XÂY DỰNG MÔ HÌNH KIỂM TRA ĐỐI CHIẾU DỮ LIỆU SỬ DỤNG NHẬN DẠNG KÝ TỰ QUANG HỌC

Nguyễn Bá Duy, Đinh Thành Nhân và Nguyễn Trung Kiên

Trường Đại học Kỹ thuật - Công nghệ Cần Thơ

Email: nbduy@ctuet.edu.vn

Thông tin chung:
Ngày nhận bài: 03.12.2023
Ngày nhận bài sửa:19.02.2024
Ngày duyệt đăng:20.02.2024

Từ khóa:
Thị giác máy tính, trích xuất thông tin ảnh, xử lý ảnh.

TÓM TẮT

Kiểm tra và đối chiếu thông tin trên văn bằng, chứng chỉ trước khi công khai là một nhiệm vụ quan trọng, trong đó việc đối chiếu thông tin từ bản scan của văn bằng, chứng chỉ với thông tin lưu trữ trong cơ sở dữ liệu là một giải pháp đơn giản hiệu quả. Trong nghiên cứu này, tác giả đề xuất mô hình cho phép trích xuất tự động thông tin từ văn bằng, chứng chỉ sử dụng kỹ thuật nhận dạng ký tự quang học để đối chiếu dữ liệu. Tác giả thực nghiệm mô hình trên tập dữ liệu gồm 200 chứng chỉ ứng dụng công nghệ thông tin của Trung tâm Ngoại ngữ - Tin học thuộc Trường Đại học Kỹ thuật - Công nghệ Cần Thơ, xây dựng một hệ thống đối chiếu dữ liệu tích hợp vào hệ thống tra cứu chứng chỉ của Trung tâm và kết quả thực nghiệm cho thấy hệ thống có thể số hóa và trích xuất thông tin với độ chính xác 89,72%. Dựa trên kết quả đạt được, tác giả đề xuất ứng dụng mô hình vào các hệ thống đối chiếu dữ liệu và đưa ra một số khuyến nghị cho các nghiên cứu tương tự trong tương lai.

1. ĐẶT VẤN ĐỀ

Cuộc cách mạng công nghiệp 4.0 và đặt biệt là công cuộc chuyển đổi số đang diễn ra mạnh mẽ trên toàn thế giới, do đó ứng dụng công nghệ thông tin nhằm nâng cao hiệu quả công tác có vai trò to lớn hơn bao giờ hết. Trong đó, việc trích xuất thông tin để so khớp dữ liệu, xác minh dữ liệu văn bằng, chứng chỉ là một trong những ứng dụng quan trọng. Thực tế, công tác này chưa được quan tâm nhiều, vì thế tác giả đề xuất mô hình kiểm tra, đối chiếu dữ liệu trên thông tin văn bằng, chứng chỉ nhằm nâng cao hiệu quả đối chiếu dữ liệu đã lưu trữ với thông tin được ghi trên văn bằng, chứng chỉ. Mô hình thực hiện so khớp, xác minh tài liệu bằng nhận dạng ký tự quang học (Optical Character Recognition - OCR). Nhận dạng ký tự quang học là việc sử dụng công nghệ để phân biệt các ký tự văn bản in ấn hoặc viết tay trong ảnh kỹ thuật số của tài liệu vật lý, như bản scan của tài liệu. Quy trình nhận dạng ký tự quang học cơ bản bao gồm phân tích văn bản của một tài liệu và phiên dịch các ký tự thành mã có thể sử dụng để xử lý dữ liệu. Phương pháp này đã được sử dụng rộng rãi như một hình thức nhập thông tin từ các bản ghi dữ liệu trên giấy và là một phương pháp phổ biến trong việc trích xuất dữ liệu văn bản từ tập tin.

Từ mô hình đề xuất, tác giả xây dựng hệ thống tích hợp vào trang tra cứu thông tin chứng chỉ ứng dụng công nghệ thông tin của Trung tâm Ngoại ngữ - Tin học thuộc trường Đại học Kỹ thuật - Công nghệ Cần

Thơ để đánh giá hiệu quả mô hình mang lại. Hệ thống tích hợp thực hiện trích xuất thông tin từ tập tin hình ảnh chứng chỉ ứng dụng công nghệ thông tin đã chọn để so khớp với thông tin đã được lưu trữ trên cơ sở dữ liệu và thực hiện cảnh báo cho người dùng tại các trường dữ liệu bị sai trước khi công khai thông tin.

2. PHƯƠNG PHÁP NGHIÊN CỨU

Trong những năm gần đây, thị giác máy tính đã có một số thành tựu trong lĩnh vực nhận dạng chữ viết in. Có rất nhiều hệ thống áp dụng thị giác máy tính để trích xuất thông tin và đạt được nhiều thành tựu.

Nhã và cộng sự (2022) xây dựng hệ thống eYKC, là phần mềm xác minh danh tính của khách hàng dựa vào hình chụp giấy tờ tùy thân, sử dụng mô hình YOLOv4 để phát hiện các trường thông tin từ giấy tờ tùy thân và đối chiếu hình ảnh người đăng ký bằng sách so khớp video hoặc hình ảnh chân dung với hình ảnh trong giấy tờ tuỳ thân.

Anand Shinde và cộng sự (2021) xây dựng hệ thống ALPR trích xuất thông tin bảng số xe hơi từ video bằng cách sử dụng Faster R-CNN để xác định bảng số xe, sử dụng Tractor được đề xuất bởi Bergmann và cộng sự (2019) để kiểm tra video có tốc độ khung hình cao nhằm xác định thông tin chính xác hơn, thực hiện nhận dạng biển số xe bằng cách cắt các bảng số thành các hộp chứa dữ liệu bảng số và trích xuất thông tin văn bản của các hộp. John Anthony C. Jose và cộng sự (2021) đề xuất mô hình ALPR để nhận diện bảng số xe hơi.

Website trích xuất thông tin văn bản từ hình ảnh được tải lên sử dụng Tesseract OCR engine được đề xuất bởi Anand Shinde và cộng sự (2021) thực hiện trích xuất thông tin bằng cách lọc nhiễu, làm mịn, chỉnh lại văn bản đúng vị trí chuẩn, chuyển sang ảnh xám từ ảnh đầu vào. Sau đó xác định các từ, các dòng và các ký tự trong ảnh, so sánh danh mục các ký tự tương đồng nhất với các ký tự đã xác định, sử dụng kết hợp từ điển và các cấu trúc ngữ pháp để nhận dạng đúng các ký tự, cuối cùng hiển thị nội dung văn bản trong hộp văn bản của website.

Bài báo trích xuất và xác định văn bản dựa trên OCR do Anshul Arora và cộng sự (2021) đã giới thiệu quy trình và các kỹ thuật để trích xuất văn bản từ ảnh, quy trình gồm 3 bước: Bước 1- Tiền xử lý (chỉnh lại văn bản đúng vị trí chuẩn, loại bỏ nhòe, chuyển thành ảnh nhị phân, xóa các dòng đóng khung). Bước 2- Nhận dạng văn bản (đề xuất sử dụng phương pháp đối sánh mẫu, nhận dạng mẫu/tương quan ảnh). Bước 3- Hậu xử lý (đề xuất phương pháp dựa trên kết cấu, phương pháp dựa trên vùng).

Hệ thống trích xuất văn bản từ ảnh sử dụng OCR của Meredita Susanty và Herminarto Nugroho (2020) đề xuất giải pháp tăng độ chính xác trích xuất văn bản của OCR ảnh đầu vào: Đầu tiên, xóa QR Code và ảnh thí sinh bằng cách cộng ảnh đầu vào với 1 ảnh đen có khung trắng tại vị trí QR Code và ảnh thí sinh. Sau đó, sử dụng phân ngưỡng nhị phân với T là 185 để chuyển ảnh đầu vào thành ảnh nhị phân có nền đen và chữ trắng. Cuối cùng sử dụng Tesseract để trích xuất nội dung văn bản. Nghiên cứu chỉ rằng Tesseract trích xuất văn bản từ ảnh nền đen chữ trắng tốt hơn ảnh nền trắng chữ đen.

Bài viết so sánh mô hình tổng hợp thích ứng và mô hình tổng hợp theo trọng số do G. Li và N. Li (2019) cùng sử dụng CNN để phân lớp đã bổ sung kỹ thuật OCR trong bước tiền xử lý dữ liệu văn bản theo Chandra và

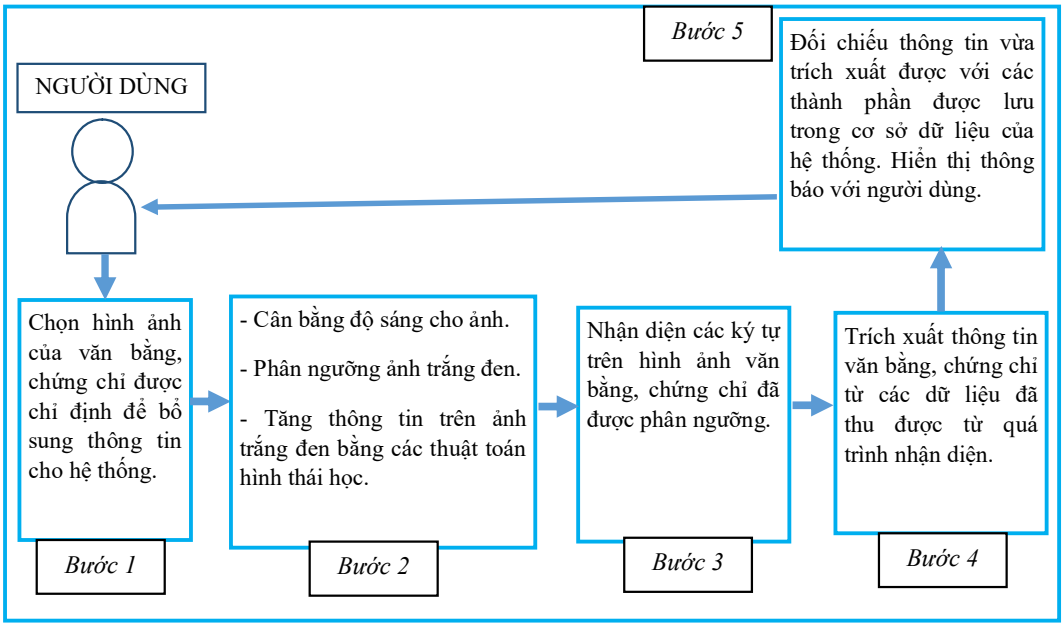
cộng sự (2020) đã đề xuất nhằm nâng cao hiệu quả phân lớp của 2 mô hình trên.

Tuy nhiên, việc áp dụng thị giác máy tính để xác nhận thông tin văn bằng, chứng chỉ được cấp cho người học hiện tại vẫn chưa được quan tâm nhiều. Do đó, chúng tôi đề xuất mô hình trích xuất và so khớp dữ liệu. Nghiên cứu này thực hiện theo một chuỗi các hoạt động bao gồm: thu thập dữ liệu, xây dựng mô hình trích xuất và cuối cùng là thử nghiệm đánh giá mô hình.

3. KẾT QUẢ

3.1. Mô hình trích xuất và so khớp dữ liệu

Trong phần này, tác giả đề xuất mô hình trích xuất thông tin tự động từ văn bằng, chứng chỉ, mô hình được thực hiện qua 5 Bước, cụ thể như Hình 1.

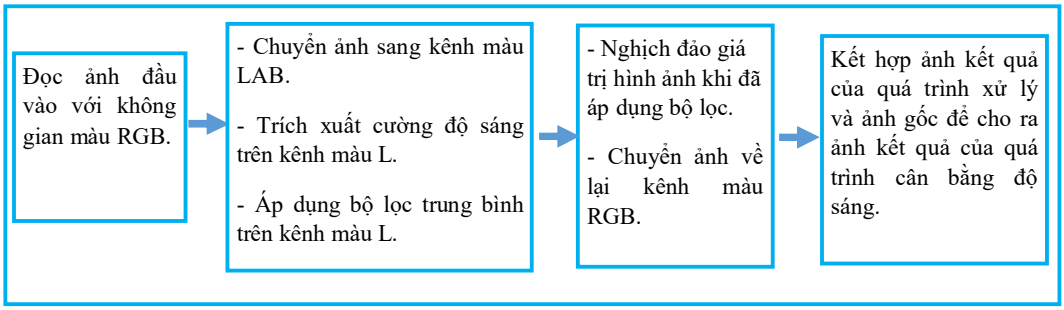


Hình 1. Kiến trúc tổng thể của hệ thống

Nguồn: Công bố của tác giả, (2023).

Đầu tiên, tại Bước 1, tác giả chọn tập tin văn bằng, chứng chỉ tương ứng với thông tin văn bằng, chứng chỉ được lưu trên cơ sở dữ liệu. Mẫu văn bằng, chứng chỉ được thường được cơ quan chức năng ban hành thống nhất đối với các cơ sở đào tạo, do đó bố cục và thông tin trên văn bằng, chứng chỉ cấp cho người học là thống nhất. Đây là một điểm thuận lợi khi trích xuất thông tin trên văn bằng, chứng chỉ.

Đối với Bước 2, tác giả áp dụng một thuật toán để cân bằng độ sáng cho ảnh, để hạn chế đến mức thấp nhất việc mất thông tin ảnh trong quá trình phân ngưỡng. Để giải quyết vấn đề này, tác giả đã áp dụng một phương pháp được trình bày bởi P. Bergmann và cộng sự (2019) để cải thiện độ sáng cho các vùng dữ liệu trên hình ảnh văn bằng, chứng chỉ. Ý tưởng chính của phương pháp này có thể được mô tả ngắn gọn như Hình 2.



Hình 2. Quá trình cân bằng độ sáng của ảnh

Nguồn: Công bố của tác giả, (2023).

Sau khi hệ thống nhận về được kết quả của quá trình cân bằng độ sáng của hình ảnh. Thuật toán đề xuất bởi Otsu sẽ được hệ thống áp dụng để hỗ trợ cho quá trình phân ngưỡng ảnh trắng đen. Thuật toán Otsu được đưa ra bởi Nobuyuki Otsu (1979), với mục đích tìm ra một ngưỡng "T" bằng cách tính toán tự động dựa trên các giá trị mức xám của từng điểm ảnh như Xiangyang Xu và cộng sự (2011). Ngưỡng "T" được sử dụng cho quá trình phân đoạn ảnh trắng đen thay cho các ngưỡng cố định không hiệu quả. Việc áp dụng thuật toán Otsu đã cải thiện rất lớn trong việc nâng cao độ chính xác của hệ thống.

Kết thúc của quá trình phân đoạn ảnh, hệ thống nhận về kết quả là một ảnh trắng đen. Việc bị nhiễu thông tin sau quá trình phân đoạn ảnh là không thể tránh khỏi. Để cải thiện vấn đề trên, chúng tôi áp dụng các thuật toán hình thái học do Luc Vincent (2018) đề xuất để loại bỏ các thông tin không cần thiết và bổ sung thêm các thông tin cần thiết cho quá trình nhận dạng. Thuật toán bao gồm tập hợp các phép toán phi tuyến tính tác động đến hình dạng hoặc hình thái của các điểm nhị phân trong ảnh dựa trên các phép toán AND, OR, XOR và NOT để biến đổi các điểm nhị phân.

Tại Bước 3, tác giả sử dụng kỹ thuật OCR theo Ray Smith (2007) để xây dựng chức năng nhận diện ký tự. Khi hệ thống đã trải qua hết các bước xử lý cơ bản hình ảnh, chức năng nhận diện ký tự sẽ được hệ thống gọi đến để tiến hành phát hiện và nhận diện ký tự trong ảnh. Trong giai đoạn này chúng tôi sử dụng thư viện Tesseract OCR theo Thomas Hegghammer (2022), có tham khảo thêm lý thuyết từ Chirag Indravadanbhai Patel và cộng sự (2012) với mục đích phát hiện vùng ảnh chứa các đối tượng thông tin trên văn bằng, chứng chỉ. Từ các vùng ảnh đã được xác định, thư viện sẽ nhận dạng các ký tự văn bản trên hình ảnh. Kết thúc quá trình trên, hệ thống nhận về được một tập các ký tự, từ đã được thể hiện trên hình ảnh. Quá trình này sẽ là cơ sở cho bước phân loại thông tin mà hệ thống cần sử dụng.

Trong Bước 4, đầu tiên tác giả xây dựng một tập tin ngữ nghĩa với định dạng XML như giới thiệu của Andrea Zisman (2000), tập tin này sẽ chịu trách nhiệm hỗ trợ hệ thống xác định được các giá trị văn bản nào là các trường mà hệ thống cần sử dụng như Arenas và cộng sự (2004). Tập tin XML do người dùng tự định nghĩa theo cấu trúc tự quy định phù hợp với thông tin trên văn bằng, chứng chỉ. Trong quá trình trích xuất thông tin, để cải thiện hiệu hiệu quả của quá trình, hệ thống sẽ liên tục cập nhật các giá trị và trường hợp bất thường trong quá trình vận hành của hệ thống. Ví dụ, đối với thông

tin "Cấp cho:" trên chứng chỉ, khi trích xuất bởi OCR có thể dẫn đến nhiều trường hợp không chính xác và tác giả sẽ xây dựng các trường hợp để điều chỉnh lại thông tin cho nội dung "Cấp cho:" bằng cách thay thế các từ được trích xuất không chính xác thành "Cấp cho:".

Cuối cùng, tại Bước 5, thực hiện đối chiếu kết quả từ quá trình trích xuất thông tin với dữ liệu đã được lưu trữ trong cơ sở dữ liệu. Trường hợp có sự bất đồng bộ tại trường thông tin nào thì sẽ hiển thị cảnh báo cho trường thông tin đó.

3.2. Thực nghiệm

Để đánh giá mô hình trên, tác giả đã xây dựng hệ thống đối chiếu thông tin và được tích hợp vào trang tra cứu chứng chỉ của Trung tâm Ngoại ngữ - Tin học để đánh giá hiệu quả của mô hình.

Trong những năm gần đây, việc công khai thông tin chứng chỉ ứng dụng công nghệ thông tin (viết tắt là chứng chỉ UDCNTT) phải đảm bảo chính xác, do đó việc kiểm tra thông tin trước khi công khai là khâu rất quan trọng. Thông tin công khai bao gồm học tên, ngày sinh, nơi sinh, điểm thi trắc nghiệm, điểm thi thực hành, số hiệu, số vào sổ của các thí sinh đạt kỳ thi sát hạch. Tuy nhiên, việc thực hiện xác nhận dữ liệu chứng chỉ của thí sinh trước khi công khai thường được kiểm tra thủ công bằng phương pháp quan sát dữ liệu đã lưu trữ so với chứng chỉ UDCNTT thực tế. Quá trình kiểm tra này thường mất từ 3 đến 4 phút cho mỗi chứng chỉ UDCNTT.

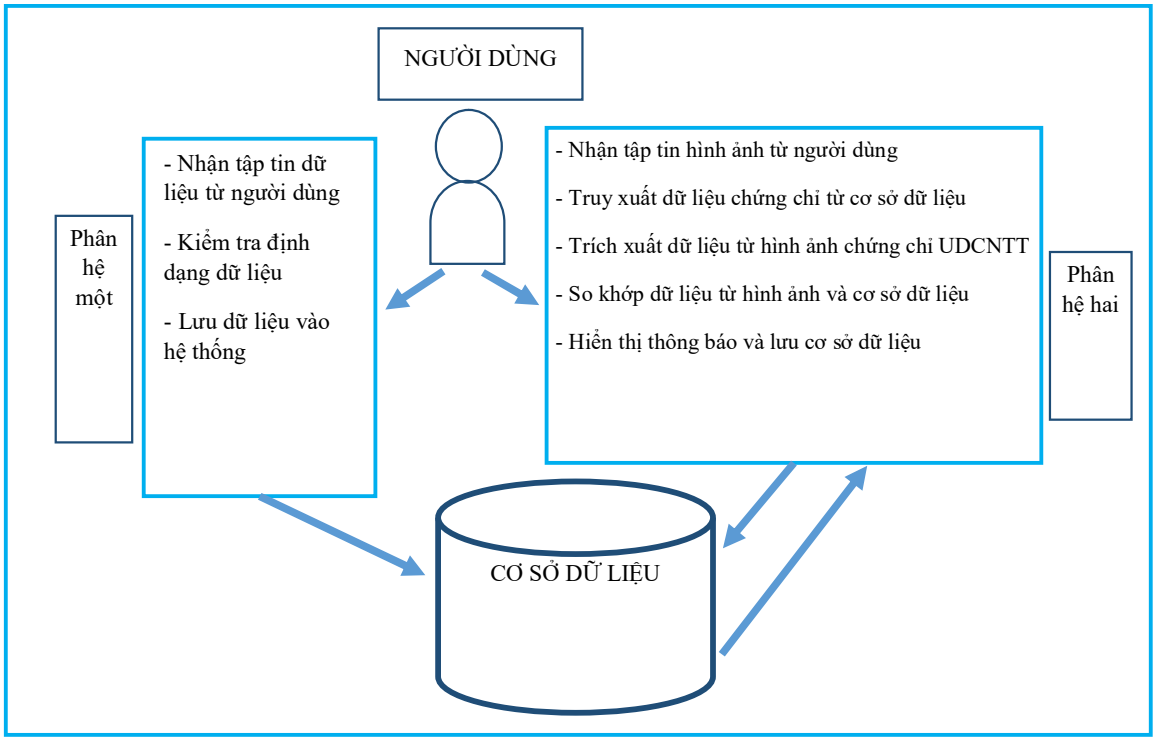
Hình 3. Trang tra cứu thông tin chứng chỉ

Nguồn: Trung tâm Ngoại ngữ - Tin học, (2023).

Bên cạnh đó, với số lượng thí sinh tham gia dự thi ngày càng lớn và việc in cấp, công bố thông tin chứng chỉ UDCNTT phải được thực hiện không quá 30 ngày kể từ ngày thi, việc kiểm tra thông tin thủ công như trước đây mất nhiều thời gian do đó có thể sẽ không được hiệu quả.

Vì thế, tác giả trình bày kiến trúc tổng quát của hệ thống lưu trữ và đối chiếu thông tin chứng chỉ UDCNTT. Trong hệ thống của tác giả thông qua hai phân hệ chính. Các phân hệ hoạt động theo trình tự đã thiết lập và phân hệ lưu trữ dữ liệu (phân hệ một) sẽ hoạt động trước, phân hệ

đối chiếu (phân hệ hai) sẽ hoạt động sau. Kết thúc thực hiện quá trình của hệ thống, dữ liệu sẽ được hoàn thiện và xác thực để đáp ứng nhu cầu khai thác sử dụng của người dùng. Kiến trúc tổng thể của hệ thống được mô tả ở Hình 4.



Hình 4. Kiến trúc tổng thể của hệ thống

Nguồn: Công bố của tác giả, (2023).

Phân hệ một: hệ thống tạo và lưu trữ thông tin chứng chỉ UDCNTT. Ở phân hệ này hệ thống nhận một tập tin Excel chứa thông tin chứng chỉ UDCNTT cấp cho học viên theo từng khóa thi. Thực hiện các kiểm tra tính hợp lệ của dữ liệu và thêm dữ liệu chứng chỉ UDCNTT vào cơ sở dữ liệu của hệ thống.

Phân hệ hai: thực hiện bổ sung trường thông tin hình ảnh chứng chỉ UDCNTT cho dữ liệu được lưu trữ tại phân hệ một. Phân hệ này cho phép người dùng tải lên một tập tin có phần mở rộng là pdf chứa hình ảnh chứng chỉ UDCNTT. Trong quá trình này, hệ thống sử dụng mô hình do chúng tôi đề xuất để đối chiếu thông tin trong hình ảnh chứng chỉ UDCNTT với các trường thông tin trong cơ sở dữ liệu để nâng cao hiệu quả của việc xác thực thông tin lưu trữ.

Tác giả xây dựng một tập tin ngữ nghĩa với định dạng XML dành cho chứng chỉ UDCNTT với cấu trúc tập tin được định nghĩa như Hình 5.

```
<?xml version="1.0" encoding="utf-8">
<!DOCTYPE chungchi [
        <!ELEMENT chungchi (capcho+, sinhngay+, noisinh+, diemtracnghiem+,
                        diemthuchanh+, sohieu+, sovaosocapchungchi+)>
        <!ELEMENT capcho (batdau, ketthuc)>
        <!ELEMENT sinhngay (batdau, ketthuc)>
        <!ELEMENT noisinh (batdau, ketthuc)>
        <!ELEMENT diemtracnghiem (batdau, ketthuc)>
        <!ELEMENT diemthuchanh (batdau, ketthuc)>
        <!ELEMENT sohieu (batdau, ketthuc)>
        <!ELEMENT sovaosocapchungchi (batdau, ketthuc)>
        <!ELEMENT batdau (#PCDATA)>
        <!ELEMENT ketthuc (#PCDATA)>
]>
```

Hình 5. Cấu trúc DTD sử dụng để định nghĩa các trường dữ liệu trên ảnh

Nguồn:Công bố của tác giả, (2023).

```
<?xml version="1.0" encoding="utf-8">
<chungchi>
        <capcho>
                <batdau> "Cấp cho:"  </batdau>
                <ketthuc> "End"  </ketthuc >
        </capcho>
        <capcho>
                <batdau> "Cấp cho:"  </batdau>
                <ketthuc> "End"  </ketthuc >
        </capcho>
        <capcho>
                <batdau> "Cáp cho:"  </batdau>
                <ketthuc> "End"  </ketthuc >
        </capcho>
        <capcho>
                <batdau> "Cap cho:"  </batdau>
                <ketthuc> "End"  </ketthuc >
        </capcho>
        <capcho>
                <batdau> "Càp cho:"  </batdau>
                <ketthuc> "End"  </ketthuc >
        </capcho>
</chungchi>
```

Hình 6. Giá trị có thể xảy ra của trường thông tin "Cấp cho:"

Nguồn: Công bố của tác giả, (2023).

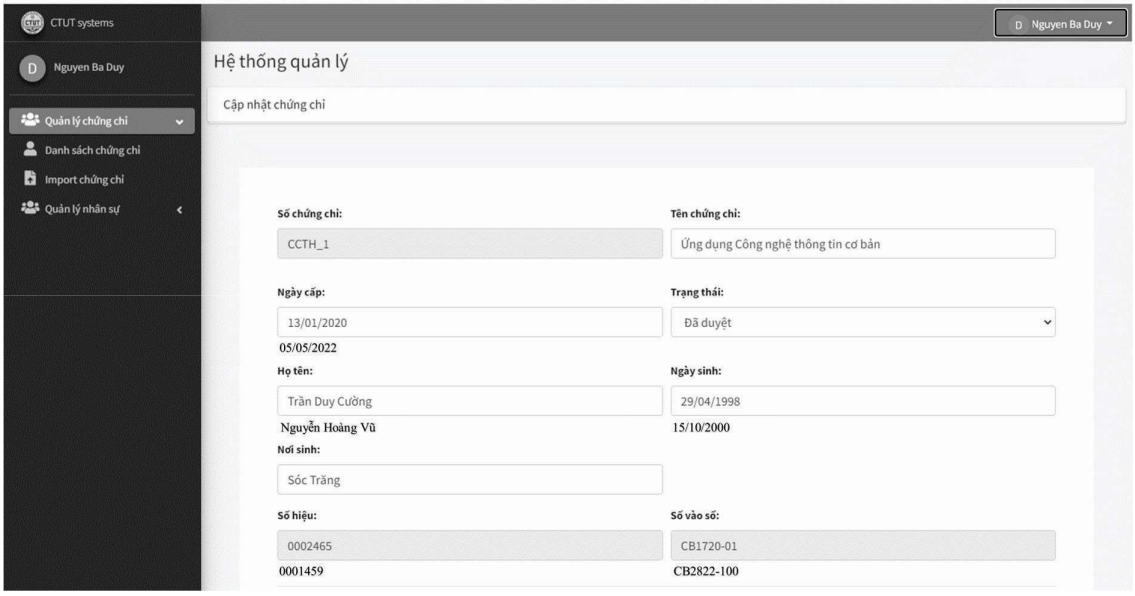
Người dùng chọn tập tin chứng chỉ UDCNTT và hệ thống sẽ tự động thực hiện Bước 2, Bước 3 trong mô hình đề xuất. Tại Bước 4, như thông tin trên ví dụ của mô hình, thông tin khi trích xuất có thể bị sai và tác giả áp dụng tập tin XML đã định nghĩa để cập nhật lại thông tin đúng như Hình 6.

Bên cạnh đó, việc nhị phân hoá ảnh không tối ưu sẽ dẫn đến độ chính xác của việc trích xuất thông tin từ chứng chỉ UDCNTT. Một minh họa cho các bước xử lý bởi hệ thống cho trường hợp dữ liệu trích xuất liên quan đến ngày cấp chứng chỉ UDCNTT bị sai do chồng lấn từ dấu mộc của Trường lên trường thông tin.

Hình 7. Quá trình nhị phân hóa ảnh không tốt trong cụm dữ liệu nhiều thông tin

Nguồn: Chứng chỉ thuộc Trung tâm Ngoại ngữ - Tin học, (2023).

Theo kết quả đo được, thời gian hệ thống xử lý và trả về kết quả so khớp thông tin từ khi người dùng chọn tập tin chứng chỉ UDCNTT không vượt quá 3 giây. Kết quả so khớp trên hệ thống sau khi tích hợp như Hình 8.



Hình 8. Hệ thống cảnh báo các trường thông tin được dự đoán bất đồng bộ

Nguồn: Công bố của tác giả, (2023).

Người dùng có thể kiểm tra lại các trường thông tin có chính xác hay chưa, chỉnh sửa nếu phát hiện sai sót. Sau giai đoạn này hệ thống sẽ hoàn thiện việc bổ sung trường thông tin cho cơ sở dữ liệu. Từ việc dữ liệu đã hoàn thiện, hệ thống sẽ sử dụng các thông tin đó để phục vụ cho các phân hệ tra cứu, thống kê, quản lý.

Để đánh giá mức độ hoạt động của hệ thống, chúng tôi chạy thử nghiệm với bộ dữ liệu gồm 200 chứng chỉ UDCNTT. Kết quả độ chính xác trong quá trình gợi ý các xung đột thông tin được mô tả tại Bảng 1.

Bảng 1. Thống kê kết quả thực nghiệm

Trường dữ liệu	Số lượng	Thành công	Thất bại	Hiệu quả
Ngày cấp	200	174	26	87.00%
Họ tên	200	179	21	89.50%
Ngày sinh	200	170	30	85.00%
Số hiệu	200	190	10	95.00%
Số vào sổ	200	184	16	92.00%
Tổng	1000	897	103	89.70%

Nguồn: Công bố của tác giả, (2023).

4. KẾT LUẬN

Qua việc so khớp dữ liệu chứng chỉ UDCNTT, thông tin được trích xuất từ mô hình tác giả đề xuất được so sánh với dữ liệu đã nhập để xác nhận tính chính xác của thông tin đã có trên cơ sở dữ liệu trước khi công bố cho người dùng tra cứu. Điều này giúp tăng cường khả năng tự động hóa, tăng cường hiệu suất làm việc và cải thiện quy trình duyệt thông tin chứng chỉ UDCNTT. Tuy nhiên, hệ thống còn gặp phải không ít khó khăn do việc gợi ý các trường thông tin sai sót trong các trường hợp đặc trưng về cấu trúc thông tin chứng chỉ UDCNTT, mặt ngữ nghĩa tiếng Việt. Để có thể cải thiện hiệu quả và ứng dụng mô hình rộng hơn trong thực tế, tác giả đề xuất thêm các hướng phát triển trong tương lai như: xây dựng các bộ ngữ nghĩa cho các trường dữ liệu; xác định, quản lý và cập nhật các thông tin sai để phục vụ cho quá trình đối chiếu dữ liệu, kiểm tra và hỗ trợ sửa lỗi chính tả, xóa nền (ảnh mộc đóng dấu, vết mực lem).

Tài liệu tham khảo

Anand Shinde, Parvinder Singh, Jay Patil, Jaideep Singh, Trupti Baraskar (2021). "Text Extraction from Images using Tesseract". International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 07.

Andrea Zisman (2000). "An overview of XML". Computing & Control Engineering Journal, 11.4: 165-167.

Anshul Arora, Rajat Singh, Ashiq Eqbal, Ankit Mangal, Prof. S.U Saoji (2021). "Extraction and detection of text from images". International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 08.

Arenas, Marcelo; Libkin, Leonid (2004). "A normal form for XML documents". ACM Transactions on Database Systems (TODS), 29.1: 195-232.

Chandra R. A. Perdana, Hanung Adi Nugroho, Igi Ardiyanto (2020). "Comparison of text-image fusion models for high school diploma certificate classification". Communications in Science and Technology 5(1) 5–9.

Chirag Indravadanbhai Patel, Atul Patel, Dharmendra Patel (2012). "Optical character recognition by open source OCR tool tesseract: A case study". International Journal of Computer Applications, 55.10: 50-56.

G. Li and N. Li (2019). "Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network, Electron". Commer. Res. 19(4) 799–800.

John Anthony C. Jose, Allysa Kate M. Brillantes, Elmer P. Dadios, Edwin Sybingco, Laurence A. Gan Lim, Alexis M. Fillone, and Robert Kerwin C. Billones (2021). "Recognition of Hybrid Graphic-Text License Plates". Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.25 No.4.

Luc Vincent (2018). "Morphological algorithms". In: Mathematical Morphology in Image Processing. CRC Press. p. 255-288.

Meredita Susanty, Herminarto Nugroho (2020). "Optical Character Recognition implementation for Admission system in universitas Pertamina". Journal SIMETRIS, Vol. 11 No. 1, P-ISSN: 2252-4983, E-ISSN: 2549-3108.

Nhã, V.T., Phụng, T.S.M., Tú, N.H., Dung, Đ.T.K. và Cường, L.Đ.P., (2022). "Xây dựng hệ thống trích xuất thông tin giấy tờ cá nhân từ hình ảnh cho hệ thống EKYC". Journal of Science & Technology , Vol. 58 - No. 2.

Nobuyuki Otsu (1979). "A Threshold Selection Method from Gray-Level Histogram". IEEE Transactions on Systems, Man, and Cybernetics, vol.9, no.1, pp. 62-66.

P. Bergmann, T. Meinhardt, and L. Leal-Taix´ e (2019). "Tracking Without Bells and Whistles". 2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 941-951, doi: 10.1109/ICCV.2019.00103.

Ray Smith (2007). "An overview of the Tesseract OCR engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). IEEE". p. 629-633.

Thomas Hegghammer (2022). "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment". Journal of Computational Social Science, 5.1: 861-882.

Xiangyang Xu, Shengzhou Xu, Lianghai Jin, Enmin Song (2011). "Characteristic analysis of Otsu threshold and its applications". Pattern recognition letters, 32.7: 956-961.

BUILDING A DATA VERIFICATION MODEL BASE ON OPTICAL CHARACTER RECOGNITION

ABSTRACT

Certificates validation by comparing the information on certificate scanned image with the data stored in the database is a simple and efficient method. In this research, we propose to build a model that allows automatic extraction of text information from certificates using optical character recognition techniques to compare data before publishing the information. We investigated the model on a data set of 200 applied information technology certificates from the Center for Foreign Languages - Informatics at Can Tho University of Technology, building a data comparison system integrated into the Center's certificate lookup system and experimental results show that it can digitize and extract information with 89.72% accuracy. Based on that, we propose the most significant system and give some recommendations for future researchs.

Keywords: *Computer Vision, image information extraction, image processing.*