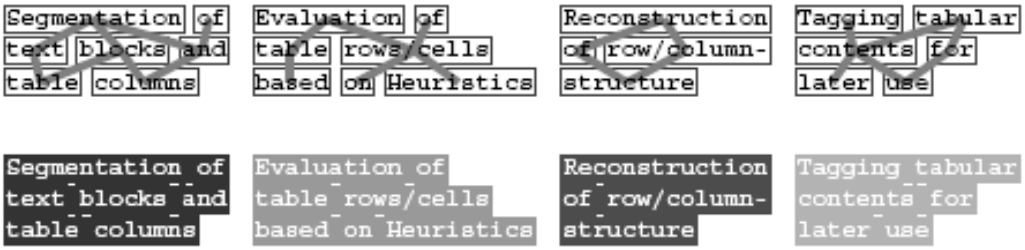
2.2.3. Những ưu điểm của thuật toán

Trong thí dụ đưa ra ở trên, điểm nổi bật của thuật toán vẫn chưa thể hiện rõ ràng vì sự phân đoạn của những khối văn bản dường như cũng giống những phương pháp có trước đây. Hình 11 minh hoạ điểm nổi bật của thuật toán khi nhận dạng cấu trúc của bảng: ở đây ta thấy mỗi khối trong hình cách nhau một khoảng cách hẹp. Do không có một từ nào nằm giữa các cột vì vậy mà các cột được phân biệt với nhau một cách rõ ràng. (Để quan sát dễ dàng hơn, mỗi cột đều được bôi một màu khác nhau để nổi bật). Ngoài những điểm mạnh đề cập trên, thuật toán còn có những đặc điểm sau:

Hình 11 Quá trình phân đoạn các cột của bảng

- Không quan tâm đến nội dung văn bản. Do đó nó có thể áp dụng cho một tài liệu kém chất lượng để thực hiện phân đoạn.
- Cho phép nhận dạng ra các cột trong bảng trong trường hợp khoảng cách giữa các cột hẹp.
- Nhận dạng cấu trúc của bảng mà không cần thông tin về tiêu đề của bảng.
- Nhận dạng cấu trúc bảng với các ô có nhiều hơn một dòng dữ liệu (Hình 12).

- Thuật toán áp dụng với các loại tài liệu phổ biến (không hạn chế một số loại bảng nào đó; không quy định luật cụ thể, không cần phải có giai đoạn học nhận dạng).



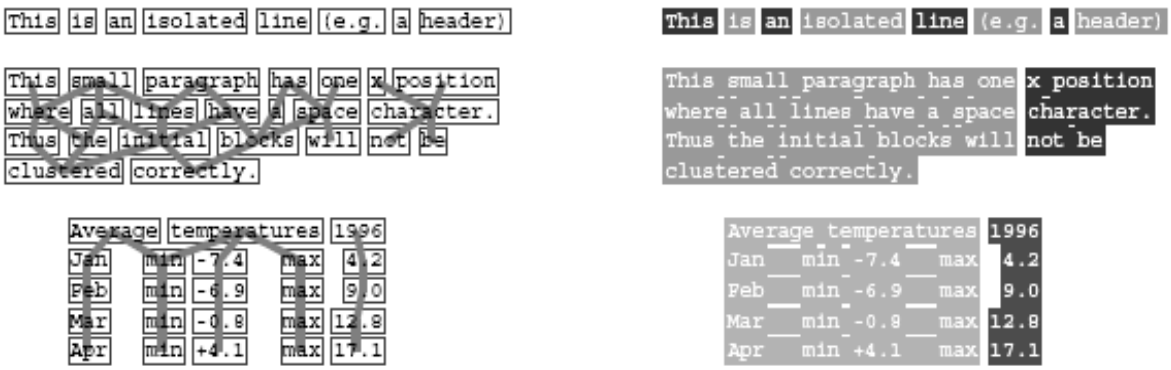
Hình 12 Trường hợp một ô của bảng chiếm nhiều dòng

2.2.4. Những mặt hạn chế của thuật toán khởi tạo

Thuật toán phân đoạn khởi tạo cũng tồn tài một số mặt hạn chế vốn có. Chẳng hạn như thuật toán sẽ coi một dòng đơn là bảng bởi vì dòng này không có những dòng là láng giềng của nó theo chiều dọc. Do đó nó sẽ coi đó là một bảng chỉ có một dòng dữ liệu trong đó mỗi một từ coi như là một cột trong bảng. Do đó khi nhận dạng một đoạn văn bản có tạo thành bảng hay không cần xem số dòng của đoạn văn bản là bao nhiêu.

Hạn chế thứ hai thường xảy ra đối với một đoạn văn bản thông thường mà đều có ký tự cách (space) tại cùng một vị trí của tất cả các dòng trong đoạn văn bản đó. Do đó đoạn văn bản đó cũng không được nhận biết đó là một khối thống nhất.

Một hạn chế khác đó là một số cột trong bảng có chung một tiêu đề. Trong trường hợp này tiêu đề chung của bảng sẽ được cho vào một khối với các cột có tiêu đề chung và thuật toán nhận biết đó chỉ là một cột. Hình 13 mô tả toàn bộ các mặt hạn chế trên.



Hình 13 Những mặt hạn chế của thuật toán

2.3. Các bước xử lý khối sau khi phân đoạn

Một số bước xử lý được đưa ra để để khắc phục những hạn chế đề cập ở trên khi nhận dạng. Trong phần này sẽ đề cập đến hai loại khối khác nhau: khối loại một là

khối chỉ bao gồm một từ trên một dòng (Hình 11), khối loại hai là tất cả các trường hợp còn lại (Hình 12). Dễ nhận thấy rằng khối loại một là một bảng đơn giản.

Phân biệt hai loại khối này sẽ giúp chúng ta dễ dàng chọn lựa từng phương pháp, kỹ thuật để phân tích từng loại khối. Phần dưới đây sẽ trình bày những phương pháp xử lý để khắc phục những trường hợp nhận dạng sai từ Hình 13.

2.3.1. Trộn các khối phân đoạn sai

Hình 13 ở trên chỉ ra một thí dụ với một đoạn văn bản thông thường mà đều có ký tự cách (space) tại cùng một vị trí của tất cả các dòng trong đoạn văn bản đó. Trong trường hợp này phương pháp phân đoạn trên đoạn văn bản đó không nhận biết đó là một khối thống nhất mà sẽ hiểu rằng đó là hai khối tách biệt nhau. Do đó ta cần có bước xử lý để nhận biết và trộn hai khối tách biệt này làm một khối thống nhất.

Trong phương pháp này chúng ta sẽ sử dụng những khối sau khi phân đoạn ở trên. Có thể thấy rõ ràng rằng các khối mà có thể trộn thành một khối chung thường nằm bên trái hoặc bên phải của nhau. Giả sử ta đã xác định được 2 khối có thể trộn với nhau, từ một khối trước tiên chúng ta sẽ đánh giá khoảng cách trung bình giữa các từ của hai khối để tìm độ rộng trung bình của ký tự cách trong đoạn văn bản. Nếu khoảng cách giữa hai khối xấp xỉ bằng độ rộng trung bình của ký tự cách thì có thể trộn hai khối đó vào làm một.

Hình 14 Trộn hai khối bị phân tách

Một lưu ý rằng khi ta xét hai khối có khả năng được trộn với nhau thì các khối đó phải thoả mãn là tất cả các dòng của khối đều có các từ nằm ngoài cùng bên trái hay bên phải có vùng bao của từ phải thẳng hàng theo chiều dọc. Tức là khi khối có một từ ở một dòng nào đó nằm thụt vào so với mép lề trái hay mép lề phải của khối (Hình 15) thì ta coi hai khối đó không có khả năng trộn với nhau.

Đối với khối loại hai chúng ta chúng ta dễ dàng tính được khoảng cách trung bình giữa các từ trên cùng một dòng, sau đó ta lấy khoảng cách đó so sánh với khoảng cách giữa hai khối. Dựa trên một số sai số đưa ra ta sẽ quyết định liệu rằng hai khối có được trộn vào với nhau hay không.

Trong trường hợp hai khối được trộn lại là hai khối loại 1 do đó ta sẽ không tính được độ rộng trung bình của các từ trong khối liền kề. Vì vậy trong trường hợp này ta sẽ tính độ rộng trung bình giữa các từ dựa vào một khối loại hai khác. Hình 14 chỉ ra hai khối được xử lý bởi kỹ thuật trên và kết quả tương ứng của nó.

2.3.2. Phân tách các cột bị trộn vào một khối

Một vấn đề khác gặp phải đó là các cột riêng biệt được trộn với nhau, chẳng hạn các cột có chung tiêu đề thường bị trộn thành một cột ở bước phân đoạn khởi tạo. Trong khi tìm ra dấu hiệu đơn giản để nhận biết các cột được tách ra ta nhận thấy rằng mối quan hệ *một – một* giữa các từ trong cột là tiêu chuẩn để đánh giá các cột được tách ra. Mối quan hệ đó phải đảm bảo là, nếu một từ W_a có chính xác một từ W_b là láng giềng dưới và W_b cũng chỉ có duy nhất W_a là láng giềng trên.

Bước tiếp theo hoàn toàn dễ hiểu: chúng ta sẽ đi phân tách tất cả các từ có quan hệ *một – một* vào thành một khối, gọi là khối con của khối đó. Do đó chúng ta không cần phải quan tâm đến khía cạnh nội dung và độ cao của khối để phân tách.

Mối quan hệ *một - một* ở trên chỉ giúp chúng ta tách được các khối con loại một (trên mỗi dòng chỉ có duy nhất một từ) do đó để tách các khối con loại hai ta phải sử dụng kỹ thuật khác.

Kết quả của quá trình phân tách sẽ được mô tả trên Hình 15 nhưng quá trình phân tách đến bước này vẫn chưa kết thúc vì cần phải xử lý một số bước nữa để tránh phân tách sai.

Hình 15 Tách các cột bị trộn

Do kỹ thuật trên áp dụng cho tất cả các khối loại hai, nhưng có một số trường hợp ta thấy rõ ràng rằng có một số lượng lớn các từ có quan hệ *một – một* nhưng chúng lại không tạo thành cột trong bảng. Tuy nhiên, trong bước xử lý ở trên chúng ta chưa áp dụng một số điều kiện ràng buộc nào để loại trừ những trường hợp đó.

Hình 16 Trộn lại các khối con bị tách

Một quy tắc đơn giản để nhận biết một cột đó là cột đó luôn đi cùng với những cột khác. Xuất phát từ các khối đã được tách ra làm khối con, chúng ta tìm đến các khối láng giềng của khối con mới được phân tách. Tìm số lượng các khối loại một bao quanh nó, độ cao của chúng, độ rộng các khoảng trắng cách ly bên trái bên phải và có thể là độ tương đồng của các từ trong cột v.v.. để đánh giá sự tồn tại của cột đó. Nếu các điều kiện trên không thoả mãn theo một tiêu chuẩn nào đó thì khối con mới được tạo ra đó sẽ được trộn ngược trở lại với khối cha nó (khi đó khối con không thoả mãn tạo thành một cột).

Cụ thể hoá quá trình nhận biết một khối con được tách riêng từ một khối cha có tạo thành một cột riêng rẽ trong bảng hay không ta sẽ đi so sánh các khối con được tách ra với nhau. Quá trình tách một khối thành các khối con sẽ chia khối cha thành các khối con được đánh số từ B_1 đến B_n. Do một khối B_i $(1 \leq i \leq n)$ bao gồm các từ liên tục nằm cạnh nhau, mỗi khối B_i có những đặc trưng (X_{Imin}, Y_{Imin}) và (X_{Imax}, Y_{Imax}). Trong đó (X_{Imin}, Y_{Imin}) là toạ độ góc trên cùng bên trái của khối và (X_{Jmax}, Y_{Jmax}) là toạ độ góc dưới cùng bên phải của khối. Vì vậy ta sẽ tìm tất cả các khối từ 1 đến n, nếu tồn tại hai khối i và j thoả mãn điều kiện như sau:

$$X_{Jmin} <= X_{Imin} < X_{Imax} <= X_{Jmax}$$

$$Y_{Jmin} <= Y_{Imin} < Y_{Imax} <= Y_{Jmax}$$

thì có nghĩa là khối i nằm trong khối j và ta sẽ thực hiện trộn hai khối i và j vào làm một khối. Quá trình sẽ tiếp tục tìm hai khối bất kỳ đến khi không có hai khối nào thoả mãn điều kiện trên thì bước tìm kiếm sẽ dừng lại. Điều kiện trên sẽ đảm bảo các khối con được tách riêng ra sẽ tạo thành một cột trong bảng hay chúng sẽ được trộn với các khối khác để tạo thành một cột của bảng khi mà khối đó không thoả mãn điều kiện tạo thành một cột riêng rẽ của bảng.

Một cách khác để nhận biết các khối con bị tách ra không tạo thành các cột trong bảng đó là dựa vào so sánh khoảng cách giữa hai khối với độ rộng trung bình của ký tự cách (khoảng cách trung bình giữa các từ trong một khối). Nhiều trường hợp do sự

trùng lặp của ký tự cách mà một khối loại hai được chia thành các khối con loại một. Do đó các khối con này phải được trộn ngược lại tạo thành một khối duy nhất.

Hình 16 chỉ ra một thí dụ một khối loại hai được phân tách thành ba khối con và kết quả sau khi phân tích ba khối này lại được trộn với khối cha tạo thành một khối duy nhất.

2.3.3. Nhóm các từ bị phân tách

Một số từ mà không có các từ làm láng giềng trên hay láng giềng dưới thì chúng có thể thuộc về một dòng phân tách (chẳng hạn dòng tiêu đề của bảng), những từ gắn vào phía cuối của một khối chưa được căn chỉnh hay những từ mô tả cho nội dung của một ô trong bảng. Những từ này sẽ được thuật toán phân đoạn khởi tạo tách ra thành các khối riêng.

Vì vậy trước tiên chúng ta cần phải tìm xem những từ bị phân tách này có nằm trong một môi trường bảng hay không, chúng có tương ứng với một ô (cell) trong bảng hay không và nếu có chúng ta cần phải xem xét chúng với toàn bộ các cột có thể có của bảng. Để đạt được điều này chúng ta sẽ từng bước đi qua từng khối và cứ ở chỗ nào có hai hoặc nhiều hơn các khối nằm kề nhau theo chiều ngang ta sẽ cho đó có thể có bảng và ta đánh giá *cấu trúc lề* bao gồm các *điểm căn lề (margin points)*.

Cấu trúc lề nắm giữ thông tin về giới hạn theo chiều dọc của các cột trong bảng và chứa hàng loạt các điểm căn lề. Các điểm căn lề này chỉ ra ranh giới bên trái, bên phải của tất cả các khối (các cột trong bảng) nằm liền kề nhau. Một điểm căn lề mới sẽ được tạo ra trong trường hợp có một điểm không nằm trong khoảng đã đưa ra. Các điểm này cũng nắm giữ thông tin liệu chúng có thể bị chặn bởi các đường biên của khối bên trái hay bên phải không (vì thế ta gọi chúng là các điểm căn lề bên trái, bên phải). Số lượng các dòng của các khối mà có liên quan đến cặp điểm căn lề trái và phải gọi là *số lượng quan hệ (reference counter)* của điểm đó. Một khoảng trắng rộng theo chiều dọc hay một khối bao phủ toàn bộ độ rộng của tài liệu sẽ đóng lại cấu trúc lề được đánh giá này.

Hình 17 Nhận biết các từ bị phân tách dựa vào các điểm phân lề

Bước tiếp theo sẽ là các điểm căn lề của tất cả các khối được xem xét. Nếu như *số lượng quan hệ* của các điểm căn lề bên trái và bên phải của một khối không đạt được một giới hạn đưa ra, thì khối này này sẽ được trộn với các khối láng giềng tương ứng theo từng phía mà xuất hiện trong một phạm vi quy định.

Tác dụng của kỹ thuật trên là nhận biết được các từ phân tách mà không thích hợp với những cột xung quoanh. Hình 17 mô tả việc đánh giá các điểm căn lề và kết quả thu được dựa vào phân tích của kỹ thuật trên.

2.4. Phân tích khối

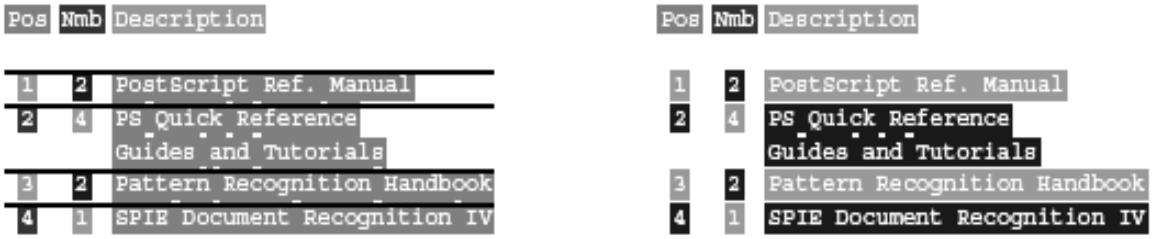
Trong khi thông thường tất cả các khối loại 2 thể hiện cho cấu trúc văn bản như là: đoạn văn bản hay đôi khi là một ô của bảng, khối loại 1 là biểu diễn của một cột trong bảng bao gồm các ô khác nhau. Để đưa ra một cấu trúc biểu diễn ở mức cao hơn từ tập hợp các loại khối trên, chúng ta cần phân chia khối loại một thành các ô của bảng. Kết quả của quá trình này được áp dụng cho Hình 15 và kết quả được đưa ra trên Hình 18.

Hình 18 Tách các khối loại 1 thành các ô của bảng

2.4.1. Khối loại 2 nằm cùng với khối loại 1

Trong trường hợp những khối loại 2 là láng giềng với khối loại 1 và ta cũng cần tách khối loại 2 thành các ô của bảng, do đó ta chỉ cần phân đoạn các dòng cho khối loại 1 thì đồng thời ta cũng tách được các ô cho khối loại 2.

Hình 19 mô tả một ví dụ về việc tách các ô trong bảng với hai cột *Pos và Nmb* là cột thuộc khối loại 1, cột *Description* là khối loại 2.



Hình 19 Tách các khối loại 2 thành các hàng trong bảng

Đầu tiên chúng ta sẽ phân đoạn khối loại 1 để tách ra các hàng trong bảng. Các hàng của bảng được phân cách với nhau bằng các đường kẻ (Hình 19 bên trái). Các đường kẻ này đồng thời cũng chia thành các hàng cho khối loại 2.

2.5. Xác định cấu trúc các cột, hàng

Sau khi đã tiến hành phân đoạn tất cả các khối cơ bản (để tách ra các ô của bảng), chúng ta vẫn cần khai thác thêm thông tin từ những khối này, xác định thêm những khối có khả năng tạo thành bảng và đặt các khối tương ứng với cột và hàng thích hợp.

Để làm việc này chúng ta sẽ sử dụng lại hệ thống ước lượng các điểm căn lề trong phần 2.3.3. Nhóm các từ bị phân tách về việc nhận biết các từ bị phân tách. Các khối láng giềng nằm theo chiều ngang tạo ra một *cấu trúc lề* bao gồm một danh sách các *điểm căn lề*.

Trong khi duyệt qua các điểm căn lề từ trái qua phải chúng ta nhận ra rằng mỗi một lần chuyển từ điểm căn lề phải sang trái xác định đường phân cách giữa hai cột của bảng và vì thế cũng tính được số lượng cột trong bảng. Trong trường hợp có những khối trải dài qua nhiều đường phân cách của hai cột thi ta coi khối đó (hay ô) chứa nhiều cột.

2.6. Kết luận chương

Chương này đã trình bày phương pháp nhận dạng bảng T-Recs, một phương pháp nhận dạng bảng với tốc độ nhanh và hiệu quả. Tư tưởng chủ yếu của thuật toán là phần phân đoạn khởi tạo, một phương pháp đơn giản nhưng thực hiện nhanh trong việc phân đoạn các khối. Chương này cũng trình bày những cải tiến của thuật toán phân đoạn khởi tạo (T–Recs) do T. G. Kieninger đề xuất trước đây nhằm giúp cho thuật toán phân đoạn các cột một cách chính xác nhất. Một số bước xử lý khối sau khi phân đoạn cũng được thêm vào nhằm khắc phục những hạn chế của thuật toán phân đoạn khởi tạo. Hơn thế nữa T-Recs cũng cho thấy nhiều ưu điểm nổi bật so với các phương pháp nhận dạng bảng khác, đặc biệt nhận dạng bảng không dựa vào dấu hiệu phân cách của bảng.

Mặc dù vậy thuật toán cũng đề ra những thách thức, đó là một số vấn đề vẫn còn tồn tại mà thuật toán chưa phân tích đúng. Trường hợp thứ nhất đó là việc tách các khối loại 2 (không nằm cùng khối loại 1) ra thành các hàng trong bảng. Thuật toán chỉ

đề ra phương pháp tách các hàng dựa vào khối loại một. Trường hợp thứ hai, thuật toán thường nhận dạng sai đối với các dòng tiêu đề của thư, chẳng hạn như địa chỉ thư, ngày gửi .v.v.. thuật toán thường nhận dạng chúng là bảng dữ liệu. Do đó những cải tiến phương pháp nhận dạng trong những trường hợp trên là cần thiết để thuật toán nhận dạng được tất cả các loại bảng.

Một trong những thách thức lớn đối với phương pháp nhận dạng bảng T-Recs là khả năng nhận dạng được môi trường bảng trong một trang tài liệu, thông thường T-Recs sẽ được thực hiện trên từng đoạn văn bản của trang tài liệu và việc phân tích trên từng đoạn văn bản đó để xem có tồn tại bảng không. Trong trường hợp một bảng dữ liệu và một đoạn văn bản nằm liền kề nhau không có sự khác biệt lớn (chẳng hạn không coi đó là hai đoạn văn bản riêng biệt) thì sẽ rất khó để xác định được bảng.

CHƯƠNG 3
THỰC NGHIỆM

3.1. T-Recs++

3.1.1. Giới thiệu

Phần này sẽ mô tả chương trình thử nghiệm T-Recs++ System nhận dạng bảng bằng thuật toán T-Recs++ được mô tả trong 2.2.2. Cải tiến các bước của thuật toán phân đoạn khởi tạo - T-Recs++. Chương trình được viết bằng ngôn ngữ Visual C# và sử dụng phần mềm ứng dụng ComponentOne để thiết kế giao diện. Chương trình gồm hai phần chính:

- Phần một là quét qua toàn bộ ảnh để nhận dạng và xây dựng các hình bao của các từ có trong ảnh.
- Phần thứ hai dựa vào các hình bao thu được từ phần một và thuật toán T-Recs++ để nhận dạng các cột có thể có của bảng trong từng trang tài liệu ảnh.

3.1.2. Mô tả chương trình

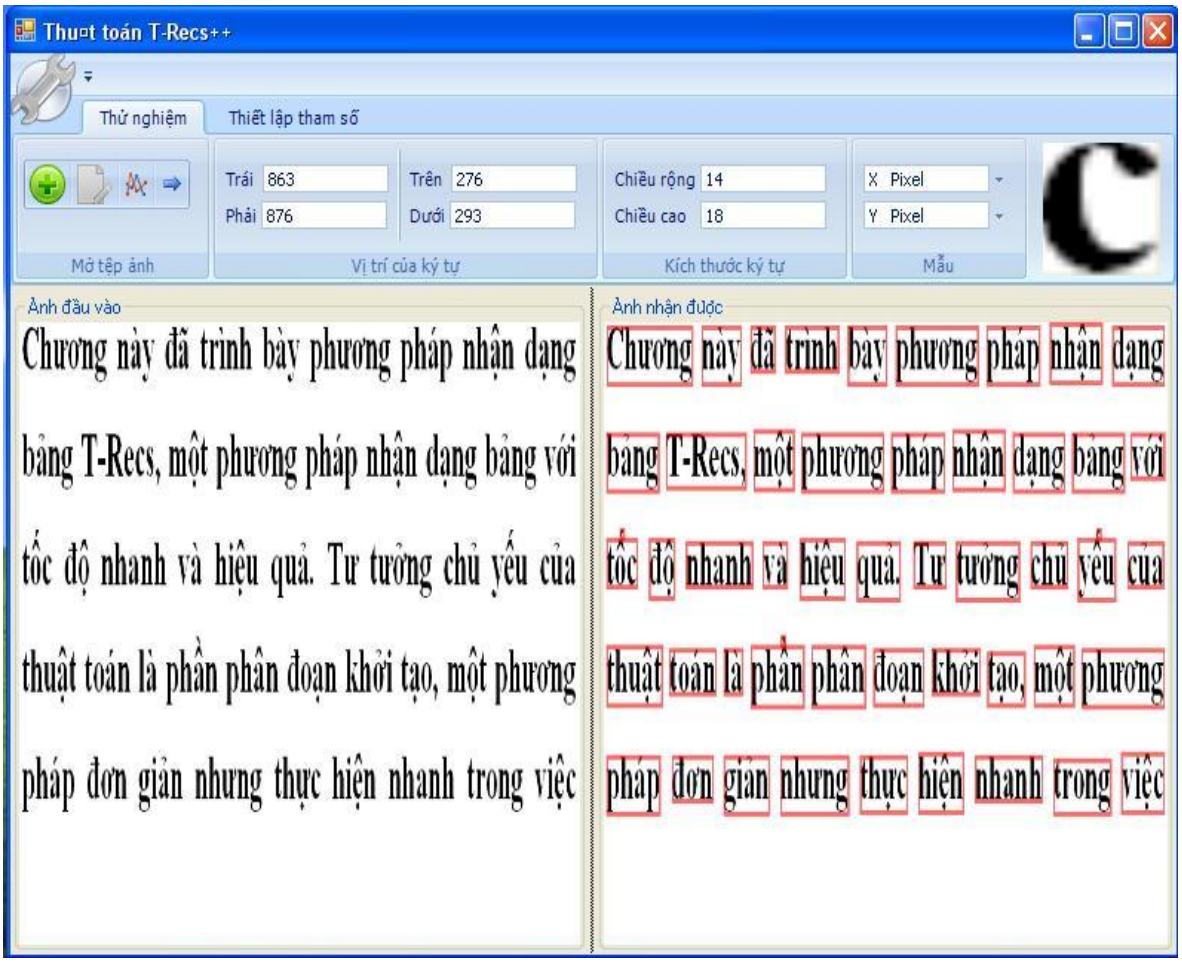
Chương trình thử nghiệm dưới đây chỉ dừng lại ở phần chính đó là nhận dạng các cột có thể có của bảng. Do thời gian hoàn thành luận án còn hạn chế nên các bước xử lý nhằm khắc phục lỗi hay bước tách các cột của bảng thành các dòng chưa đưa vào trong chương trình. Chương trình cũng chỉ áp dụng nhận dạng các bảng khi chúng không có các đường kẻ. Chương trình hoạt động bao gồm các bước như sau:

1) Tài liệu ảnh được tải vào chương trình bằng việc người sử dụng chọn một tệp ảnh nhị phân (bmp) để mở. Khi đó tài liệu ảnh sẽ được quét để nhận dạng số dòng văn bản có trong tài liệu.
2) Sau khi tài liệu được quét để nhận dạng số dòng văn bản có trong tài liệu. Chương trình sẽ thực hiện quét lần lượt qua tất cả các dòng, tại mỗi dòng sẽ nhận dạng từng ký tự và nhận dạng từng từ trên mỗi dòng. Từ đó xây dựng hình bao cho mỗi từ trên từng dòng.
3) Dựa vào thông tin hình bao của các từ, chương trình sẽ xây dựng các từ nằm trong cùng một khối bằng thuật toán T-Recs++. Thuật toán sẽ quét từ

trên xuống dưới và nhận biết các đoạn văn bản khác nhau, sau đó thực hiện thuật toán T-Recs++ trên các đoạn văn bản khác nhau đó. Cách nhận biết các đoạn văn bản khác nhau đó là dựa vào khoảng trắng giữa các dòng văn bản. Và kết quả chương trình đưa ra ảnh bao gồm các từ thuộc cùng một khối (một cột).

Chương trình có một tham số cần phải thiết lập (đặt mặc định là 5), đó là tham số *số điểm ảnh lớn nhất giữa hai ký tự trong một từ*, bởi vì tham số này phụ thuộc vào kích cỡ của phông chữ. Tham số này giúp xác định các ký tự thuộc cùng một từ. Dựa vào tham số này để chương trình nhận biết khi hai ký tự cách nhau một khoảng như thế nào thì nhóm chúng lại làm một từ.

Giao diện của chương trình được mô tả trên Hình 20:



Hình 20 Giao diện chương trình T-Recs++

Nhấn vào nút ***Mở tệp ảnh*** để chọn một ảnh nhị phân để mở.

Nhấn vào nút ***Đặt tham số*** để thiết lập tham số *số điểm ảnh tối đa giữa hai ký tự* (được đặt mặc định là 5).

Nhấn vào nút ***XD hình bao*** để nhận dạng hình bao cho các từ trong ảnh và đồng thời chương trình sẽ vẽ ra môt hình chữ nhật nhỏ nhất bao từ.

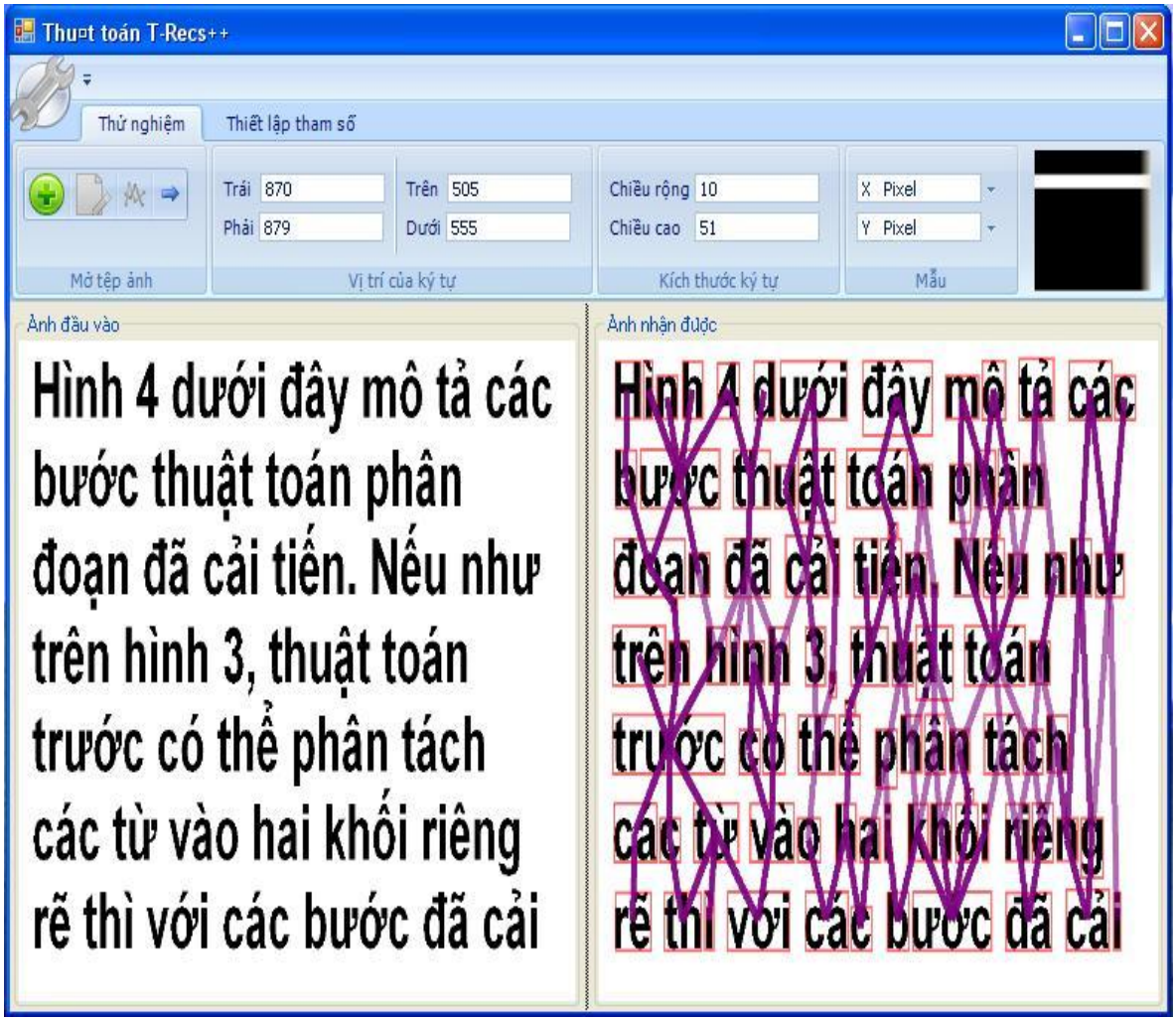
Nhấn vào nút ***Nhận dạng*** để nhận dạng các cột có thể có của bảng trong ảnh.

Nút ***Ký tự tiếp theo*** cho phép nhận dạng từng ký tự của ảnh.

Thông tin về toạ độ, chiều rộng, chiều cao, hình dạng của từng ký tự sẽ hiển thị phía trên khi nhận dạng.

3.1.3. Một số kết quả thử nghiệm

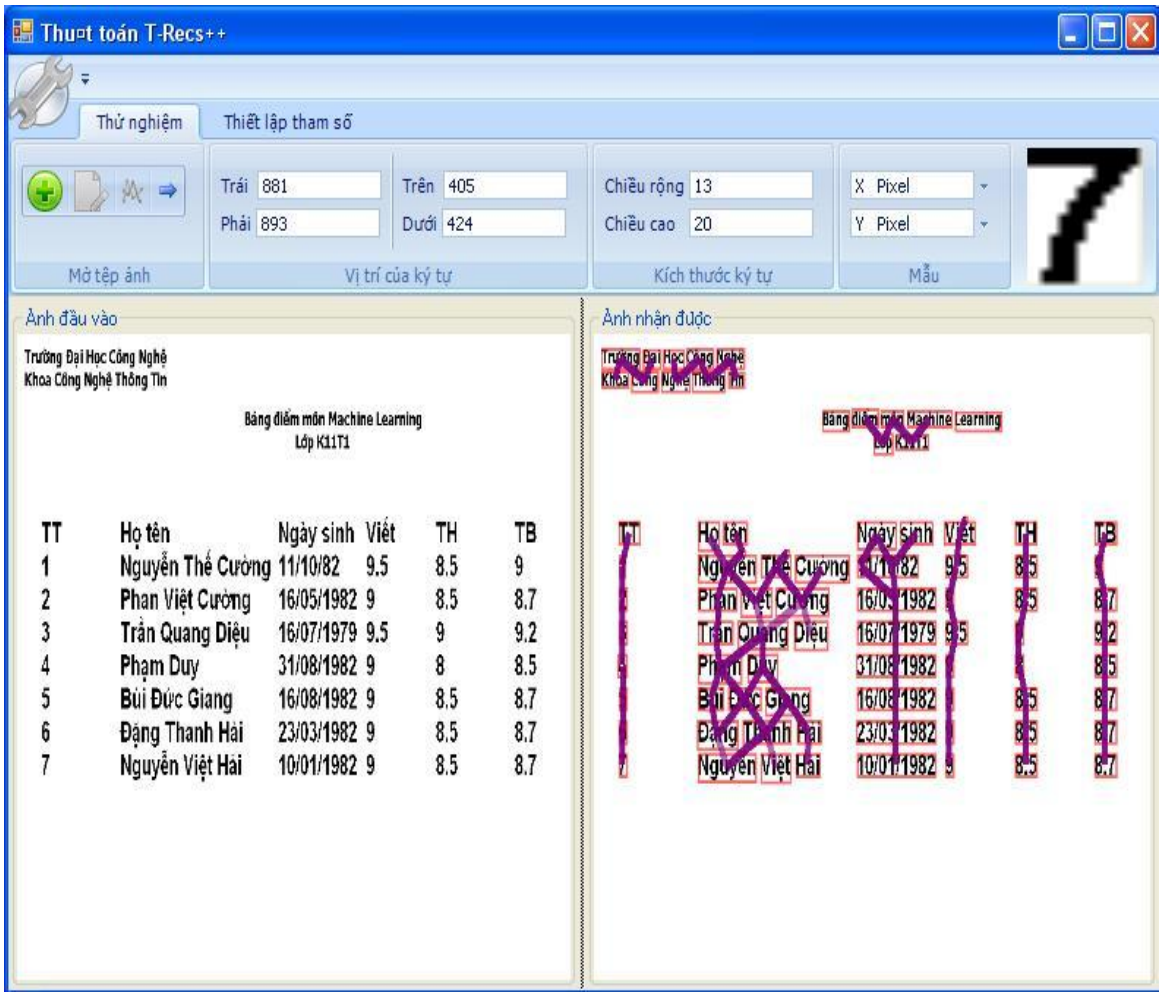
Hình 21 là kết quả nhận dạng đối với một đoạn văn bản thông thường. Với một đoạn văn bản thông thường, T-Recs++ chỉ xây dựng được một khối duy nhất.



Hình 21 Kết quả xây dựng khối của T-Recs++

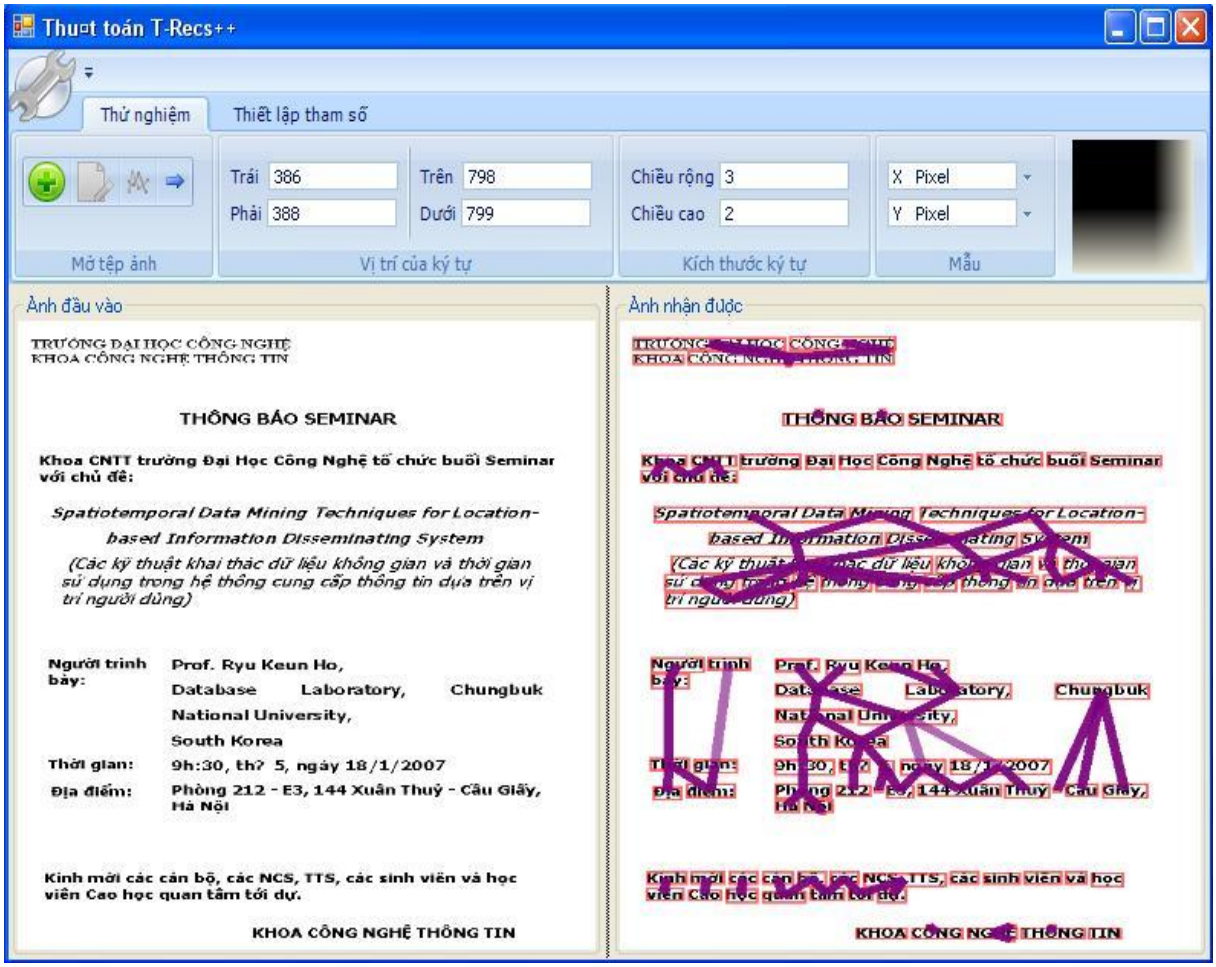
Một trường hợp khác có tồn tại môi trường bảng như trên Hình 22. Đầu tiên chương trình nhận thấy ảnh có ba đoạn văn bản và thực hiện thuật toán T-Recs++ trên ba đoạn văn bản này. Mặc dù hai đoạn văn bản phía trên không phải là bảng và có ký tự cách trùng lặp ở một vị trí, và thuật toán đã nhận dạng những đoạn văn bản này có

nhiều hơn một cột dữ liệu. Tuy nhiên đó không phải là một vấn đề khó, dựa vào đánh giá độ rộng trung bình của ký tự cách ta có thể trộn lại các khối bị phân tách vào thành một khối duy nhất. Trong tệp ảnh trên Hình 22 chỉ có mỗi đoạn văn bản thứ ba là môi trường bảng và thuật toán đã nhận dạng chính xác 6 cột của bảng.



Hình 22 Trường hợp nhận dạng có môi trường bảng

Một kết quả nhận dạng khác phức tạp hơn được chỉ ra trên Hình 23. Chỉ có một môi trường bảng duy nhất. Các trường hợp chỉ có một dòng văn bản hay có ký tự cách trùng lặp có thể dễ dàng xử lý để nhận biết không có môi trường bảng.



Hình 23 Trường hợp nhận dạng đối với một văn bản thông báo

KẾT LUẬN

Phân tích tài liệu ảnh là một lĩnh vực đã được đề xuất và phát triển trong một thời gian khá lâu, một số sản phẩm thương mại về hệ Phân tích tài liệu ảnh cũng đã xuất hiện trên thị trường. Tuy nhiên ngành nhận dạng đã và đang rất phát triển bởi những thách thức đặt ra đối với những vấn đề mới và đòi hỏi những cải tiến để nâng cao tính chính xác và hiệu quả của các hệ Phân tích tài liệu ảnh. Yêu cầu đối với một hệ Phân tích tài liệu không chỉ đơn giản là chuyển đổi nội dung của tài liệu ảnh sang định dạng tài liệu có thể soạn thảo được mà còn phải nhận dạng ra cấu trúc nội dung lưu trữ trong từng trang tài liệu. Nhận dạng bảng, nhận dạng biểu mẫu là bài toán điển hình về nhận dạng cấu trúc trong tài liệu ảnh. Trong khuôn khổ của mình, luận án đã đi vào nghiên cứu thuật toán nhận dạng bảng trong tài liệu ảnh kỹ thuật.

Thuật toán nhận dạng bảng được đề cập trong luận án dựa trên thuật toán T-Recs do G. Kieninger đề xuất, tuy nhiên các bước thực hiện mà Kieninger đưa ra vẫn còn nhiều hạn chế và nhận dạng sai trong một số trường hợp. Luận án đã đưa ra những cải tiến các bước thực hiện của thuật toán, đặt tên là T-Recs++. Một số thuật toán nhận dạng bảng trước đây dựa trên dấu hiện phân cách các ô trong bảng, chẳng hạn như là các đường kẻ, khoảng trắng .v.v.. Tuy nhiên T-Recs++ là phương pháp nhận dạng bảng không dựa trên một dấu hiệu phân cách nào, kể cả trong trường hợp khoảng cách giữa hai cột trong bảng cách nhau một khoảng cách hẹp. Đó chính là một ưu điểm nổi bật của thuật toán. Trong một khối văn bản thuật toán nhận dạng khá chính xác các cột dữ liệu nếu có của bảng. Một vấn đề còn tồn tại của thuật toán đó là khả năng nhận dạng các dòng của bảng, trong thuật toán này việc nhận dạng ra dòng của bảng phụ thuộc vào khối loại một. Trong trường hợp bảng chỉ bao gồm khối loại hai phương pháp tách các dòng của bảng nhận dạng phải dựa vào dấu hiệu phân tách nào đó. Kết quả thực nghiệm ở trên cho thấy thuật toán T-Recs++ có khả năng nhận dạng chính xác gần như hoàn toàn các cột có của bảng, kể cả trong trường hợp khoảng cách giữa các cột của bảng hẹp. Kết quả thực nghiệm cho thấy độ chính xác trong việc nhận dạng các cột của bảng từ 92% đến 96%. Một số lỗi nhận dạng có thể xuất hiện là trường hợp trùng lặp ký tự cách tại cùng một vị trí trên các dòng văn bản, một số từ nằm ở các vị trí bất thường của đoạn văn bản cũng tạo thành cột hay trường hợp các dòng đơn .v.v..

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Tô Văn Khánh & Ngô Quốc Tạo: "Áp dụng phương pháp T–Recs vào nhận dạng bảng". Hội nghị khoa học Viện CNTT, 12-2006.

Tiếng Anh

[2] O'Gorman & Kasturi: "Document image analysis", 1997

[3] Kasturi, O'Gorman, Govindaraju: "Document image analysis: A primer", 2002

[4] Arcelli C, Sanniti di Baja G 1985: "A width-independent fast thinning algorithm". *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-7: 463–474.

[5] Arcelli C, Sanniti di Baja G 1993 Euclidean skeleton via center-of-maximal-disc extraction. *Image Vision Comput.* 11: 163–173

[6] Fukunaga K, Hostetler L D 1975 K-nearest-neighbour Bayes-risk estimation. *IEEE Trans. Inf. Theor.* 21: 285-293

[7] Murthy B K, Deshpande W R 1998 Optical character recognition (OCR) for Indian languages. Proc. Int. Conf. on Computervision, Graphics, Vision, Image Process. ICVGIP, New Delhi

[8] G S Lehal1, Renu Dhir: "A Range Free Skew Detection Technique for Digitized Gurmukhi Script Document"

[9] Pavlidis T, Zhou J 1991 Page segmentation by white streams. Proc. 1st Int. Conf. on Document Analysis and Recognition ICDAR), St. Malo, France, pp 945–953

[10] O'Gorman L 1993 The document spectrum for structural page layout analysis. *IEEE Trans. Pattern Anal. Machine Intelli.* AMI-15: 1162–73

[11] Nartker T A, Rice S V, Kanai J 1994 OCR Accuracy. UNLV's Second Annual Test. Technical Journal INFORM, University of Nevada, Las Vegas

[12] Sawaki M, Hagita K 1998 Text-line extraction and character recognition of document headlines with graphical design using omplimentary similarity measure. IEEE Trans. Pattern Anal. Machine Intell. PAMI-20: 1103–1109

[13] Wilson C L, Geist J, Garris M D, Chellapa R 1996 Design, integration, and evaluation of form-based handprint and OCR ystems. Technical Report, NISTIR5932, National Institute of Standards & Technology, US; download from http://www.itl.nist.gov/iad/894.03/pubs.html

[14] Fletcher A, Kasturi R 1988 A robust algorithm for text string separation from mixed text /graphics images. *IEEE Trans. Pattern nal. Machine Intell.* PAMI-10: 910–918

[15] Thomas G.Kieninger, "Table Structure Recognition Based On Robust Block Segmentation", 1998.

[16] Thomas G.Kieninger and Andreas Dengel, "Applying The T-Recs Table Recognition System To The Business Letter Domain". In Proceeding of the Sixth International Conference On Document Analysis and Recognition -ICDAR'01, 2001

[17] T. Hu, “New Methods for Robust and Efficient Recognition of the Logical Structures in Documents”. PHD thesis, Institure of Informatics of the University of Fribourg, Switzerland, 1994.

[18] A. S. Condit, “Autotag: A tool for creating, structured document collection from printed materials,” Master’s thesis, Dept. of Computers Science, University of Nevada, Las Vegas, 1995.

[19] A. Dengel, “About the logical partitioning of document images”, in Proc. Of Int’t Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, Apr. 1994