Áp dụng giải pháp nhận dạng ký tự quang học (OCR) trong biên mục tài liệu lưu trữ

OCR (Optical Character Recognition) – nhận dạng ký tự quang học là quá trình phần mềm máy tính xử lý và chuyển các file ảnh (thường là sản phẩm của máy scanner) thành các file văn bản dạng ký tự (text).

Nền tảng lý thuyết của OCR là các nghiên cứu về nhận dạng mẫu, trí tuệ nhận tạo và hiển thị bằng máy. Hiện nay, công nghệ OCR được ứng dụng rộng rãi, hiệu quả ở nhiều lĩnh vực, ở nhiều quốc gia.

1. Hiện trạng giải pháp kỹ thuật nhận dạng ký tự quang học

Hiện nay, giải pháp nhận dạng ký tự quang học đang phát triển mạnh về mặt ứng dụng và liên tục có nhiều cải tiến mới để tăng tính ứng dụng và hiệu quả của sản phẩm đầu ra. Đi đầu trong công nghệ OCR phải kể đến hai hãng phát triển và cải tiến phần mềm nhận dạng ký tự là Google và ABBYY.

Google trên nền tảng **Tesseract** (*Tesseract OCR engine*) được phát triển bởi HP Labs trong giai đoạn 1985-1995, sử dụng mã nguồn mở, có chất lượng nhận dạng chính xác cao, với nhiều định dạng file ảnh và có thể nhận dạng hơn 60 ngôn ngữ khác nhau.

Trong khi, **ABBYY** được coi là hãng tiên phong trong lĩnh vực OCR. ABBYY cho ra đời phần mềm nhận dạng ký tự quang học với tên gọi ABBYY có khả năng nhận dạng 190 ngôn ngữ. Đặc biệt, đối với ký tự La-tinh và tiếng Nga, công nghệ OCR của ABBYY có thể đạt hiệu quả nhận dạng đến 99% cho một file ảnh chất lượng tốt.

Tuy nhiên, việc nhận dạng ký tự tiếng Việt (loại hình ngôn ngữ có "dấu") vẫn là thách thức đối với sự phát triển của công nghệ OCR trên thế giới. Hiện tại, ABBYY đang tiến hành nghiên cứu và triển khai công nghệ nhận dạng tiếng Việt với độ chính xác trên 90% cho một file ảnh có chất lượng tốt. Song mới chỉ dừng lại ở việc nhận dạng ký tự tiếng Việt được soạn thảo bằng vi tính hoặc công nghiệp in ấn. Mà chưa thể vươn tới các loại hình chữ viết tay, thậm chí hiệu quả đạt được còn rất khiêm tốn đối với các loại tài liệu cũ, đa dạng về phông chữ hoặc sử dụng kỹ thuật in lạc hậu. Ngoài ra, giải pháp OCR của ABBYY còn thiếu tính cạnh tranh bởi giá thành cao, là sản phẩm đóng gói của nước ngoài, chưa thực sự phù hợp với các loại hình tài liệu của Việt Nam, cũng như khó khăn trong việc tích hợp vào các công nghệ khác.

Ở Việt Nam cũng có một vài hãng phần mềm đầu tư xây dựng công nghệ OCR, mà tiêu biểu có thể kể:

Phần mềm **VnDOCR 4.0 Professional**, chương trình nhận dạng chữ Việt in, được phát triển bởi nhóm chuyên gia phát triển phần mềm của Phòng Nhận dạng và Công nghệ tri thức, Viện Công nghệ thông tin – Viện Khoa học và Công nghệ Việt Nam. VnDOCR sử dụng chương trình điều khiển máy quét, để quét ảnh từ tài liệu in dưới dạng ảnh đen trắng (line Art, Black and White – B&W, độ phân giải 300dpi (dots per

inch), sau đó chuyển qua chế độ nhận dạng. Kết quả nhận dạng chữ Việt độ chính xác đạt khoảng trên 90% tùy vào chất lượng bản quét.

Phần mềm **VietOCR** được phát triển dựa trên nền tảng mã nguồn mở Tesseract, với công nghệ Java/.NET, hỗ trợ nhận dạng cho các dạng ảnh PDF, TIFF, JPEG, GIF, PNG, và BMP. Khả năng nhận dạng của VietOCR có thể đạt tới 95% đối với file ảnh có chất lượng tốt.

Nhìn chung, công nghệ OCR đã tạo ra giải pháp kỹ thuật mới, mang tính đột phá trong việc xây dựng cơ sở dữ liệu điện tử. Tuy nhiên, các phần mềm hiện nay chỉ đạt được hiệu quả cao đối với các file văn bản có chất lượng tốt, cùng một số loại hình ngôn ngữ La-tinh.

Đặc biệt với tiếng Việt – loại hình ngôn ngữ có "dấu" (sắc, huyền, hỏi, ngã, nặng), các giải pháp OCR hiện tại bộc lộ rất nhiều hạn chế, cụ thể:

1. Việc nhận dạng dấu trong tiếng Việt đạt hiệu quả thấp. Khi ứng dụng các phần mềm các dấu trong tiếng Việt thường bị đảo vị trí hoặc nhận sai.

2. Các văn bản tiếng Việt thường rất đa dạng về ký tự, chữ đánh máy, chữ viết tay, con số, tiếng Việt xen lẫn ngoại ngữ,… hoặc trên cùng văn bản tồn tại nhiều font chữ, nhiều kiểu định dạng văn bản khác nhau, khiến các phần mền khó xử lý hoặc nhận dạng lỗi.

3. Ngoài ra, các loại hình văn bản hành chính, ngoài phần ký tự còn có con dấu, chữ ký, chữ ký nháy,… đặt ra yêu câu phải khoanh vùng nhận diện, xử lý bóc tách thông tin là thách thức đối với các giải pháp OCR mà chưa có phần mềm nào có thể thực hiện được ở độ chính xác cao.

Một hạn chế rất lớn mà tất cả các phần mềm OCR hiện nay gặp phải là thiếu hẳn tính năng rút trích thông tin (IE)(1) từ văn bản, biểu mẫu. Hầu hết các phần mềm mới chỉ dừng lại việc nhận dạng toàn văn mà không thể rút trích thông tin theo các mẫu/định dạng, các trường theo nhu cầu của người sử dụng – một nhu cầu rất thiết yếu đối với những người làm công tác văn thư, lưu trữ.

2. Giải pháp nhận dạng ký tự quang học QHOCR 1.0

QHOCR 1.0 là phần mềm nhận dạng ký tự quang học được xây dựng và phát triển bởi nhóm Nghiên cứu của Trung tâm Lưu trữ quốc gia II cùng với các chuyên gia công nghệ thông tin. Phần mềm là sự kết hợp các kỹ thuật xử lý căn bản của Tesseract, các kỹ thuật xử lý ảnh nâng cao, xử lý ngôn ngữ tự nhiên, xử lý chữ viết tay rời rạc và con số. Trên cơ sở kế thừa các tính năng ưu việt của công nghệ OCR, phần mềm giải quyết được các hạn chế của các chương trình hiện có và đặc biệt phù hợp với việc nhận dạng ký tự và rút trích thông tin từ văn bản hành chính ở Việt Nam.

Giải pháp QHOCR 1.0 có một số kỹ thuật tiêu biểu như:

Các kỹ thuật xử lý ảnh căn bản

Đơn giản hóa bằng màu sắc tức là phân cụm ảnh gốc theo màu sắc. Bằng việc nhóm các vùng có màu sắc gần nhau thành một sẽ giúp hỗ trợ nhận dạng ảnh nền đặc trưng và phân vùng ảnh văn bản.

Ảnh trước xử lý *Ảnh sau khi*

Phân cụm theo màu sắc và hình dạng giúp nhận diện các khối ảnh của văn bản cần phải xử lý.

Trong ví dụ: Từ ảnh gốc các mảng, khối sẽ được phân cụm, nhận diện

Chuyển trắng đen làm nổi bật văn bản: Sau khi phân đoạn và nhận diện mảng khối bằng các kỹ thuật xử lý phân cụm ảnh, kỹ thuật này sẽ được áp dụng nhằm loại bỏ ảnh nhiễu, chuyển ảnh xám và ảnh nhị phân, làm nổi bật văn bản.

Trong ví dụ: Ảnh gốc đã được xóa nhiễu, làm nổi bật các dòng chữ.

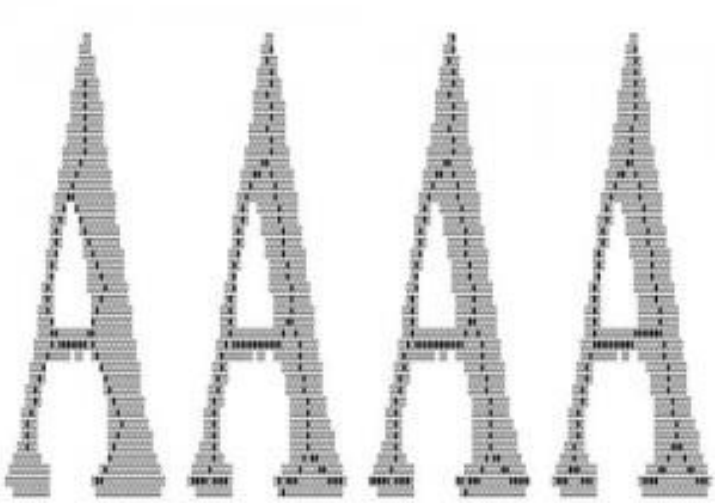
Xử lý ảnh trong văn bản

Để nhận diện văn bản, QHOCR 1.0 tiếp tục áp dụng các kỹ thuật xử lý ảnh cao cấp khác phục vụ riêng cho nhận diện văn bản:

Xác định dòng

Trong ví dụ: các dòng văn bản sẽ được nhận diện dù trong file ảnh chúng bị bẻ cong.

Xác định trục xương của ký tự: Các ký tự sẽ được nhận dạng thông qua phát hiện các trục xương tạo nên hình dáng của ký tự.

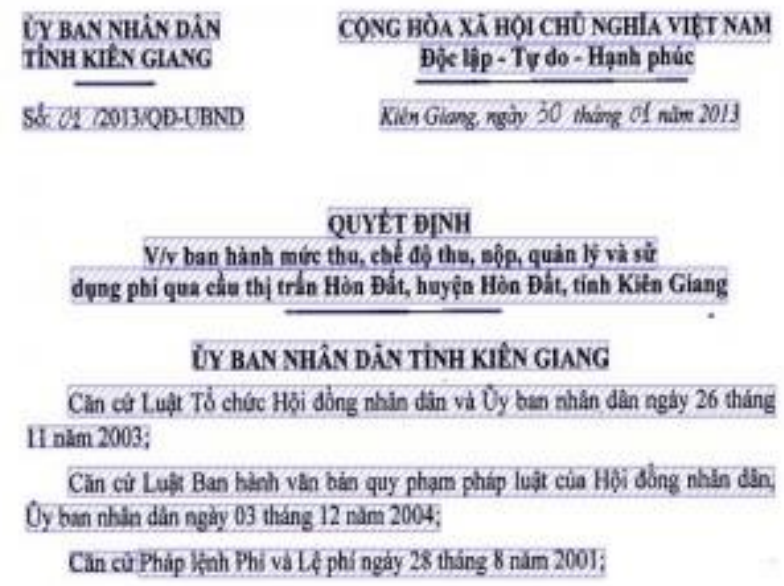


Trong ví dụ: Xương của ký tự A sẽ được nhận diện và xử lý.

Các điểm ưu việt trong xử lý ảnh của giải pháp QHOCR 1.0

Các giải pháp OCR của thế giới và Việt Nam hiện nay chỉ dừng lại ở nhận dạng toàn văn bản và chuyển sang dạng text toàn văn bản của một file ảnh, lỗi chính tả tiếng Việt thường xuyên gặp phải trong các giải pháp kể trên. QHOCR 1.0 đã giải quyết được những hạn chế đó và điểm ưu việt của QHOCR 1.0 là rút trích thông tin chi tiết theo yêu cầu.

Ưu điểm đầu tiên mà QHOCR 1.0 có được là rút trích thông tin chi tiết. Ví dụ, trong một file ảnh văn bản hành chính, muốn rút trích thông tin theo từng dòng như Cộng hòa xã hội chủ nghĩa Việt Nam, dòng Quyết định… áp dụng phần mềm QHOCR 1.0 đều cho kết quả chính xác cả về thông tin và chính tả.



ỦY BAN NHÂN DÂN
TỈNH KIÊN GIANG

Số: 01 /2013/QĐ-UBND

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Kiên Giang, ngày 30 tháng 01 năm 2013

QUYẾT ĐỊNH
V/v ban hành mức thu, chế độ thu, nộp, quản lý và sử dụng phí qua cầu thị trấn Hòn Đất, huyện Hòn Đất, tỉnh Kiên Giang

ỦY BAN NHÂN DÂN TỈNH KIÊN GIANG

Căn cứ Luật Tổ chức Hội đồng nhân dân và Ủy ban nhân dân ngày 26 tháng 11 năm 2003;

Căn cứ Luật Ban hành văn bản quy phạm pháp luật của Hội đồng nhân dân, Ủy ban nhân dân ngày 03 tháng 12 năm 2004;

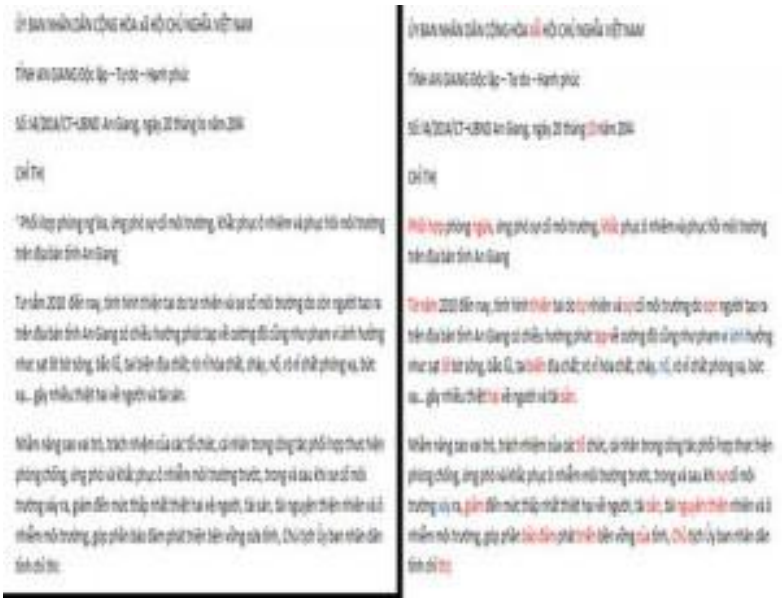
Căn cứ Pháp lệnh Phí và Lệ phí ngày 28 tháng 8 năm 2001;

Trong ví dụ:

Các ký tự đặc biệt:“â”, “ồ”, “ứ”, “ị”, “ế” được khoanh vùng chính xác vì vậy độ chính xác dòng trong văn bản tiếng Việt của giải pháp là 100%.

Các ảnh văn bản mờ, từ bị đứt nét, chữ viết tay được giải quyết trong vấn đề gom cụm. Sở dĩ có kết quả trên là nhờ QHOCR 1.0 đã làm chủ hàng loạt kỹ thuật xử lý ảnh thông minh như: xoay biểu mẫu, xóa hình nền, xác định khung thông tin trọng điểm…

QHOCR 1.0 thể hiện tính ưu điểm khi xử lý tốt chính tả tiếng Việt – một hạn chế của các phần mềm nhận dạng ký tự hiện có. Khi áp dụng các phần mềm OCR cho ra sản phẩm thường bị lỗi chính tả như ví dụ minh họa bên trái. Còn sử dụng QHOCR 1.0 thì các lỗi chính tả sẽ được khắc phục như ví dụ minh họa bên phải:



Chữ màu đỏ là chữ đã được sửa lỗi chính tả. Chữ màu xanh là chữ sai lỗi chính tả chưa sửa được.

Bằng việc áp dụng các thuật toán về chính tả tiếng Việt và các xác suất của các cụm từ đứng cạnh nhau, QHOCR1.0 sau đó sẽ tiếp tục xử lý và cho ra một văn bản theo định dạng Word như hình vẽ:

ỦY BAN NHÂN DÂN
TỈNH AN GIANG

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

Số: 14/2014/CT-UBND

An Giang, ngày 20 tháng 10 năm 2014

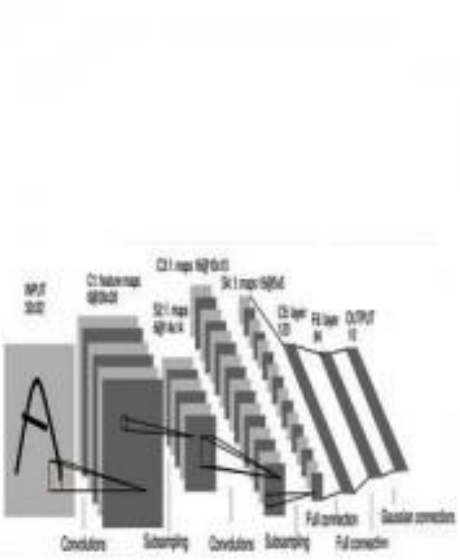
CHỈ THỊ

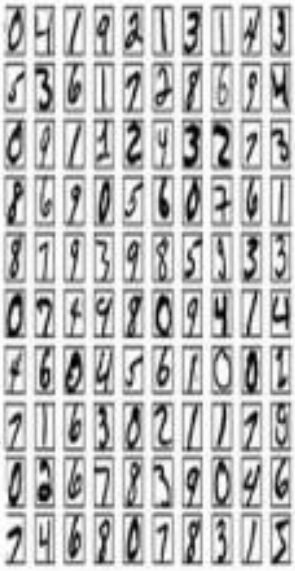
Phối hợp phòng ngừa, ứng phó sự cố môi trường, khắc phục ô nhiễm và phục hồi môi trường trên địa bàn tỉnh An Giang

Từ năm 2010 đến nay, tình hình thiên tai do tự nhiên và sự cố môi trường do con người tạo ra trên địa bàn tỉnh An Giang có chiều hướng phức tạp về cường độ cũng như phạm vi ảnh hưởng như: Sạt lở bờ sông, bão lũ, tai biến địa chất; rò rỉ hóa chất, cháy, nổ, rò rỉ chất phóng xạ, bức xạ... gây nhiều thiệt hại về người và tài sản.

Nhằm nâng cao vai trò, trách nhiệm của các tổ chức, cá nhân trong công tác phối hợp thực hiện phòng chống, ứng phó và khắc phục ô nhiễm môi trường trước, trong và sau khi sự cố môi trường xảy ra, giảm đến mức thấp nhất thiệt hại về người, tài sản, tài nguyên thiên nhiên và ô nhiễm môi trường, góp phần bảo đảm phát triển bền vững của tỉnh, Chủ tịch Ủy ban nhân dân tỉnh chỉ thị:

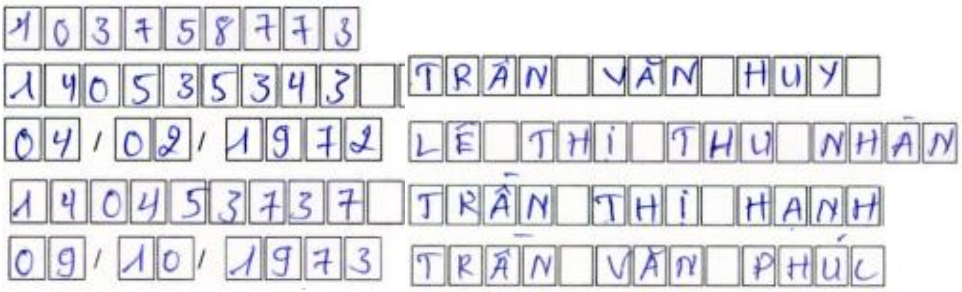
Đối với mô hình học DeepLearning cho tiếng Việt: Mô hình học DeepLearning hiện đang là một trong các kỹ thuật tốt nhất hiện tại trong lĩnh vực nhận dạng.





QHOCR 1.0 đã áp dụng mô hình này vào nhận dạng chữ số viết tay, nhận dạng các mẫu chữ tiếng Việt cũ, mờ, nhận dạng chữ viết tay rời.

Ví dụ về các mẫu mà QHOCR1.0 đã giải quyết.



Độ chính xác đạt được là: 98%. Tuy nhiên, ứng dụng này đòi hỏi thời gian xử lý vì phải xử lý ảnh và khoanh vùng.

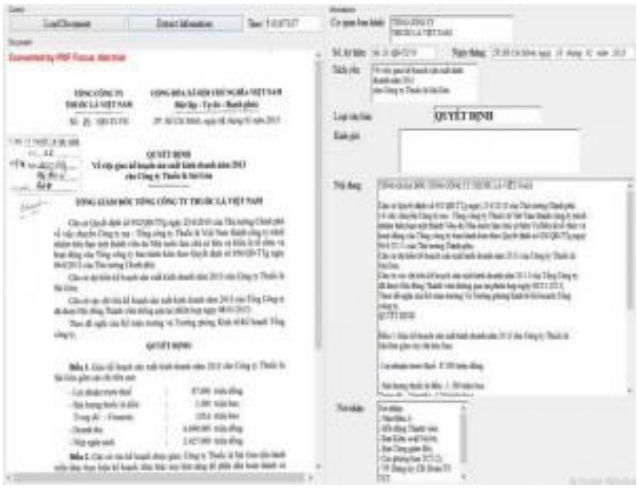
Ngoài những ưu điểm của QHOCR 1.0 vừa kể trên, QHOCR 1.0 còn có các ưu điểm khác như: hỗ trợ phần rút trích thông tin từ ảnh văn bản; hỗ trợ tốt nhận dạng ảnh văn bản màu; khoanh vùng con dấu; xử lý nét viết bút bi; xây dựng các dữ liệu học riêng cho tiếng Việt và phông chữ tiếng Việt.

3. Áp dụng QHOCR 1.0 vào quy trình biên mục tài liệu lưu trữ

Trong công tác văn thư lưu trữ, biên mục hồ sơ và biên mục văn bản là hai hoạt động thường xuyên và rất quan trọng. Đây là cơ sở để xây dựng hệ thống công cụ tra cứu điện tử và quản lý khoa học tài liệu lưu trữ. Tuy nhiên, hiện nay, công tác biên mục tài liệu lưu trữ (bao gồm cả biên mục hồ sơ và biên mục văn bản) vẫn đang được thực hiện thủ công, chưa có phần mềm nào có thể đáp ứng được các yêu cầu của một quy trình biên mục, đặc biệt là đối với biên mục văn bản.

Khi ứng dụng QHOCR 1.0 với việc rút trích thông tin từ file ảnh thành các dữ liệu dạng text phù hợp các trường thông tin trong biên mục tài liệu lưu trữ sẽ tạo bước đi mới đối với việc ứng dụng công nghệ thông tin vào công tác biên mục tài liệu. QHOCR 1.0 có khả năng nhận dạng ký tự và rút trích thông tin với độ chính xác đạt xấp xỉ 97% trên cả các file ảnh có chất lượng trung bình.

Hình ảnh dưới đây thể hiện giải pháp rút trích thông tin (IE) của phần mềm QHOCR 1.0 theo các trường thông tin của quy trình biên mục văn bản:



Phần mềm QHOCR 1.0 cho phép tự động nhận dạng các vùng thông tin quan trọng trong văn bản như: số ký hiệu, cơ quan ban hành, nơi nhận, loại văn bản, trích yếu văn bản, ngày ban hành, …Hệ thống có khả năng nhận dạng các chữ số viết tay trong phần ngày tháng, số ký hiệu,…Hệ thống tự động khoanh vùng, xác định tên loại văn bản để rút trích các trường thông tin cần thiết và sửa lỗi chính tả các thông tin cần rút trích. Chính vì thế, khi áp dụng phần mềm QHOCR 1.0 trong công tác biên mục văn bản, sẽ giảm thao tác nhập máy thủ công như hiện nay. Theo tính toán, trong một phút, phần mềm này có thể thực hiện được 4 văn bản bằng khoảng 10 người nhập dữ liệu thông thường trong biên mục văn bản.

QHOCR 1.0 có tính ứng dụng cao trong nhiều lĩnh vực khi muốn rút trích thông tin từ các file ảnh văn bản. Nó đặc biệt cần thiết trong công tác số hóa, biên mục hồ sơ và xây dựng hệ thống tìm kiếm các văn bản hành chính quốc gia.

Kết luận

Phần mềm nhận dạng ký tự quang học (OCR) và giải pháp rút trích thông tin (IE) có giá trị thực tiễn cao trong đời sống xã hội khi nhu cầu ứng dụng các sản phẩm công nghệ thông tin càng nhiều trong nhiều lĩnh vực. OCR và IE hỗ trợ con người trong việc tiết kiệm thời gian và công sức khi muốn trích dẫn thông tin từ một file ảnh văn bản. Khi OCR và IE chưa ra đời, thì việc lấy thông tin từ một file ảnh văn bản phải thực hiện thủ công bằng cách là đọc file ảnh và nhập máy các thông tin cần lấy từ file ảnh đó. Đây có thể nói là một hoạt động mang tính thủ công trong thời đại công nghệ số phát triển.

Tuy nhiên, OCR và IE của thế giới mang tính phổ biến lại không thể ứng dụng một cách hiệu quả vào nhận dạnh file ảnh văn bản tiếng Việt vì đặc điểm ngôn ngữ và sự đa dạng trong các loại hình văn bản ở Việt Nam.

Từ những nghiên cứu ưu và khuyết của OCR và EI của thế giới trong ứng dụng đối văn bản tiếng Việt, chúng tôi cho ra đời QHOCR 1.0. Phần mềm này có thể xem là hoàn toàn mới do những cải tiến và tính ứng dụng cao trong xử lý nhận dạng hình ảnh văn bản hành chính. Đây là một giải pháp tốt cho nhận dạng tiếng Việt vì nó có 3 khả năng: xử lý ảnh theo sát định dạng của biểu mẫu, đặc biệt là các loại hình văn bản của Việt

Nam; tích hợp kết quả về chính tả, xử lý ngôn ngữ tự nhiên dành cho tiếng Việt; mô hình học Deep Learning cho tiếng Việt gồm cả văn bản viết tay rời, phông cũ, số.

Tóm lại, QHOCR 1.0 là một phần mềm cải tiến ứng dụng rút trích thông tin trong file ảnh văn bản tiếng Việt. Nó đặc biệt thích hợp ứng dụng vào biên mục tài liệu văn bản vì tính ưu việt trong rút trích các trường thông tin theo quy trình biên mục. Phần mềm này được sử dụng sẽ làm tăng năng suất lao động, giảm số lượng lao động, tiết kiệm thời gian và chi phí trong biên mục tài liệu lưu trữ./.

Vũ Văn Tâm – Trung tâm Lưu trữ quốc gia II

Bùi Ngọc Lê – Đại học Hoa Sen