

# PHÁT TRIỂN MỘT ỨNG DỤNG DI ĐỘNG HỖ TRỢ HỌC TIẾNG ANH BẰNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG

DEVELOP A MOBILE APPLICATION TO SUPPORT ENGLISH LEARNING BY OPTICAL CHARACTER RECOGNITION TECHNOLOGY

Trần Tiến Dũng<sup>1,\*</sup>,  
 Khuất Trọng Hiếu<sup>2</sup>, Trần Đông Kiên<sup>3</sup>

## TÓM TẮT

Sinh viên trong giờ học tiếng Anh thường tốn nhiều thời gian tra cứu các từ mới trong tài liệu hoặc trên bảng. Chúng tôi đã phát triển một ứng dụng di động tên là HaUI-OCR cho phép dịch các tài liệu tiếng Anh bằng chức năng chụp ảnh của điện thoại di động hệ điều hành Android. Công nghệ phát triển phần mềm và công nghệ nhận dạng ký tự quang OCR đã được chúng tôi tích hợp với nhau để tạo thành một hệ thống phần mềm hoàn chỉnh. Phần mềm hỗ trợ học tiếng Anh tại lớp là một giải pháp góp phần phát triển Giáo dục 4.0 tại Đại học Công nghiệp Hà Nội.

**Từ khoá:** Phần mềm, nhận dạng ký tự quang, học tiếng Anh.

## ABSTRACT

It usually takes significant time for a student at the English class to translate the new words in printed documents and board. We developed a mobile application called HaUI-OCR that allows its users to translate English documents by the camera of an Android mobile. Software engineering and Optical Character Recognition (OCR) technologies were integrated to form a complete software system. The software for English learners at the class is a solution that contributes to the development of Education 4.0 at Hanoi University of Industry.

**Keywords:** Software, optical character recognition, English learning.

<sup>1</sup>Bộ môn Công nghệ Phần mềm, Khoa CNTT, Đại học Công nghiệp Hà Nội

<sup>2</sup>Lớp KTPM-K9, Khoa CNTT, Đại học Công nghiệp Hà Nội

<sup>3</sup>Lớp KTPM-K8, Khoa CNTT, Đại học Công nghiệp Hà Nội

\*Email: trantd.vn@gmail.com

Ngày nhận bài: 08/01/2018

Ngày nhận bài sửa sau phản biện: 05/4/2018

Ngày chấp nhận đăng: 21/8/2018

Phản biện khoa học: TS. Nguyễn Bá Nghiễn

## 1. GIỚI THIỆU

OCR là một thuật ngữ viết tắt từ cụm từ Optical Character Recognition, một công nghệ nhận dạng ký tự quang được áp dụng để phát hiện các văn bản trên một file ảnh. Các sản phẩm OCR dựa trên các nghiên cứu trong lĩnh vực nhận dạng mẫu, trí tuệ nhận tạo và thị giác máy tính để xử lý các định dạng dữ liệu như hình 1. Hiện nay, công nghệ OCR đang được nghiên cứu ứng dụng rộng rãi trong lĩnh

vực nhận dạng ảnh, một nền tảng của công nghiệp 4.0 [1, 2]. Hai hãng Google và ABBYY đang đầu tư vào nghiên cứu phát triển các hệ thống phần mềm OCR như Tesseract và ABBYY FineReader. Các thư viện phần mềm OCR của các hãng này đã mang lại một giải pháp kỹ thuật mới cho các nhà phát triển phần mềm [3]. Tuy nhiên, các phần mềm OCR hầu hết chỉ hoạt động hiệu quả với các file hình ảnh văn bản có chất lượng tốt và trên một vài ngôn ngữ Latin nhất định. Với các ngôn ngữ có "đấu" (sắc, huyền, hỏi, ngã, nặng) như tiếng Việt, các phần mềm OCR vẫn còn thiếu nhiều ưu điểm [4, 5]. Đặc biệt, lĩnh vực phần mềm OCR tiếng Việt dành cho đào tạo đang phát triển khá khiêm tốn ở thị trường Việt Nam. Ví dụ, các phần mềm như VnDOCR, VietOCR có thể làm việc tốt với tiếng Việt, song chưa có phiên bản phục vụ việc học tập cho sinh viên, nhất là hỗ trợ học tiếng Anh.



Hình 1. Mô hình phần mềm OCR tổng quát. Cho phép nhận dạng văn bản từ .pnd, .jpeg, .pdf, .gif và các file dữ liệu ảnh khác. Nguồn www.infourok.ru

Hầu hết sinh viên sử dụng điện thoại thông minh hằng ngày để chụp ảnh, ghi hình bài giảng, tra cứu tài liệu từ Internet và thực hiện các hoạt động học tập trực tuyến theo xu thế cách mạng 4.0. Tuy nhiên, việc tra cứu và biên dịch các tài liệu tiếng Anh bằng phần mềm di động có hiệu quả chưa cao vì các từ mới tại lớp chủ yếu ở trên sách in, hoặc các file ảnh [6, 7]. Người dùng đang tra từ bằng cách nhập lại từng từ vào các phần mềm điện thoại di động như Google Translate, The Free Dictionary. Điều này đã góp phần làm giảm hiệu quả cho giờ học tiếng Anh tại lớp.

Chúng tôi phát triển một phần mềm di động hệ điều hành Android, tên là HaUI-OCR, nhận dạng văn bản trong các hình ảnh. Phần mềm đã được xây dựng dựa trên thư viện phát triển phần mềm nhận dạng ký tự quang Tesseract-OCR. Nó sử dụng các API của Google Translate để biên dịch tài liệu theo một kịch bản ở hình 2. Phần mềm cho người học tiếng Anh tại lớp là một giải pháp góp phần phát triển Giáo dục 4.0 tại Đại học Công nghiệp Hà Nội.

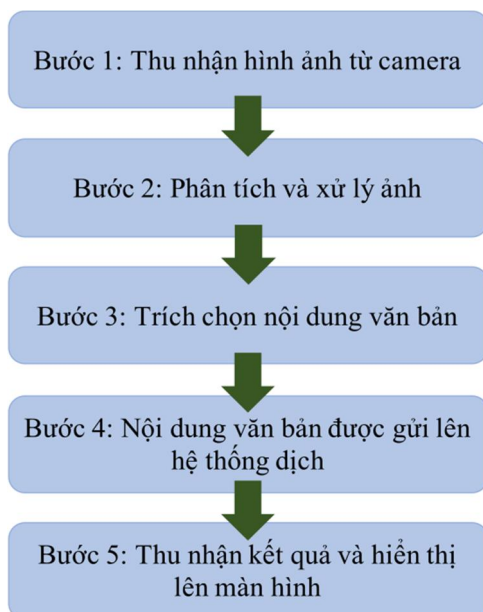


Hình 2. Kịch bản sử dụng phần mềm HaUI-OCR

## 2. PHƯƠNG PHÁP VÀ TƯ LIỆU

### 2.1. Nguyên lý hoạt động của HaUI-OCR

Phần mềm HaUI-OCR sử dụng camera trên điện thoại để thu nhận hình ảnh, truyền dữ liệu về mô đun nhận dạng ký tự OCR theo sơ đồ thuật toán ở hình 3. Kỹ thuật nhận dạng ký tự OCR đã được trình bày chi tiết trong [8, 9]. Chúng tôi đã sử dụng thư viện nguồn mở Croplmage trên GitHub để cài đặt mô đun nhận dạng và trích rút dữ liệu văn bản từ 01 hình ảnh. Mô đun Tess-Twoo của thư viện này đã được import vào mã nguồn để có thể nhúng được thư viện này vào code.

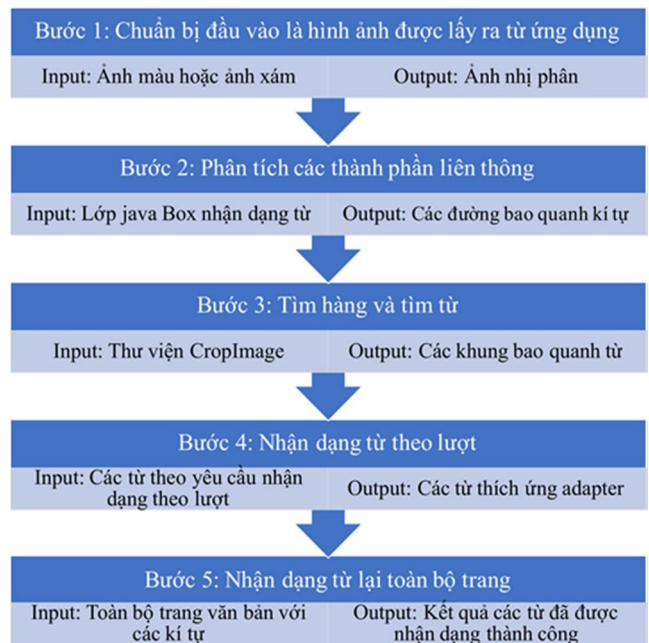


Hình 3. Sơ đồ thuật toán biên dịch hình ảnh văn bản

Tess-Two có chứa 2 công cụ biên dịch chính là Tesseract-OCR và Leptonica. Hai công cụ này có thể sử dụng trên nền tảng mã nguồn mở Android. Nó sẽ cung cấp 01 API để truy cập vào các API Tesseract-OCR và Leptonica, tương ứng là các máy nhận dạng ký tự quang đa nền tảng và thư viện xử lý ảnh. Trong đó các chức năng của Tesseract-OCR làm nhiệm vụ nhận dạng được chữ viết trên hình ảnh trong phần mềm. Nó có thể nhận dạng được nhiều ngôn ngữ khác nhau trên thế giới đặc biệt là ngôn ngữ theo hệ Latin.

Chúng tôi đã cài đặt 1 lớp đối tượng TessBaseAPI đóng gói các chức năng liên quan đến Tesseract-OCR. Các API của Tesseract sẽ được truy cập thông qua các đối tượng thuộc TessBaseAPI. Để khởi tạo các đối tượng này, phần mềm cần được người dùng cung cấp đường dẫn tới thư mục chứa tập tin dữ liệu ngôn ngữ và loại ngôn ngữ được sử dụng. Ở đây, chúng tôi dùng thư mục bên trong thư mục Data của ứng dụng và ngôn ngữ là tiếng Việt.

### 2.2. Ứng dụng thuật toán nhận dạng ký tự



Hình 4. Sơ đồ ứng dụng thuật toán nhận dạng ký tự từ thư viện Tesseract-OCR

Mô đun nhận dạng ký tự được vận hành theo một thuật toán có sẵn trong thư viện Tesseract-OCR, việc ứng dụng thuật toán được thể hiện trên hình 4. Các bước sử dụng các thuật toán đã được cài đặt sẵn trong thư viện Tesseract như sau:

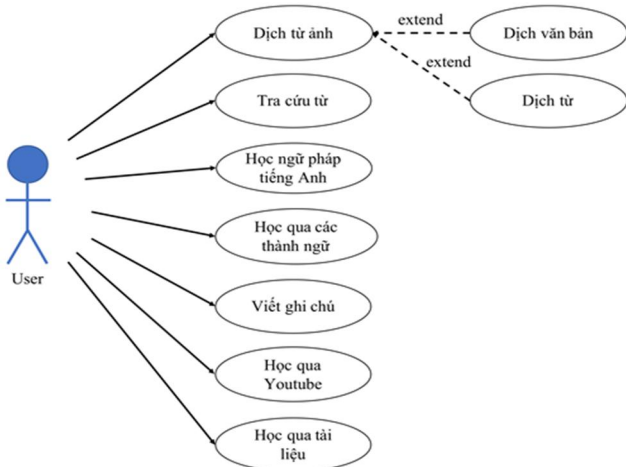
- *Bước 1:* Lấy một hình từ thư viện CroplmageView - là thư viện mà tạo ra các khung hình được cắt ra từ ảnh gốc. Hình này là dữ liệu đầu vào và được chuyển về định dạng bitmap với kiểu ảnh đen trắng.
- *Bước 2:* Phân tích phần liên thông tạo ra các khung có đường bao quanh với kích thước xác định.
- *Bước 3:* Phân tích các khung bao quanh để phân thành các hàng và tìm ra các từ.

• **Bước 4:** Nhận dạng các từ theo lượt. Các từ thỏa mãn yêu cầu trong giai đoạn này sẽ được chuyển sang bộ phân loại thích ứng để làm dữ liệu huấn luyện.

• **Bước 5:** Nhận dạng lại toàn bộ trang. Giai đoạn này sẽ quét hết toàn bộ trang, các từ không được nhận diện chính xác ở giai đoạn đầu sẽ được nhận diện lại lần nữa.

**2.3. Các chức năng chính của hệ thống**

Dựa trên việc quan sát các hoạt động học tiếng Anh thực tế tại lớp của sinh viên chuyên ngành Kỹ thuật phần mềm, Khoa Công nghệ thông tin, chúng tôi đã thu thập yêu cầu và thiết kế một hệ thống phần mềm đáp ứng các ca sử dụng của người dùng như hình 5.



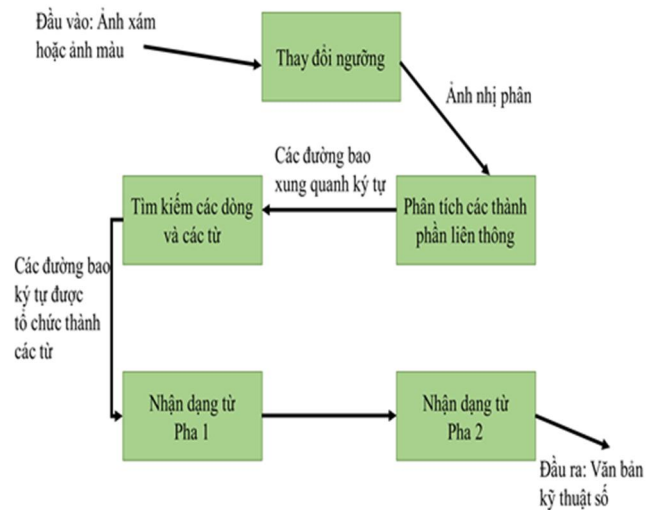
Hình 5. Biểu đồ ca sử dụng hệ thống phần mềm

**2.4. Các công cụ phát triển phần mềm**

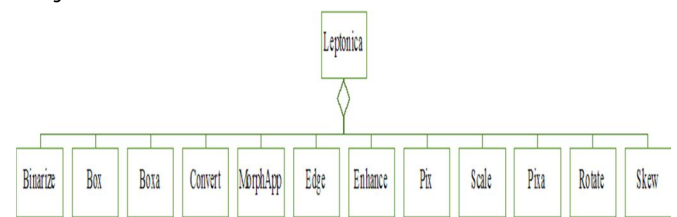
**a) Android Studio.** Đây là môi trường phát triển tích hợp (IDE), chính thức dành cho phát triển nền tảng Android. Nó hoạt động trên các hệ điều hành Windows, Mac OS X và Linux. Đây là công cụ phát triển phần mềm chính thức của Google cho các ứng dụng chạy trên hệ điều hành Android. Chúng tôi đã sử dụng Android Studio thay thế cho Android Development Tools (ADT) dựa trên Eclipse [10].

**b) Tesseract và Leptonica.** Tesseract là 1 thư viện OCR nguồn mở có thể tích hợp với các ứng dụng Android. Nó tạo ra các module có chức năng OCR nhận dạng trên ảnh và sử dụng các hàm API do Google cung cấp để tạo ra mô đun có chức năng dịch văn bản sang ngôn ngữ tùy chọn [11]. Nguyên lý hoạt động của thư viện Tesseract được mô tả như hình 6. Leptonica là một thư viện nguồn mở hữu ích cho các ứng dụng xử lý ảnh [5, 12]. Cấu trúc của thư viện Leptonica được mô tả trong hình 7. Các nhóm chức năng chính của Leptonica bao gồm:

- Phép biến đổi Affine (rộng, dịch, xoay, cắt) trên hình ảnh độ sâu pixel tùy ý.
- Chuyển đổi dự báo và Bilinear.
- Hình thái nhị phân và màu xám, các bộ lọc thứ tự xếp hạng, và convolution.
- Seedfill và các thành phần kết nối.



Hình 6. Nguyên lý hoạt động Tesseract-OCR. Đường nét liên biểu diễn luồng thông tin

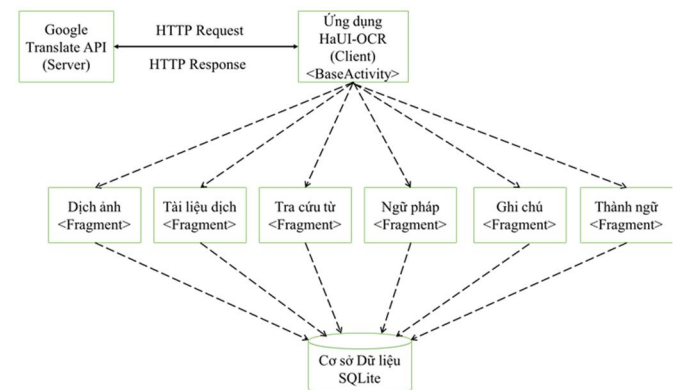


Hình 7. Biểu đồ thành phần Leptonica. Mũi tên thể hiện lớp Leptonica được cấu tạo từ các lớp thành phần khác

**c) Dữ liệu từ điển.** Phần mềm đã dùng dữ liệu từ điển của Google Translate, một dịch vụ phần mềm trực tuyến có các APIs cho phép các nhà phát triển phần mềm lập trình chức năng dịch tự động một văn bản sang một ngôn ngữ khác. Một tài liệu có kích thước lớn có thể tải lên máy chủ Google để biên dịch [13]. Các bản dịch có thể hiệu đính được và sự hiệu đính này sẽ được sử dụng trong các lần dịch sau [12, 14]. Khả năng học từ sự hiệu đính của người dùng đã giúp cho chất lượng biên dịch của Google Translate ngày một cải thiện [15, 16].

**3. KẾT QUẢ**

**3.1. Sản phẩm phần mềm HaUI-OCR**

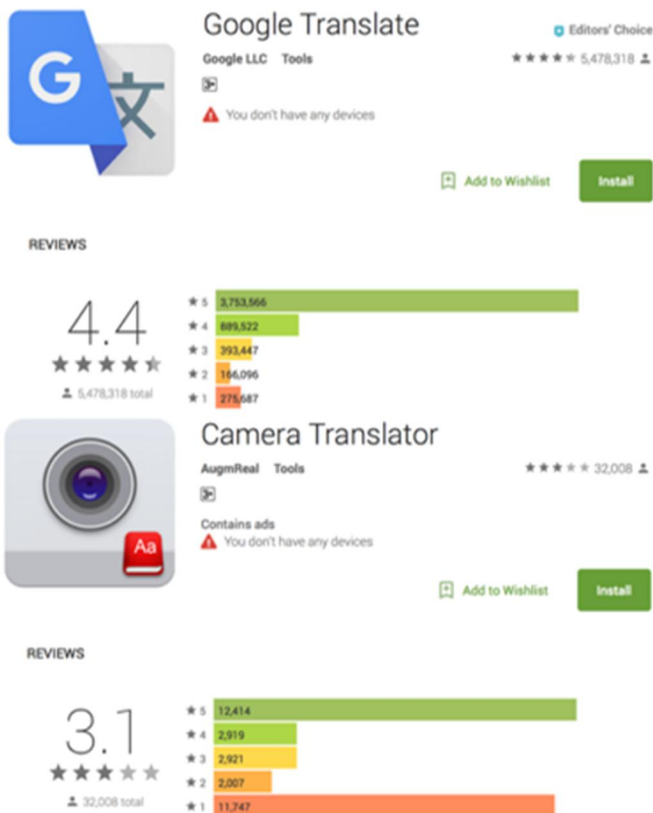


Hình 8. Biểu đồ kiến trúc hệ thống HaUI-OCR. Đường nét liền và đường nét đứt tương ứng với luồng thông tin và quan hệ phụ thuộc

Mục đích của chúng tôi là tạo ra một ứng dụng điện thoại di động phục vụ việc học tiếng Anh chuyên ngành Công nghệ thông tin. Điểm nổi bật của ứng dụng là hỗ trợ chức năng Việt hóa giáo trình tiếng Anh của Đại học Công nghiệp Hà Nội, chuyển tài liệu tiếng Anh dạng hình ảnh sang tài liệu văn bản tiếng Việt. Kiến trúc phần mềm được biểu diễn trong hình 8.

### 3.2. So sánh tính năng phần mềm

HaUI-OCR cho phép tra cứu bản dịch của tài liệu giáo trình tiếng Anh. Phần mềm cho phép dịch từ trong ảnh, ghi chú là lưu lại lịch sử biên dịch. Nó cũng đóng vai trò giống như một quyển từ điển tiếng Anh, cho phép người dùng tra cứu từng từ riêng lẻ. Ngoài các tính năng biên dịch, nó cho phép học tiếng Anh online trên Youtube. Để có cơ sở phát triển các phiên bản phần mềm tốt hơn, chúng tôi đã so sánh phiên bản phần mềm hiện tại với các ứng dụng khác cùng lĩnh vực. Chúng tôi đã chọn 02 ứng dụng tiêu biểu là Google Translate và Camera Translator làm chuẩn để so sánh vì chúng có lượng người dùng lớn, số điểm đánh giá cao, lại được phát triển bởi các hãng phần mềm danh tiếng. Thông tin về hai phần mềm này trong Google Play được biểu diễn trong hình 9.



Hình 9. Thông tin về 02 ứng dụng chuẩn Camera Translator, Google Translate. Việc lựa chọn ứng dụng chuẩn dựa trên lượng người dùng, số điểm cho ứng dụng. Số liệu ngày 10/03/2018 trên Google Play

Do được tích hợp nhiều tính năng dành riêng cho nhóm người dùng là sinh viên ngành kỹ thuật, phần mềm có các chức năng hỗ trợ việc học tập nhiều hơn Google Translate và Camera Translator (bảng 1). Tuy nhiên, do nguồn lực hạn

chế, HaUI-OCR mới chỉ hoạt động tốt trên nền tảng di động phổ biến Android (bảng 2), chưa có các phiên bản dành cho iPhone, Window phone.

Bảng 1. So sánh chức năng của HaUI-OCR với hai ứng dụng chuẩn Google Translate và Camera Translator. HaUI-OCR có hỗ trợ chức năng học tập nhiều hơn hai ứng dụng chuẩn

Chức năng \ Ứng dụng	Google Translate	Camera Translator	HaUI-OCR
Ghi chú	X		X
Dịch từ	X	X	X
Dịch ảnh	X	X	X
Dịch bằng giọng nói	X	X	
Dịch hình vẽ	X		
Từ điển	X		X
Dịch giáo trình			X
Học online			X

Bảng 2. So sánh khả năng hoạt động đa nền tảng di động của HaUI-OCR, Google Translate và Camera Translator

Đa nền tảng \ Ứng dụng	Google Translate	Camera Translator	HaUI-OCR
Android	X	X	X
iOS	X	X	
Windows Phone	X		

### 4. KẾT LUẬN

Xuất phát từ một vấn đề thực tế trong giờ học tiếng Anh chuyên ngành của sinh viên Công nghệ thông tin, chúng tôi đã áp dụng công nghệ nhận dạng ký tự quang Tesseract-OCR trên nguồn dữ liệu từ điển của Google Translate để tạo ra phần mềm HaUI-OCR. Ứng dụng di động này đã sử dụng camera trên điện thoại để thu dữ liệu hình ảnh văn bản đầu vào, chuyển cho thư viện Tesseract-OCR nhận dạng sang văn bản, rồi gửi văn bản đó cho Google Translate biên dịch. Như vậy, nó đã phối hợp các dịch vụ có sẵn của một hệ sinh thái phần mềm để tạo thành một sản phẩm phần mềm mới. Các công cụ phát triển phần mềm mới nhất của Google đã được áp dụng trong phát triển ứng dụng này nhằm giúp người dùng biên dịch các tài liệu tiếng Anh. Chúng tôi đã so sánh phiên bản đầu của phần mềm HaUI-OCR với các phần mềm khác tiêu biểu trên Google Play để có cơ sở phát triển các phiên bản sau tốt hơn. Do được tích hợp nhiều tính năng dành riêng cho nhóm người dùng sinh viên ngành kỹ thuật, phần mềm có nhiều chức năng hỗ trợ việc học tập hơn Google Translate và Camera Translator. Do nguồn lực hạn chế, HaUI-OCR mới chỉ hoạt động tốt trên nền tảng di động phổ biến Android. Sản phẩm của chúng tôi đã hỗ trợ việc học tiếng Anh chuyên ngành của sinh viên CNTT tại lớp, góp phần phát triển Giáo dục 4.0 tại Đại học Công nghiệp Hà Nội.

**TÀI LIỆU THAM KHẢO**

- [1]. S. Mori, C. Y. Suen, and K. Yamamoto, 1992. *Historical review of OCR research and development*. Proceedings of the IEEE, vol. 80, no. 7, pp. 1029-1058.
- [2]. Messina, C., 2006. *Open source OCR in the wild*.
- [3]. S. Impedovo, R. Modugno, A. Ferrante, and E. Stasolla, 2009. *New Trends in Digital Scanning Processes*. Presented at the 2009 10th International Conference on Document Analysis and Recognition, pp. 1071–1075.
- [4]. B. T. Hồ và M. C. Lương. *Về xử lý tiếng Việt trong công nghệ thông tin*.
- [5]. Vũ Văn Tâm, B.N.L., 2016. *Áp dụng giải pháp nhận dạng ký tự quang (OCR) trong biên mục tài liệu lưu trữ*.
- [6]. Thạc Bình Cường, H.X.N., 2004. *Tiếng Anh chuyên ngành Công nghệ thông tin*.
- [7]. Trung, T.C.V., 2015. *Giáo trình Tiếng Anh chuyên ngành Công nghệ thông tin*.
- [8]. R. Smith, 2007. *An Overview of the Tesseract OCR Engine*. Presented at the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, 2007, pp. 629–633.
- [9]. R. W. Smith, 2013. *History of the Tesseract OCR engine: what worked and what didn't*. Presented at the IS&T/SPIE Electronic Imaging, 2013, vol. 8658, p. 865802.
- [10]. Sơn, L.H., 2016. *Hướng dẫn cài đặt và cấu hình Android Studio*.
- [11]. R. Smith, D. Antonova, and D. S. Lee, 2009. *Adapting the Tesseract open source OCR engine for multilingual OCR*. Presented at the International Workshop, New York, New York, USA, p. 1.
- [12]. Vlasov, P., 2016. *How to detect text language offline*.
- [13]. Shin, Y.J., 2017. *Google Translate Image Translation*.
- [14]. M. Faruqui and S. Kumar, 2015. *Multilingual Open Relation Extraction Using Cross-lingual Projection*. arXiv.org, vol. cs.CL. 23-Mar-2015.
- [15]. Franz Och, 2006. *Statistical machine translation live*.
- [16]. M. Krack, J. Pugalia, J. Wong, and W. Wong. *System and method for automatic language translation for applications*. US8862456B2.