RÚT TRÍCH VĂN BẢN TỪ TẬP TIN HÌNH ẢNH VỚI TESSERACT

TRẦN THANH PHƯỚC
Khoa Công nghệ Thông tin – Trường ĐHCN Thực phẩm Tp.HCM

TÓM TẮT

Rút trích văn bản từ tập tin hình ảnh đang là một trong những bài toán quan trọng trong xử lý ảnh hiện nay. Trong bài báo này, chúng tôi bước đầu tìm hiểu các phương pháp trích lọc văn bản từ hình ảnh của một số công trình liên quan đồng thời cũng tìm hiểu, hiệu chỉnh công cụ mã nguồn mở Tesseract để thực hiện trích lọc văn bản tiếng Anh từ tập tin hình ảnh. Kết quả thử nghiệm bước đầu cho thấy công cụ này rút trích khá tốt các văn bản từ tập tin hình ảnh chứa văn bản được đánh máy.

Từ khóa: Rút trích văn bản, tập tin hình ảnh, Tesseract.

EXTRACTING TEXT FROM IMAGE FILES USING TESSERACT

ABSTRACT

Extracting text from the image file is one of the important problems in image processing. In this paper, we initially study the methods of text extracting from images from a number of related works. Besides, we also learn and adjust the Tesseract, an open source tool, to perform extracting English texts from the image file. Initial test results show that this tool quite extracted the text from the image file containing the typed text.

Key words: Extracting text, image files, tesseract.

1. Giới thiệu

Hiện nay, nhu cầu về việc rút trích từ ngữ từ hình ảnh đang ngày càng phát triển, bên cạnh sự gia tăng về nhu cầu là sự phát triển của công nghệ nhận dạng ký tự quang học (Optical Character Recognition) hay còn được gọi tắt là **OCR**. Đây là một công nghệ giúp chuyển đổi hình ảnh của chữ viết tay hoặc đánh máy thành các ký tự đã được mã hóa trong máy tính.

Giả sử chúng ta cần chỉnh sửa một số tài liệu giấy như: Các bài viết trên tạp chí, tờ rơi, hoặc một tập tin PDF hình ảnh. Rõ ràng, chúng ta không thể sử dụng một máy quét để chuyển các tài liệu này thành tập tin văn bản để có thể chỉnh sửa (ví dụ như trình soạn thảo Microsoft Word). Tất cả những gì máy quét có thể làm là tạo ra một hình ảnh hoặc một bản chụp của các tài liệu. Để giải nén và sử dụng lại dữ liệu từ tài liệu được quét, hình ảnh máy ảnh hoặc hình ảnh của các tập tin PDF, chúng ta cần một phần mềm OCR. Nó sẽ xuất ra kí tự trên hình ảnh, ghép chúng thành từ và sau đó ghép các từ thành câu. Nhờ vậy, chúng ta có thể truy cập và chỉnh sửa nội dung của tài liệu gốc.

Tương tự, những tài liệu cổ đang bị hư hại theo thời gian và việc viết tay hay đánh máy lại những tài liệu này sẽ tốn rất nhiều chi phí, thời gian và không đảm bảo được độ chính xác cũng như là sự an toàn cho tài liệu nền. Việc này rất cần một công nghệ lấy từ ngữ từ hình ảnh chụp.

Trong bài báo này, chúng tôi sẽ tìm hiểu, chỉnh sửa công cụ Tesseract để thực hiện việc rút trích các văn bản từ tập tin hình ảnh. Bài báo được trình bày như sau: Phần 2, chúng tôi sẽ trình bày các công trình liên quan đến việc rút trích văn bản. Ở phần 3, chúng tôi sẽ trình bày công cụ Tesseract cũng như cách rút trích văn bản của công cụ này. Phần thử nghiệm sẽ được chúng tôi trình bày ở phần 4 và phần 5 sẽ trình bày kết luận.

2. Công trình liên quan

Có nhiều phương pháp để tạo ra một phần mềm dạng OCR, độ chính xác của các phương pháp này phụ thuộc vào công nghệ tạo nên phần mềm. Các phương pháp này đạt được độ tin cậy trong các hình ảnh có chất lượng tốt và vừa.

Độ chính xác của việc rút trích văn bản là điều quan trọng nhất. Nhóm tác giả Kirill Safronov [1] cho rằng một số sai sót trong quá trình chuyển đổi thường không quá quan trọng trừ các trường hợp như rút trích số serial từ ảnh chụp,...

Để khắc phục tình trạng kết quả xuất ra không chính xác của công nghệ OCR, nhiều công nghệ khác đã ra đời, tác giả A. Vinutha M H [2] đã ứng dụng định hướng robot (Optical Character Recognition Based Auto Navigation of Robot). Việc định hướng của robot dựa vào bảng tính hiệu như là một cột mốc đánh dấu đường đi tiếp theo của robot. Định hướng tự động của các robot trong một vùng lớn đòi hỏi nhiều bảng tín hiệu khác nhau với mô hình nhận dạng duy nhất. Ngoài ra, hệ thống này còn cho phép nhận diện vị trí tên riêng.

Bên cạnh việc cải thiện độ chính xác, cần có sự thay đổi kích thước của thiết bị nhận dạng, tác giả Ali Ahmadi [3] đã đề cập trong nghiên cứu của mình, tốc độ xử lý và độ chính xác cao là yêu cầu lớn hiện nay của các thiết bị nhận dạng ký tự dạng nhỏ, ví dụ như bút biết nhận dạng. Nhưng dù có nhiều mặt hàng loại này được chào bán trên thị trường nhưng nó vẫn không đáp ứng nhu cầu sử dụng và kích thước thiết bị.

Ngoài sự đa dạng trong cách thức nhận dạng, OCR còn đa dạng về cách dùng, nó được chia thành hai cách, dùng online và dùng offline, tác giả Priya Sharma [4] có nhận xét về hai cách dùng này như sau: (1) Nhận dạng offline: nhận dạng các văn bản in ra giấy hoặc các bản viết tay và nó đòi hỏi quá trình scan trên mặt giấy hoặc mặt vật liệu có chữ. Cách này thường đòi hỏi con người phải thực hiện một số thao tác như phân loại, lưu trữ và chỉnh sửa văn bản trước khi scan. (2) Nhận dạng online: thường chỉ được dùng cho nhận dạng chữ viết tay được lưu trữ ở dạng kỹ thuật số, thông thường để scan dạng này chúng ta thường dùng một loại bút đặc biệt nhưng do sự thành công của các nghiên cứu gần đây mà giờ đã có các thiết bị khác thay thế. Việc nhận dạng online nhằm giúp con người giao tiếp với máy tính tốt hơn bằng cách viết tay thay vì gõ phím.

Trong bài báo này, chúng ta sẽ tìm hiểu về một công cụ OCR điển hình và là một trong những nền tảng quan trọng, đó là Tesseract.

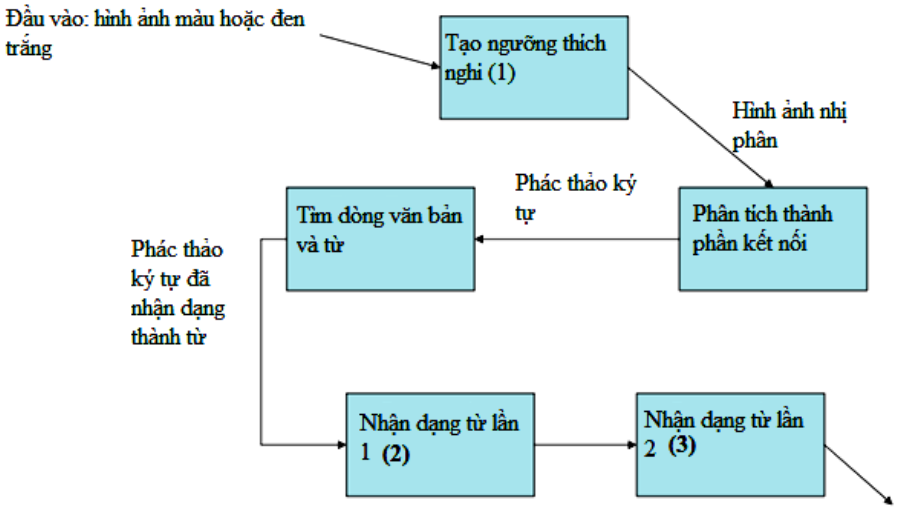
3. Rút trích văn bản từ tập tin hình ảnh với Tesseract

3.1. Giới thiệu Tesseract

Tesseract là một công cụ OCR mã nguồn mở được nghiên cứu và phát triển bởi HP trong giai đoạn 1984-1994. Nó được biết như là một phần mềm thêm vào cho dòng sản phẩm máy quét của HP. Trong giai đoạn này, nó vẫn còn rất sơ khai và chỉ được dùng để cải thiện chất lượng của các bản in. Nó được phát triển cho đến năm 1994 thì ngưng. Sau khi được cải thiện độ chính xác, nó được HP đưa vào cuộc kiểm tra thường niên về độ chính xác của các công cụ OCR và nó đã thể hiện được sự vượt trội của mình. Kể từ năm 2006, nó đã được cải thiện rộng rãi bởi Google.

Tesseract hoạt động trên Linux, Windows (với VC + + Express hoặc Cygwin) và Mac OSX. Chúng ta có thể tải về tại địa chỉ *http://code.google.com/p/tesseract-ocr.*

3.2. Cấu trúc của Tesseract



Hình 1. Cấu trúc của Tesseract

Tạo ngưỡng thích nghi giúp loại bỏ các yếu tố nền của hình ảnh (ví dụ như ánh sáng, bóng,…) và giúp phân tích các pixel thành ảnh nhị phân.

Nhận dạng được tiến hành qua một quá trình với hai lần nhận dạng. Lần thứ nhất: nhận ra lần lượt từng từ. Mỗi từ có nghĩa là đạt yêu cầu và được thông qua và được lưu vào dữ liệu. Lần thứ hai, khi phân loại thích ứng, công cụ sẽ nhận dạng lại các từ không được nhận dạng tốt ở lần trước đó.

3.3. Xác định dòng và từ

Xác định dòng

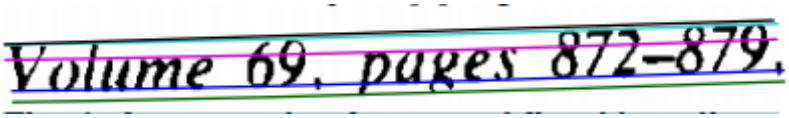
Mục đích của bước này là nhận dạng các dòng của các hình ảnh bị nghiêng, giúp giảm sự mất thông tin khi nhận dạng ảnh nghiêng. Các bộ phận quan trọng của quá trình

này là lọc dãy màu (còn được gọi là blobs) và xây dựng dòng. Bước này cũng giúp loại bỏ các văn bản có drop-cap.

Thiết lập dòng cơ sở

Khi dòng văn bản được tìm thấy, các dòng cơ sở được thiết lập chính xác hơn bằng cách sử dụng một đường có tên là spline toàn phương (là dòng mà được kết hợp từ nhiều đoạn). Nó giúp Tesseract xử lý các trang có đường cơ sở là đường cong.

Các dòng cơ sở được thiết lập bằng cách phân vùng các blobs thành các nhóm có thể thay thế thích hợp liên tục trong đường cơ sở thẳng ban đầu. Một spline toàn phương được thiết lập cho phân vùng dày đặc nhất, (giả định là đường cơ sở) của một hình có phương ít nhất. Spline có lợi thế là tính toán ổn định, nhược điểm là sự gián đoạn có thể xảy ra khi nhiều phân đoạn spline được yêu cầu.



Hình 2. Ví dụ về một đường cơ sở dạng cong

Cắt nhỏ từ

Tesseract sẽ xác định xem có các ký tự dính với nhau trong một từ hay không. Nếu có nó sẽ cắt nhỏ các ký tự ra thành các ký tự riêng lẻ.

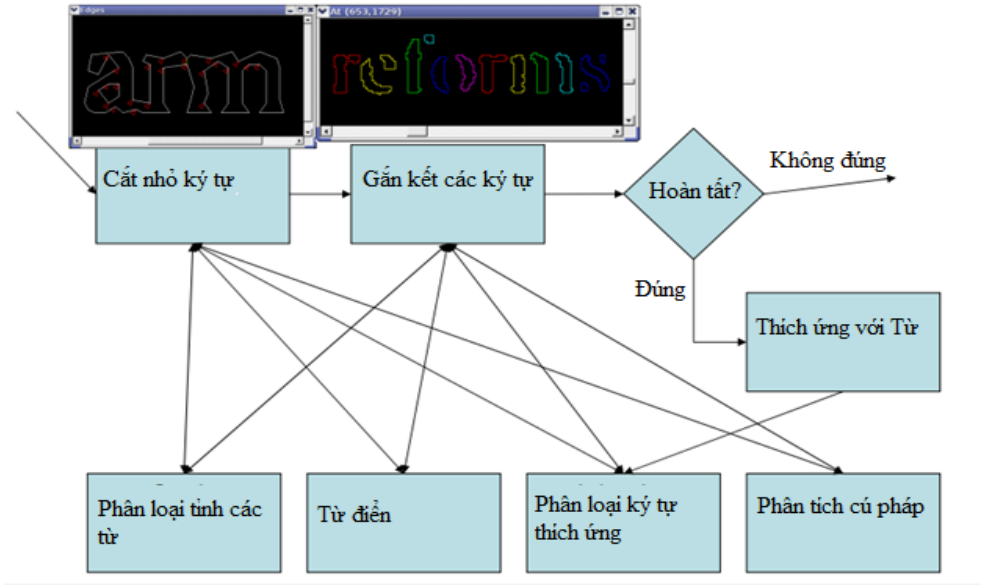
Hình 3. Ví dụ về cắt các ký tự bị dính

Nhận dạng khoảng cách giữa chữ hoặc số

Xác định khoảng cách giữa các số hoặc giữa các chữ là một vấn đề khá phức tạp. Tesseract giải quyết những vấn đề này bằng cách đo khoảng cách trong một phạm vi hạn chế theo chiều dọc giữa dòng cơ sở và dòng trung bình.

Nhận dạng từ

Quá trình nhận dạng một từ là quá trình phân tích một từ được chia ra thành các ký tự như thế nào.



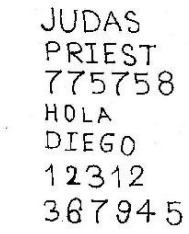
Hình 4. Quá trình nhận dạng từ

Khi kết quả xuất ra một từ mà nó không thỏa mãn nhu cầu thì Teseract cố gắng cải thiện kết quả này bằng cách cắt nhỏ các từ có nghĩa không tốt nhất. Nếu việc cắt nhỏ không làm tăng chất lượng từ thì nó sẽ phục hồi lại từ trước đó.

4. Một số thử nghiệm

Chúng tôi tiến hành thử nghiệm trên ba loại hình ảnh: Hình chụp từ chữ viết tay (1), hình chụp từ chữ đánh máy (2) và hình từ tập tin pdf (3).

Hình chữ viết tay



Hình 5. Một ví dụ về hình chứa chữ viết tay

- Kết quả:

JUDA$
PRIEST
775758
HOLA
DIEGO
12312
387945

- Tỉ lệ sai: 1/33 chiếm 3,03%.

Hình chữ đánh máy

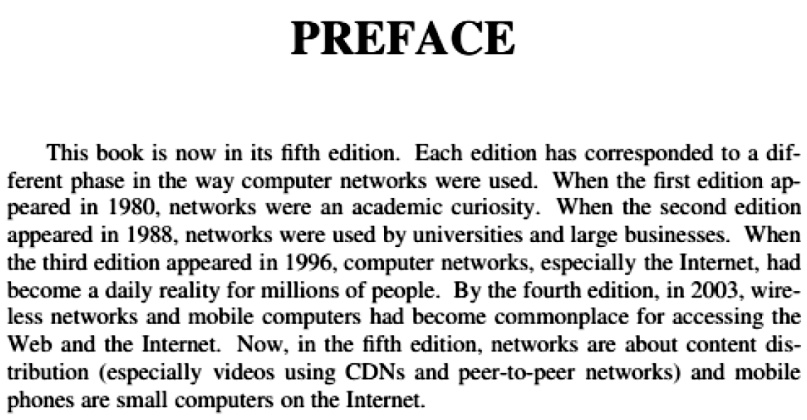
Hình 6. Một ví dụ về hình chứa chữ đánh máy

- Kết quả:

ESTA67
ES767
UNA4567
PRU EBA5887

- Tỉ lệ sai: 1/28 chiếm 3,57%

Hình ảnh tập tin.pdf



PREFACE

This book is now in its fifth edition. Each edition has corresponded to a different phase in the way computer networks were used. When the first edition appeared in 1980, networks were an academic curiosity. When the second edition appeared in 1988, networks were used by universities and large businesses. When the third edition appeared in 1996, computer networks, especially the Internet, had become a daily reality for millions of people. By the fourth edition, in 2003, wireless networks and mobile computers had become commonplace for accessing the Web and the Internet. Now, in the fifth edition, networks are about content distribution (especially videos using CDNs and peer-to-peer networks) and mobile phones are small computers on the Internet.

Hình 7. Một ví dụ về hình dạng pdf

- Kết quả:

PREFACE
This book is now In "5 mm edllmn Eden edmon has cormsponded In a d|f—
teaenr phase rn me way camplllnt networks were used When the firs! edman ap
peared in man. networks weae an academic cum: Iy When me second edmorr
appeared In 1933. networks were used by unlvcrslues and large businesses When
lhe nrrrd ednmn appeared in 1995, compuler networks. especially lhe Inrer-rrer,
had
become a duly reamy rar mrnmna cl penplc By lhe rrnrrnr edllmn. in 2003. wu':—
less nclwmks and mohllc compumeus had become commonplace for accessing rhe
Web and me unerrrer. Now, In [he mun edllkm, networks are about content
u1bullan(espeda.I|y videos using cum and pecuopccr networks) and mobile
phones are small mmpulers on the xnrerner

- Tỉ lệ sai trên 50% so với văn bản gốc. Văn bản càng dài độ chính xác càng giảm dần.

5. Kết luận

Trong bài viết này, chúng tôi giới thiệu về công cụ OCR với mã nguồn mở - Tesseract. Công cụ này dùng để nhận dạng kí tự trên một tập tin hình và chuyển kí tự thành tập tin thành văn bản. Bên cạnh những ưu điểm vượt trội của mình, Tesseract cũng có một số những hạn chế như nhầm lẫn giữa chữ hoa và chữ thường, nhầm lẫn giữa các kí tự có hình dáng tương tự, đúng từ nhưng sai trong ngữ cảnh.

Hướng tiếp theo, chúng tôi sẽ tiếp tục nghiên cứu để nâng cao chất lượng cho bài toán rút trích văn bản tiếng Anh từ tập tin hình ảnh, đồng thời bắt đầu nghiên cứu rút trích văn bản cho tiếng Việt có dấu.

TÀI LIỆU THAM KHẢO

1. Kirill Safronov: Optical Character Recognition Using Optimisation Algorithms. Institute for Process Control and Robotics (IPR) University of Karlsruhe Karlsruhe, Germany (2007).

2. Vinutha MH, Sweatha KN and Sreepriya Kurup: Optical Character Recognition Based Auto Navigation of Robot (2013).

3. Ali Ahmadi, Yoshinori Shirakawa, Md.Anwarul Abedin, Kazuhiro Takemura, Kazuhiro Kamimura, Hans Jürgen Mattausch, and Tetsushi Koide: Real-time Character Reconition System Using Associative Memory Base Hardware, Japan.

4. Priya Sharm, Randhir Singh: Performance of English Character Recognition with and without Noise, India (2013).