

# Cải Thiện Kết Quả Nhận Dạng Ký Tự Quang Học bằng Tiền Xử Lý Dữ Liệu

Nguyễn Minh Thuận

MSSV: 19120388

Khoa Công nghệ Thông tin

Trường Đại học Khoa học Tự Nhiên

Đại học Quốc gia TP.HCM

Thành phố Hồ Chí Minh, Việt Nam

19120388@student.hcmus.edu.vn

Trần Anh Huy

MSSV: 19120082

Khoa Công nghệ Thông tin

Trường Đại học Khoa học Tự Nhiên

Đại học Quốc gia TP.HCM

Thành phố Hồ Chí Minh, Việt Nam

19120082@student.hcmus.edu.vn

Đỗ Hoài Nam

MSSV: 19120296

Khoa Công nghệ Thông tin

Trường Đại học Khoa học Tự Nhiên

Đại học Quốc gia TP.HCM

Thành phố Hồ Chí Minh, Việt Nam

19120296@student.hcmus.edu.vn

**Tóm tắt nội dung**—Trong thập niên thứ hai của thế kỷ XXI, hình ảnh tài liệu chụp bằng thiết bị di động trở nên phổ biến hơn vì tính tiện lợi và nhanh chóng nhưng chất lượng ảnh lại không phù hợp cho việc nhận dạng ký tự quang học, làm kết quả kém chính xác. Trong bài báo này, chúng tôi đề xuất phương pháp tiền xử lý dữ liệu trước khi tiến hành nhận dạng gồm phân tách nội dung với nền, điều chỉnh độ nghiêng và làm phẳng ảnh. Phương pháp này được thử nghiệm trên tập dữ liệu các hình ảnh tài liệu chụp bằng máy ảnh kỹ thuật số cho kết quả đầy hứa hẹn với tỷ lệ lỗi ký tự trung bình là 6.72%, giảm gần 10 lần so với nhận dạng hình ảnh chưa qua xử lý. Nghiên cứu này có thể áp dụng vào các lĩnh vực cần nhận dạng với độ chính xác cao như xây dựng thư viện số, lưu trữ văn bản hành chính, nhận dạng giấy tờ,...

**Index Terms**—Optical Character Recognition, image pre-processing, recognition accuracy, mobile-captured documents.

## I. GIỚI THIỆU

Nhận dạng ký tự quang học là một trong những đề tài nghiên cứu vô cùng hấp dẫn, đã được giới khoa học nghiên cứu trong thời gian dài vì có nhiều ứng dụng thực tiễn trong nhiều lĩnh vực. Nó giúp việc chuyển đổi ảnh scan, ảnh chụp tài liệu thành dạng văn bản để có thể lưu trữ, sử dụng dễ dàng trong máy tính. Phương pháp nhận dạng truyền thống chỉ tập trung tối ưu cho ảnh scan tài liệu, những hình ảnh này có đặc điểm chung là văn bản cần nhận dạng phẳng, ngay ngắn, thẳng hàng và được phân tách với nền một cách rõ ràng. Tuy nhiên trong thập niên thứ hai của thế kỷ XXI, với sự phát triển mạnh mẽ của công nghệ, chất lượng ảnh chụp bởi máy ảnh kỹ thuật số, điện thoại thông minh đã được cải thiện đáng kể, vì vậy những hình ảnh tài liệu được chụp bởi những thiết bị này đã trở nên phổ biến vì tính di động, tiện lợi và nhanh chóng. Bên cạnh những ưu điểm đó, những hình ảnh này vẫn có các khuyết điểm như tài liệu chụp bị biến dạng (ví dụ như chụp những quyển sách dày), nghiêng và độ sáng không đồng đều. Những yếu tố này là thách thức lớn cho phương pháp nhận dạng truyền thống vốn chỉ hoạt động hiệu quả trên các hình ảnh tài liệu có chất lượng cao.

Để nội dung trong các ảnh chụp tài liệu bằng thiết bị di động được nhận dạng với độ chính xác cao hơn, nhóm nghiên cứu đề xuất các phương pháp tiền xử lý hình ảnh nhằm nâng cao chất lượng trước khi tiến hành nhận dạng. Các phương pháp

này gồm: nhị phân hóa, điều chỉnh độ nghiêng và làm phẳng hình ảnh tài liệu.

Các phương pháp này được thử nghiệm trên tập dữ liệu The IUPR Dataset of Camera-Captured Document Images [1] gồm các hình ảnh tài liệu chụp bằng máy ảnh kỹ thuật số, sau đó nhận dạng bằng công cụ Tesseract 5.0 cho kết quả có tỷ lệ lỗi ký tự trung bình là 6.72%, giảm gần 10 lần so với nhận dạng hình ảnh chưa qua xử lý. Kết quả này quan trọng vì đã làm giảm đáng kể tỷ lệ lỗi, nâng cao độ chính xác của các hình ảnh chụp tài liệu khi được nhận dạng bằng các công cụ nhận dạng ký tự quang học.

Phần còn lại của bài báo sẽ theo cấu trúc như sau. Trong Phần II, chúng tôi giới thiệu các công trình nghiên cứu liên quan trong lĩnh vực cải thiện kết quả nhận dạng ký tự quang học. Phần III và Phần IV lần lượt trình bày về khái niệm nhận dạng ký tự quang học và mục tiêu của việc tiền xử lý dữ liệu. Trong phần V, chúng tôi trình bày kỹ thuật nhị phân hóa hình ảnh nhằm phân tách nội dung văn bản và nền. Phần VI mô tả phương pháp xoay hình ảnh để điều chỉnh độ nghiêng. Phương pháp xử lý cuối cùng là làm phẳng hình ảnh được trình bày trong Phần VII. Phần VIII được dành cho kết quả thử nghiệm và đánh giá các phương pháp đề xuất. Cuối cùng, chúng tôi đưa ra kết luận và định hướng nghiên cứu ở tương lai trong phần IX.

## II. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Các nghiên cứu về cải thiện kết quả nhận dạng ký tự quang học đã được công bố trong nhiều năm qua. Một số hướng tiếp cận cải thiện kết quả bằng cách sửa lỗi văn bản có được sau khi nhận dạng. Trong [2] Karen Kukich đưa ra phương pháp phát hiện lỗi sai bằng chuỗi liên tiếp các từ trong văn bản (n-gram) hay từ điển và thay thế chúng bằng các từ gần giống có trong từ điển bằng cách sử dụng phương pháp thống kê. Còn Xiang Tong và David A.Evans [3] đề xuất phương pháp sử dụng ngữ cảnh của chính văn bản để sửa các từ bị nhận dạng sai dựa trên mô hình ngôn ngữ thống kê. Hay gần đây, Jie Mei và các cộng sự [4] sử dụng các đặc điểm phong phú về từ vựng, ngữ nghĩa và ngữ cảnh để phát hiện lỗi sai và tìm từ thay thế bằng kho ngữ liệu Google Web 1T và tiêu đề bài viết Wikipedia tiếng Anh. Tuy nhiên, những phương pháp này phụ thuộc vào ngôn ngữ và

từ điển nên có thể không áp dụng được trên các ngôn ngữ khác. Một phương pháp khác là kết hợp nhiều kết quả nhận dạng để tự phát hiện và sửa lỗi mà không phụ thuộc vào ngôn ngữ [5], [6].

Một số hướng tiếp cận khác cải thiện kết quả bằng cách xử lý dữ liệu trước khi tiến hành nhận dạng. Trong [7] Hirobumi Nishida đề xuất phương pháp khôi phục độ phân giải cao cho hình ảnh bằng nội suy kết hợp với bổ sung các nét còn thiếu của ký tự dựa trên đặc điểm địa hình, phương pháp này đặc biệt hiệu quả với các ký tự có cấu trúc phức tạp như chữ Kanji. Wojciech Bieniecki và các cộng sự [8] đề xuất phương pháp điều chỉnh độ nghiêng và làm phẳng tài liệu dựa trên tọa độ của các dòng văn bản. Ngoài ra, còn có một số phương pháp khác như loại bỏ phần viền gây nhiễu bên ngoài khung trang, loại bỏ phông nền hay phân chia và phân loại tài liệu thành nhiều phần nhỏ [9]–[11]. Một phương pháp gần đây là sử dụng mạng thần kinh tích chập (CNN) tự động lựa chọn các phương pháp xử lý dựa trên chất lượng hình ảnh [12].

### III. NHẬN DẠNG KÝ TỰ QUANG HỌC

Nhận dạng ký tự quang học - Optical Character Recognition (OCR) là việc sử dụng công nghệ để tự động nhận dạng các ký tự văn bản in hoặc viết tay trong hình ảnh kỹ thuật số của tài liệu vật lý, chẳng hạn như tài liệu giấy được quét bằng máy scan và chuyển đổi thành dữ liệu văn bản mà máy tính có thể đọc được.

Nhận dạng ký tự quang học có hai ưu điểm chính đó là khả năng tăng năng suất chuyển đổi vì được thực hiện tự động và khả năng lưu trữ văn bản hiệu quả, do đó, công nghệ này được áp dụng trong hầu hết các lĩnh vực từ tài chính, giáo dục đến cả cơ quan chính phủ. Các ứng dụng phổ biến của nó là xây dựng thư viện kỹ thuật số (lưu trữ văn bản, công thức toán học, bảng nhạc, ...), nhận dạng các đối tượng trên bản đồ số, công cụ đọc văn bản cho người khiếm thị, nhận dạng tài liệu viết tay, ...

Nhận dạng ký tự quang học gồm những quá trình cơ bản như sau:

- Tiền xử lý dữ liệu, quá trình này cải thiện dữ liệu hình ảnh đầu vào nhằm nâng cao tỷ lệ nhận dạng chính xác. Qua đó, các biến dạng không mong muốn được loại bỏ và các đặc điểm hình ảnh được thay đổi phù hợp cho việc nhận dạng.
- Nhận dạng ký tự, có hai phương pháp cơ bản được sử dụng là so khớp ma trận (Matrix Matching) và trích xuất đặc trưng (Feature Extraction). Trong đó, so khớp ma trận là phương pháp đơn giản, phổ biến hơn. Hai phương pháp này được mô tả như sau [13]:

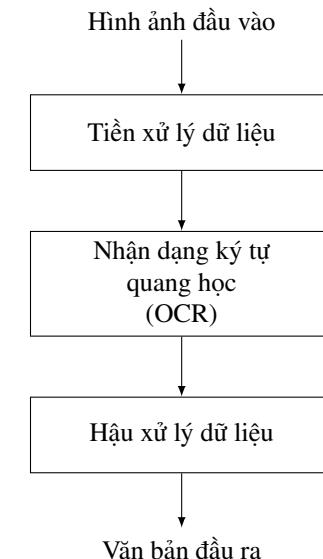
So khớp ma trận so sánh những gì mà hệ thống OCR nhìn thấy (xem như một ký tự) với thư viện ma trận ký tự hoặc các mẫu quy định. Khi một hình ảnh khớp với một trong các ma trận chấm theo quy định trong một mức độ giống nhau nhất định, máy tính sẽ gắn nhãn hình ảnh đó là ký tự ASCII tương ứng.

Trích xuất đặc trưng là phương pháp nhận dạng ký tự quang học mà không cần khớp nghiêm ngặt với các mẫu quy định. Trong phương pháp này, máy tính tìm kiếm các đặc điểm chung như vùng mở, hình khép kín, đường chéo,

giao điểm đường thẳng,... Phương pháp này linh hoạt hơn nhiều so với so khớp ma trận ma trận.

- Hậu xử lý dữ liệu, đây là quá trình cải thiện văn bản sau khi nhận dạng nhằm tăng độ chính xác, các phương pháp thường được áp dụng là thống kê, sử dụng ngữ cảnh văn bản hay kết hợp nhiều văn bản nhận dạng nhằm phát hiện và sửa chữa lỗi sai.

Các quá trình nhận dạng ký tự quang học cơ bản được thể hiện trong Hình 1.



Hình 1. Quá trình nhận dạng ký tự quang học

### IV. MỤC TIÊU CỦA TIỀN XỬ LÝ DỮ LIỆU

Wojciech Bieniecki đã phân tích sự cần thiết của quá trình tiền xử lý ảnh trong nhận dạng ký tự quang học bằng cách đo độ chính xác khi nhận dạng của phần mềm FineReader 7.0 với hình ảnh một trang báo được chỉnh sửa độ nhiễu hạt, độ phân giải, độ biến dạng và độ sáng khác nhau [14].

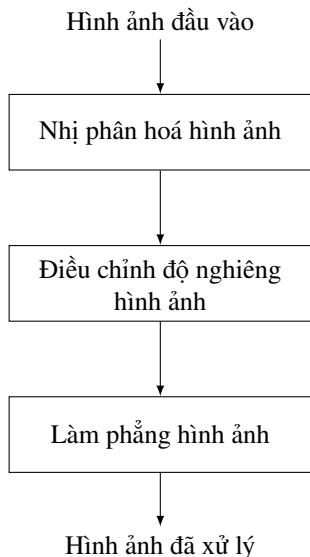
Thí nghiệm đã rút ra các kết luận như sau:

- Độ phân giải hình ảnh hưởng đáng kể đến độ chính xác, đặc biệt khi dưới 300 DPI. Nhưng số lỗi sai có thể giảm đáng kể nếu những hình ảnh này được tăng độ phân giải bằng thuật toán.
- Độ nhiễu hạt có tác động lớn đến chất lượng kết quả nhận dạng, nhưng với độ nhiễu thấp (10%) và độ phân giải cao (600 DPI) thì độ chính xác không bị giảm nhiều.
- Độ biến dạng hình học có ảnh hưởng lớn nhất đến độ chính xác. Những biến dạng này làm các dòng văn bản không được phân tách đúng, việc nhận dạng hoàn toàn thất bại nếu hình ảnh chỉ xoay đi 10 độ.
- Độ sáng không đều ít ảnh hưởng đến việc phân tách vùng văn bản lẫn việc nhận dạng ký tự.

Độ biến dạng cao, độ sáng không đồng đều là đặc điểm thường thấy ở các hình ảnh chụp bởi máy ảnh kỹ thuật số hay điện thoại thông minh do khoảng cách giữa tài liệu và thiết bị,

trong khi độ phân giải và độ nhiễu hạt thường ở mức phù hợp cho việc nhận dạng. Vì vậy việc hiệu chỉnh, làm giảm độ biến dạng của các hình ảnh loại này là quan trọng nhất, đây cũng là điều nghiên cứu của chúng tôi tập trung vào.

Quá trình tiền xử lý dữ liệu mà chúng tôi đề xuất được thể hiện trong Hình 2.



Hình 2. Quá trình tiền xử lý dữ liệu

## V. NHỊ PHÂN HOÁ HÌNH ẢNH

Với ảnh tài liệu được chụp từ thiết bị di động đã được xử lý sang ảnh thang độ xám, chúng tôi nhận thấy rằng ngoài các điểm ảnh trắng và đen còn có các điểm ảnh với nhiều mức độ xám khác nhau. Điều này làm cho nền và văn bản không được phân tách rõ ràng. Do đó, việc phân tách nền và văn bản là nhiệm vụ cần thiết phải làm trước tiên trong quá trình tiền xử lý. Vì vậy, chúng tôi đề xuất phương pháp nhị phân hóa hình ảnh sử dụng quy trình phân ngưỡng nhằm giải quyết vấn đề này.

Fương pháp nhị phân hóa hình ảnh tạo ra các phiên bản nhị phân (đen và trắng) của ảnh gốc ban đầu. Trong đó điểm ảnh đen được xem là văn bản và điểm ảnh trắng được xem là nền tài liệu. Bằng cách này, nền và văn bản của tài liệu sẽ được phân tách một cách rõ ràng.

### A. Phân ngưỡng toàn cục (Global thresholding)

Một trong những phương pháp đơn giản nhất để nhị phân hóa hình ảnh là sử dụng Phân ngưỡng toàn cục. Trong phương pháp này, một giá trị ngưỡng được chọn và áp dụng cho toàn bộ ảnh tài liệu. Mỗi điểm ảnh được đặt thành màu đen nếu nó nhỏ hơn giá trị ngưỡng, hoặc được đặt thành màu trắng nếu nó lớn hơn giá trị ngưỡng.

Có một vấn đề là độ sáng cũng như độ tương phản của ảnh chụp rất đa dạng, nên việc cố định một giá trị ngưỡng thường dẫn đến nhiều vấn đề khác nhau. Để khắc phục điểm yếu này, phương pháp Otsu [15] thường được dùng để xác định giá trị ngưỡng cho từng ảnh. Phương pháp này hoạt động tương đối ổn định trên tài liệu được chụp một cách rõ ràng.

### B. Phân ngưỡng cục bộ (Local thresholding)

Trong thực tế, ảnh chụp thường xuyên gặp các vấn đề về độ sáng. Các vấn đề này thường là ánh sáng không đồng đều, bị hắt sáng,... do môi trường chụp không đảm bảo. Do đó, việc sử dụng một giá trị ngưỡng cho toàn bộ ảnh như phân ngưỡng toàn cục có thể làm cho nền và văn bản bị lẫn vào nhau ở một phần hay toàn bộ hình ảnh. Đây là một vấn đề lớn cần khắc phục vì nó có thể làm mất một phần văn bản trong tài liệu, điều này có thể nhìn thấy rõ ở Hình 3.

Để giải quyết vấn đề trên, chúng tôi quyết định sử dụng Phân ngưỡng cục bộ [16]. Giá trị ngưỡng của từng điểm ảnh không được lấy toàn cục mà sẽ phụ thuộc vào một phần nhỏ của ảnh xung quanh điểm ảnh đang xét. Vì thế sẽ có nhiều giá trị ngưỡng khác nhau trên từng vùng ảnh khác nhau trên cùng một ảnh. Điều này giúp kết quả nhị phân hóa tốt hơn rất nhiều đối với ảnh có ánh sáng không đồng nhất, điều này được minh họa ở Hình 3e.

- no prior: due to the missing prior, forged text-lines are considered to be as frequent as genuine ones. It is, however, more likely that genuine text-lines are much more frequent than forged ones.
- text-line length: for longer text-lines, the skew angle can be measured much more accurately than for shorter ones. Also, the skew angles of shorter lines are more sensitive to noise. In an extreme case, for a line with a length of 100 px consisting of two connected components, one supplemental pixel at the bottom of one component would change the angle by at least  $\pm 0.57^\circ$ .

(a) Hình ảnh đầu vào

- no prior: due to the missing prior, forged text-lines are considered to be as frequent as genuine ones. It is, however, more likely that genuine text-lines are much more frequent than forged ones.
- text-line length: for longer text-lines, the skew angle can be measured much more accurately than for shorter ones. Also, the skew angles of shorter lines are more sensitive to noise. In an extreme case, for a line with a length of 100 px consisting of two connected components, one supplemental pixel at the bottom of one component would change the angle by at least  $\pm 0.57^\circ$ .

(b) Phân ngưỡng toàn cục với giá trị ngưỡng = 63

- no prior: due to the missing prior, forged text-lines are considered to be as frequent as genuine ones. It is, however, more likely that genuine text-lines are much more frequent than forged ones.
- text-line length: for longer text-lines, the skew angle can be measured much more accurately than for shorter ones. Also, the skew angles of shorter lines are more sensitive to noise. In an extreme case, for a line with a length of 100 px consisting of two connected components, one supplemental pixel at the bottom of one component would change the angle by at least  $\pm 0.57^\circ$ .

(c) Phân ngưỡng toàn cục với giá trị ngưỡng = 95

- no prior: due to the missing prior, forged text-lines are considered to be as frequent as genuine ones. It is, however, more likely that genuine text-lines are much more frequent than forged ones.
- text-line length: for longer text-lines, the skew angle can be measured much more accurately than for shorter ones. Also, the skew angles of shorter lines are more sensitive to noise. In an extreme case, for a line with a length of 100 px consisting of two connected components, one supplemental pixel at the bottom of one component would change the angle by at least  $\pm 0.57^\circ$ .

(d) Phân ngưỡng toàn cục với giá trị ngưỡng = 127

- no prior: due to the missing prior, forged text-lines are considered to be as frequent as genuine ones. It is, however, more likely that genuine text-lines are much more frequent than forged ones.
- text-line length: for longer text-lines, the skew angle can be measured much more accurately than for shorter ones. Also, the skew angles of shorter lines are more sensitive to noise. In an extreme case, for a line with a length of 100 px consisting of two connected components, one supplemental pixel at the bottom of one component would change the angle by at least  $\pm 0.57^\circ$ .

(e) Phân ngưỡng cục bộ

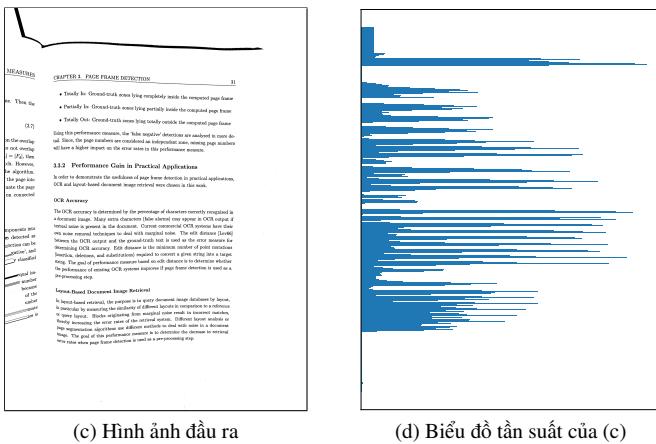
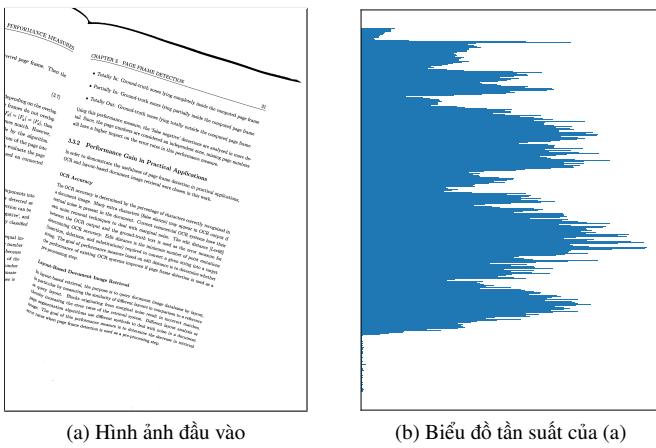
Hình 3. Nhị phân hóa hình ảnh

## VI. ĐIỀU CHỈNH ĐỘ NGHIÊNG HÌNH ẢNH

Trong thực tế khi chụp tài liệu bằng thiết bị di động, có một vấn đề rất phổ biến và khó lòng tránh khỏi là hướng của tài liệu được chụp không phù hợp để tiến hành nhận dạng. Vì vậy, việc loại bỏ độ nghiêng của hình ảnh trước khi nhận dạng là bước bắt buộc trong quá trình tiền xử lý nhằm làm cho các dòng văn bản phải song song với viền của hình ảnh.

Nhằm thực hiện công việc trên, chúng tôi tìm góc nghiêng của tài liệu so với trực tung và dựa trên góc nghiêng này để xoay ảnh tài liệu về đúng vị trí. Để tìm ra góc nghiêng của tài liệu, chúng tôi nêu ra hai phương pháp: biểu đồ tần suất và quang phổ hình ảnh.

### A. Phương pháp biểu đồ tần suất



Hình 4. Hình ảnh và biểu đồ tần suất tương ứng

Với ảnh tài liệu sau khi được nhị phân hóa, mỗi điểm ảnh chỉ có thể thuộc một trong hai màu: đen hoặc trắng. Biểu đồ tần suất là biểu đồ mà mỗi cột của nó là tổng số lần xuất hiện các điểm ảnh đen trên một hàng ngang tương ứng.

Nếu ảnh tài liệu không nghiêng, do sự đan xen giữa các dòng văn bản và khoảng trắng giữa dòng, nên biểu đồ có các cột rất cao (tương ứng với các dòng văn bản) đan xen với các cột rất thấp (tương ứng với khoảng trắng). Nhìn chung biểu đồ có sự biến thiên rất lớn khi so với trường hợp ảnh nghiêng, điều này

được dễ dàng nhìn thấy trong Hình 4. Dựa vào nhận xét trên, chúng tôi sử dụng phương pháp vét cạn [17]. Bằng cách xoay hình ảnh với các góc khác nhau, trong mỗi trường hợp, chúng tôi đánh giá tổng độ biến thiên của biểu đồ bằng công thức (1)

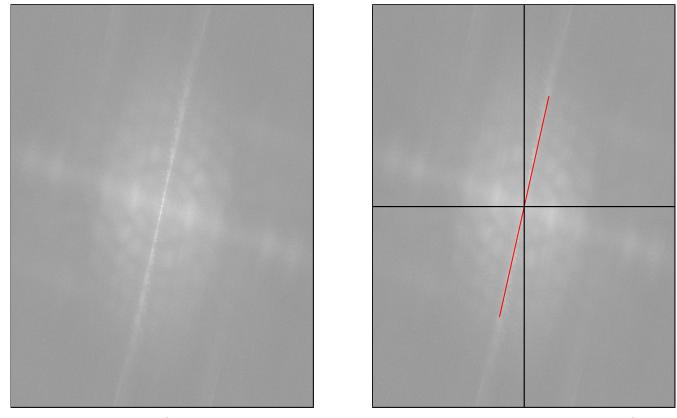
$$S = \sum_{i=1}^n (H_i - H_{i-1})^2 \quad (1)$$

với  $H_i$  là giá trị của biểu đồ tại cột  $i$ . Với việc chọn góc xoay có giá trị  $S$  tương ứng lớn nhất, thì đây có thể được xem là vị trí đúng của ảnh tài liệu.

### B. Phương pháp quang phổ hình ảnh (Image spectrum)

Một trong những phương pháp để tìm ra góc nghiêng của tài liệu là sử dụng quang phổ hình ảnh. Để tìm được quang phổ hình ảnh, chúng tôi sử dụng 2D Discrete Fourier Transform (DFT) để tìm miền tần số. Fast Fourier Transform [18] là một thuật toán nhanh để tính toán DFT. Từ quang phổ hình ảnh, chúng ta có thể dễ dàng tìm được góc lệch của ảnh.

Điểm sáng nhất trong quang phổ chỉ ra hai trục vuông góc tương tự như hai trục trong hệ trục toạ độ Oxy. Để tăng độ chính xác cho việc nhận dạng, chúng tôi chia quang phổ thành bốn phần tương ứng bốn phần của hệ trục toạ độ. Dựa trên bốn phần hình ảnh và thuật toán nhận dạng đường thẳng phù hợp, góc giữa đường thẳng và trực tung sẽ được chỉ ra và đây cũng chính là góc lệch của tài liệu. Hình 5 minh họa quang phổ hình ảnh và góc lệch tương ứng.

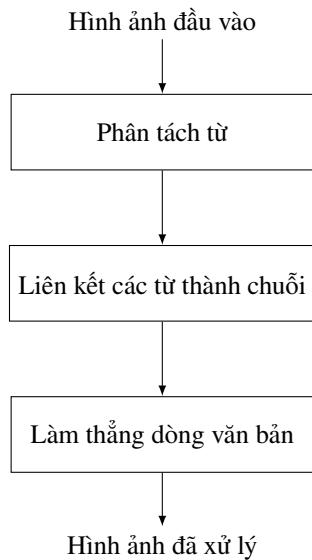


Hình 5. Quang phổ và độ nghiêng của Hình 4a

## VII. LÀM PHẲNG HÌNH ẢNH

Trải qua hai phương pháp tiền xử lý được trình bày trong Phần V và VI, kết quả OCR đã được cải thiện phần nào nhưng vẫn còn tồn tại rất nhiều lỗi nhận dạng do các nguyên nhân khác nhau. Sau quá trình tìm hiểu, chúng tôi nhận ra nguyên nhân chính là do các dòng văn bản bị biến dạng, điều này không thể tránh khỏi do tài liệu được chụp từ thiết bị di động. Ngoài ra còn những nguyên nhân khác quan như sự biến dạng trang sách trong một quyển sách dày cũng có thể làm cong dòng văn bản. Điều này làm cho việc nhận dạng thứ tự của các từ trong câu không chính xác. Song song đó, dòng văn bản bị biến dạng

cũng làm các kí tự biến dạng theo, làm gia tăng khó khăn cho hệ thống nhận dạng ký tự quang học. Do đó, chúng tôi đề xuất một phương pháp làm phẳng hình ảnh đơn giản trải qua các bước được thể hiện như Hình 6.



Hình 6. Quá trình làm phẳng hình ảnh

#### A. Phân tách từ

Trong vấn đề làm phẳng hình ảnh, việc đầu tiên phải làm là phân tách các từ để sẵn sàng cho các bước tiếp theo. Để thực hiện được công việc trên, chúng tôi sử dụng thuật toán Border Following [19] được cài đặt bởi OpenCV. Trong quá trình sử dụng thuật toán trên, chúng tôi nhận thấy có một vài vấn đề ảnh hưởng đến kết quả phân tách từ. Đầu tiên là các chữ cái, dấu câu có liên quan không được nhận dạng chính xác là một từ. Nhằm giải quyết vấn đề này, chúng tôi sử dụng phương pháp giãn nở (dilation) [20] để liên kết các kí tự, dấu câu gần nhau thành một thành phần gắn kết và được nhận dạng là một từ.

Ngoài ra, để cải thiện kết quả phân tách từ, chúng tôi còn thực hiện loại bỏ các thành phần có diện tích rất nhỏ hoặc rất lớn, cũng như các thành phần có chiều dài hoặc chiều rộng rất lớn. Việc loại bỏ như trên nhằm mục đích lọc nhiễu sinh ra từ việc chụp tài liệu từ các thiết bị di động, cũng như loại bỏ những thành phần không phải là từ như hình ảnh, biểu đồ, đường kẻ,...

#### B. Liên kết các từ thành chuỗi

Để liên kết các từ thành một chuỗi, chúng tôi đề xuất phương pháp tìm kiếm liền kề. Phương pháp này được thực hiện bằng cách tìm kiếm và liên kết các từ gần với nhau theo một điều kiện nhất định thành một chuỗi, các chuỗi này chính là các dòng trong văn bản.

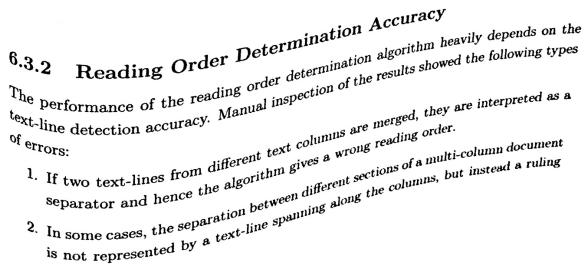
Trong quá trình tìm kiếm và liên kết các từ gần nhau, chúng tôi xét từng từ và tìm kiếm hàng xóm gần nhất của từ đó. Công việc này gồm hai trường hợp xảy ra:

- Từ đang xét chưa thuộc chuỗi nào. Trong trường hợp này, chúng tôi ước lượng hướng của từ là một hàm bậc nhất -

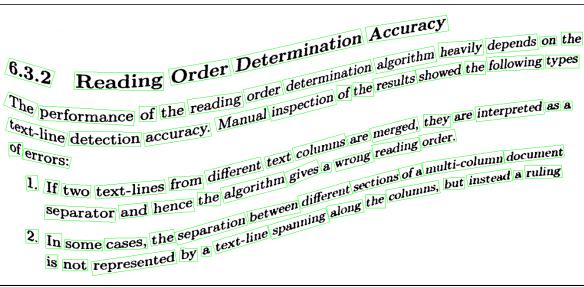
chính là trục chính của từ. Hàm này được xác định bằng moment [21] của từ hiện tại.

- Từ đang xét đang thuộc về một chuỗi. Trong trường hợp này, hàm ước lượng hướng của từ cũng chính là hàm ước lượng của chuỗi. Hàm này phải là một hàm trơn và đi qua tất cả các từ của chuỗi ấy. Trong trường hợp này, chúng tôi sử dụng hàm nội suy Spline [22].

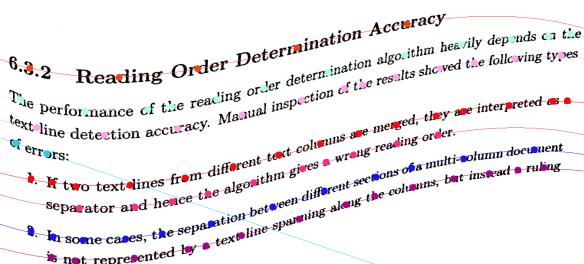
Từ từ hiện tại đang xét, chúng tôi đi dọc hàm ước lượng theo hướng bên phải cho đến khi gặp một từ mới. Đây chính là hàng xóm của từ đang xét. Nếu một từ là hàng xóm của hai hay nhiều từ, thì chúng tôi quyết định chọn từ có khoảng cách ngắn nhất.



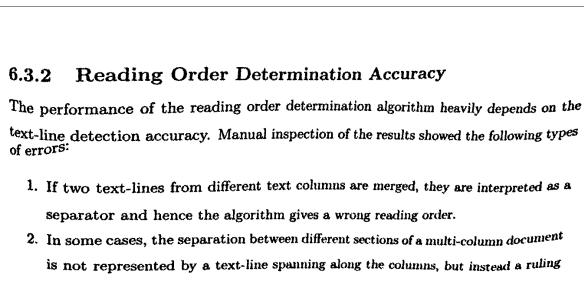
(a) Hình ảnh đầu vào



(b) Phân tách từ



(c) Liên kết các từ thành chuỗi



(d) Làm thẳng dòng văn bản

Hình 7. Minh họa quá trình làm phẳng hình ảnh

### C. Làm thẳng dòng văn bản

Sau khi đã tìm được chuỗi cùng với hàm ước lượng, chúng tôi xác định vị trí của dòng mong đợi bằng phương pháp như sau:

- 1) Tìm vị trí đầu tiên của dòng. Chúng tôi chọn vị trí này bằng cách đi dọc theo hàm ước lượng và lấy điểm có vị trí trái nhất của chuỗi.
- 2) Vạch ra dòng cơ sở (base line). Từ vị trí đầu tiên của dòng, chúng tôi vạch ra một đường thẳng theo phương ngang và xác định đây là vị trí mong đợi của dòng văn bản.
- 3) Di chuyển các từ về vị trí mong đợi. Từ dòng cơ sở và hàm ước lượng, chúng tôi tính toán độ chênh lệch và di chuyển từng điểm ảnh của các từ thuộc dòng ấy về dòng cơ sở.

Quá trình làm phẳng hình ảnh được minh họa trong Hình 7.

### VIII. KẾT QUẢ

Để đánh giá kết quả nhận dạng ký tự quang học, chúng tôi dựa vào tỷ lệ lỗi ký tự (character error rate - CER) của văn bản sau khi nhận dạng và so sánh những thông số này giữa việc nhận dạng trên tập dữ liệu gốc và tập dữ liệu sau khi áp dụng phương pháp tiền xử lý.

Công thức tính tỷ lệ lỗi ký tự [23] được định nghĩa như sau:

$$CER = \frac{i + s + d}{n}$$

Với  $n$  là tổng số lượng ký tự trong văn bản nguồn;  $i, s, d$  lần lượt là số thao tác thêm, thay thế, xóa một ký tự cần thiết để chuyển đổi thành văn bản đích (khoảng cách Levenshtein).

Chúng tôi chứng minh tính hiệu quả của các phương pháp được đề xuất để cải thiện chất lượng hình ảnh và độ chính xác OCR bằng cách thử nghiệm trên tập dữ liệu The IUPR Dataset of Camera-Captured Document Images [1] gồm 5 trong tổng số 100 ảnh chụp trắng đen (grayscale) tài liệu bằng máy ảnh kỹ thuật số ở định dạng PNG. Sau khi xử lý bằng phương pháp đề xuất, hình ảnh được nhận dạng bằng công cụ Tesseract 5.0 và tính toán CER dựa trên văn bản đúng (ground truth) kèm theo trong tập dữ liệu.

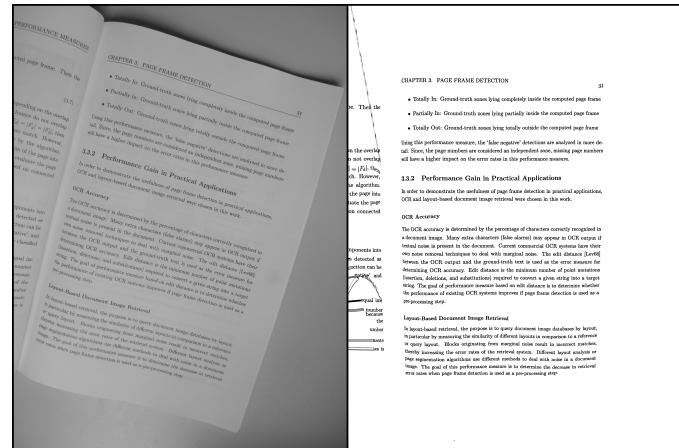
Hình 8 cho thấy hình ảnh mới của tài liệu được tạo ra sau khi áp dụng ba phương pháp xử lý đề xuất, cùng với hình ảnh gốc của tài liệu bên trái.

Kết quả thử nghiệm đánh giá tỷ lệ lỗi ký tự sau khi áp dụng các phương pháp được mô tả trong bảng I dưới đây.

Bảng I  
SO SÁNH TỶ LỆ LỖI KÝ TỰ (CER)

Phương pháp xử lý	Tỷ lệ lỗi ký tự trung bình (Avg CER)
Không	66.88%
Nhi phân hóa hình ảnh	53.09%
Điều chỉnh độ nghiêng hình ảnh	43.43%
Làm phẳng hình ảnh	25.95%
Cả ba phương pháp	6.72%

Từ bảng này, chúng tôi có thể kết luận rằng các phương pháp được trình bày trong bài báo rất hiệu quả trong việc cải thiện độ chính xác của nhận dạng ký tự quang học trong tất cả các trường hợp được xử lý. Tỷ lệ lỗi ký tự trung bình đã giảm từ 66.88% chỉ còn 6.72% sau khi áp dụng cả ba phương pháp mà chúng tôi đề xuất.



Hình 8. Hình ảnh tài liệu trước và sau khi xử lý bằng ba phương pháp

### IX. KẾT LUẬN

Trong bài báo này, chúng tôi đã chỉ ra ba phương pháp tiền xử lý hình ảnh quan trọng nhằm cải thiện kết quả nhận dạng ký tự quang học từ hình ảnh tài liệu. Văn bản trong hình ảnh được xác định chính xác hơn bằng phương pháp nhị phân hóa, tình trạng nhận dạng lệch dòng văn bản hầu như không còn nhờ phương pháp điều chỉnh độ nghiêng và làm phẳng hình ảnh tài liệu, các phương pháp này đã làm giảm tỷ lệ lỗi ký tự khi nhận dạng xuống thấp đáng kể.

Nghiên cứu này có ý nghĩa quan trọng vì các máy ảnh kỹ thuật số, điện thoại thông minh đang nhanh chóng trở thành nguồn thu thập thông tin hình ảnh chính, kể cả trong lĩnh vực quét hình ảnh văn bản. Nghiên cứu có thể áp dụng vào các lĩnh vực cần nhận dạng với độ chính xác cao như xây dựng thư viện số, lưu trữ văn bản hành chính, nhận dạng giấy tờ, ...

Trong tương lai, chúng tôi sẽ phát triển một phần mềm ứng dụng sử dụng các phương pháp trong bài báo này có chức năng xử lý hình ảnh tài liệu đưa vào nhằm nâng cao độ chính xác khi được nhận dạng bằng các công cụ nhận dạng ký tự quang học.

### LỜI CẢM ƠN

Chúng tôi xin trân trọng gửi lời cảm ơn đặc biệt sâu sắc và chân thành đến PGS.TS Vũ Hải Quân, PGS.TS Trần Minh Triết và Th.S. Nguyễn Thành An đã tận tình định hướng, giúp đỡ nhóm trong suốt quá trình nghiên cứu, là cơ sở để đề tài này được hoàn thành một cách tốt nhất.

### TÀI LIỆU

- [1] S. S. Bukhari, F. Shafait, and T. M. Breuel, "The IUPR dataset of camera-captured document images," in *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 2011, pp. 164–171.
- [2] K. Kukich, "Techniques for automatically correcting words in text," *Acm Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.
- [3] X. Tong and D. A. Evans, "A statistical approach to automatic OCR error correction in context," in *Fourth workshop on very large corpora*, 1996.
- [4] J. Mei, A. Islam, A. Moh'd, Y. Wu, and E. Milios, "Statistical learning for OCR error correction," *Information Processing & Management*, vol. 54, no. 6, pp. 874–887, 2018.
- [5] I. Z. Yalniz and R. Manmatha, "A fast alignment scheme for automatic OCR evaluation of books," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 754–758.

- [6] W. B. Lund, D. J. Kennard, and E. K. Ringger, “Combining multiple thresholding binarization values to improve OCR output,” in *Document Recognition and Retrieval XX*, vol. 8658. International Society for Optics and Photonics, 2013, p. 86580R.
- [7] H. Nishida, “Restoring high-resolution text images to improve legibility and OCR accuracy,” in *Document Recognition and Retrieval XII*, vol. 5676. International Society for Optics and Photonics, 2005, pp. 136–147.
- [8] W. Bieniecki, S. Grabowski, and W. Rozenberg, “Image preprocessing for improving OCR accuracy,” in *2007 international conference on perspective technologies and methods in MEMS design*. IEEE, 2007, pp. 75–80.
- [9] S. S. Bukhari, F. Shafait, and T. M. Breuel, “Border noise removal of camera-captured document images using page frame detection,” in *International Workshop on Camera-Based Document Analysis and Recognition*. Springer, 2011, pp. 126–137.
- [10] M. Shen and H. Lei, “Improving OCR performance with background image elimination,” in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2015, pp. 1566–1570.
- [11] Y. Yang, K. Summers, and M. Turner, “A text image enhancement system based on segmentation and classification methods,” in *Proceedings of the 1st ACM workshop on Hardcopy document processing*, 2004, pp. 33–40.
- [12] Q. A. Bui, D. Mollard, and S. Tabbone, “Selecting automatically preprocessing methods to improve OCR performances,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 169–174.
- [13] “OCR Introduction,” <http://www.dataid.com/aboutocr.htm>, accessed: 2021-06-01.
- [14] W. Bieniecki, “Analiza wymagań dla metod przetwarzania wstępnego obrazów w automatycznym rozpoznawaniu tekstu,” *Automatyka/Akademia Górnictwo-Hutnicza im. Stanisława Staszica w Krakowie*, vol. 9, pp. 525–532, 2005.
- [15] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [16] “Miscellaneous image transformations.” [Online]. Available: [https://docs.opencv.org/4.5.2/d7/d1b/group\\_imgproc\\_misc.html#ga72b913f352e4a1b1b397736707afcde3](https://docs.opencv.org/4.5.2/d7/d1b/group_imgproc_misc.html#ga72b913f352e4a1b1b397736707afcde3)
- [17] C. Anand, “Detect correct skew in images using python,” Nov 2016. [Online]. Available: <https://avilpage.com/2016/11/detect-correct-skew-images-python.html>
- [18] “Performance optimization of dft.” [Online]. Available: [https://docs.opencv.org/4.x/de/dbc/tutorial\\_py\\_fourier\\_transform.html](https://docs.opencv.org/4.x/de/dbc/tutorial_py_fourier_transform.html)
- [19] S. Suzuki *et al.*, “Topological structural analysis of digitized binary images by border following,” *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985.
- [20] “Image filtering.” [Online]. Available: [https://docs.opencv.org/3.4/d4/d86/group\\_imgproc\\_filter.html#ga4ff0f3318642c4f469d0e11f242f3b6c](https://docs.opencv.org/3.4/d4/d86/group_imgproc_filter.html#ga4ff0f3318642c4f469d0e11f242f3b6c)
- [21] “Structural analysis and shape descriptors.” [Online]. Available: [https://docs.opencv.org/3.4/d3/dc0/group\\_imgproc\\_shape.html#ga556a180f43cab22649c23ada36a8a139](https://docs.opencv.org/3.4/d3/dc0/group_imgproc_shape.html#ga556a180f43cab22649c23ada36a8a139)
- [22] “scipy.interpolate.univariate\_spline.” [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.UnivariateSpline.html>
- [23] “2.3 Computing error rates - Text Digitisation,” <https://sites.google.com/site/textdigitisation/qualitymeasures/computingerrorrates>, accessed: 2021-06-10.