**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**NGÀNH KHOA HỌC MÁY TÍNH**

**Nguyễn Trọng Triết – 1753113**

**Nguyễn Vương Đạt Vũ – 1753131**

**KHÓA LUẬN TỐT NGHIỆP**

**NHẬN DIỆN ĐƠN THUỐC SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU**

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN

CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

**GIÁO VIÊN HƯỚNG DẪN**

ThS. Lê Ngọc Thành

TP. Hồ Chí Minh, Ngày Tháng Năm 2021

MỤC LỤC

[**LỜI CẢM ƠN** 3](#_Toc53755474)

[**DANH MỤC HÌNH ẢNH** 4](#_Toc53755475)

[**DANH MỤC BẢNG BIỂU** 5](#_Toc53755476)

[**DANH MỤC THUẬT NGỮ** 6](#_Toc53755477)

[**TÓM TẮT KHÓA LUẬN** 7](#_Toc53755478)

[**CHƯƠNG 1: MỞ ĐẦU** 8](#_Toc53755479)

[**1.1** **Giới thiệu đề tài** 8](#_Toc53755480)

[**1.2** **Lý do chọn đề tài** 8](#_Toc53755481)

[**1.3** **Nghiên cứu liên quan** 9](#_Toc53755482)

[**1.4** **Hướng tiếp cận và giải quyết vấn đề** 9](#_Toc53755483)

[**1.5** **Mục tiêu luận văn** 9](#_Toc53755484)

[**1.6** **Phạm vi đề tài** 10](#_Toc53755485)

[**1.7** **Cấu trúc của luận văn** 10](#_Toc53755486)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 11](#_Toc53755487)

# **LỜI CẢM ƠN**

Lời đầu tiên, chúng em xin trân trọng gửi lời cảm ơn sâu sắc và chân thành nhất tới ThS. Lê Ngọc Thành đã chỉ dạy và hướng dẫn nhiệt tình trong quá trình học cũng như hoàn thành luận văn.

Chúng em xin chân thành cảm ơn quý thầy cô và ban giám hiệu trường đại học Khoa Học Tự Nhiên, đặc biệt là thầy cô khoa Công nghệ thông tin đã tận tình giảng dạy chúng em trong suốt 4 năm đại học. Đó không chỉ là kiến thức mà còn là hành trang để chúng em bước vào đời.

Chúng em xin cảm ơn quý nhà trường đã tạo điều kiện thuận lợi cho chúng em học tập và thực hiện đề tài tốt nghiệp này.

Chúng em xin cảm ơn sự quan tâm giúp đỡ và ủng hộ của các anh chị bạn bè trong quá trình thực hiện khóa đề tài. Mặc dù đã cố gắng hoàn thành khóa luận trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót.

Chúng em rất mong nhận được sự thông cảm, góp ý và tận tình chỉ bảo của quý thầy cô và các bạn.

TP. Hồ Chí Minh, Ngày Tháng Năm 2021

# **DANH MỤC HÌNH ẢNH**

# **DANH MỤC BẢNG BIỂU**

# **DANH MỤC THUẬT NGỮ**

# **TÓM TẮT KHÓA LUẬN**

Trong 10 năm gần đây, với sự phát triển của phương pháp học sâu nói riêng cũng như ngành trí tuệ nhân tạo nói chung đã tạo ra sự thu hút sự quan tâm đặc biệt đối với các sinh viên ngành công nghệ thông tin, các nghiên cứu sinh, các nhà khoa học… Trong đó, bài toán nhận diện ký tự quang học, đang được nhóm chúng em quan tâm. Luận văn này trình bày về nghiên cứu và xây dựng 1 hệ thống nhận diện và trính xuất thông tin của đơn thuốc, trong đó gồm thông tin của thuốc(được so sánh với ngân hàng thuốc Việt Nam). Trên cơ sở các nghiên cứu với điều kiện thực tế của các công cụ xử lý ngôn ngữ tiếng Việt, nhóm đưa ra đề xuất sử dụng các mô hình học máy để nhận diện, phân tách, trích xuất và so sánh ký tự nhằm đưa ra tỷ lệ chính xác nhất. Thực nghiệm và đánh giá các mô hình cho thấy kết quả khả quan tạo tiền đề cho các phát triển tiếp theo trong tương lai.

# **CHƯƠNG 1: MỞ ĐẦU**

## Giới thiệu đề tài

Đơn thuốc là tài liệu chỉ định dùng thuốc của bác sĩ cho người bệnh; Là cơ sở pháp lý cho việc chỉ định sử dụng thuốc, bán thuốc và cấp thuốc theo đơn. Các công việc liên quan tới đơn thuốc rất quan trọng, đặc biệt là việc lấy thông tin của thuốc. Việc thường thấy là bệnh nhân mang đơn thuốc bác sĩ đã in ra các cửa hàng thuốc để tìm mua thuốc hoặc là người bán thuốc tìm thông tin của thuốc qua đơn thuốc.

Trên thực tế, các công việc kiểm tra tên thuốc bằng máy vẫn chưa phổ biến và việc này vẫn đang thực hiện bằng cách thủ công bằng cách các dược sĩ sẽ nhập tên thuốc lên hệ thống để kiểm tra. Nếu số lượng đơn thuốc lớn thì cần rất nhiều thời gian để nhập tay từng loại thuốc hoặc cần thêm nhiều nguồn nhân lực để đẩy nhanh tiến độ.

Ngoài ra, khi kiểm tra thuốc bằng cách thủ công có thể dẫn đến sai sót cao đồng thời kéo theo rủi ro sai tên thuốc có thể gây nguy hiểm cho bệnh nhân.

Dựa vào sự phát triển của công nghệ 4.0, với sự giúp đỡ của các thiết bị công nghệ sẽ giúp con người giải quyết các công việc cần độ chính xác cao và cần sự nhanh chóng.

## Lý do chọn đề tài

Nhằm giúp ngành y tế phát triển, nhóm chúng em quyết định nghiên cứu một giải pháp giúp cho việc lấy thông tin trên đơn thuốc qua nhưng tấm hình đơn thuốc bình thường, chất lượng cao. Sau đó, có thể trích xuất tên loại thuốc rồi so sánh với ngân hàng thuốc của bộ y tế để đưa ra tên thuốc chính xác nhất. Công việc này có thể giúp cho người bệnh tìm hiểu các loại thuốc mình sử dụng và giúp cho các dược sĩ lấy được thông tin thuốc rồi kiểm tra với hệ thống thuốc trong cửa hàng.

## Nghiên cứu liên quan

Đọc và trích xuất văn bản thông qua hình ảnh là giải pháp không mới và đã có nhiều công trình nghiên cứu đã được sử dụng vào đời sống. Tuy nhiên, áp dụng công nghệ này vào ngành y tế nhằm trích xuất thông tin đơn thuốc vẫn chưa được phổ biến.

Hiện nay có rất nhiều thư viện nhận diện ký tự quang học phổ biến như OCR engines ABBYY Finereader, OCRopus, Tesseract, Calamari, EasyOCR... Nhưng đối với các ký tự có dấu của Tiếng Việt thì các thư viện này hỗ trợ nhận diện ký tự với độ chính xác không cao. Để đạt được độ chính xác cao thì người dung cần huấn luyện mô hình bộ dữ liệu Tiếng Việt riêng. Và để hỗ trợ cho khả năng đọc được tốt hơn thì hình ảnh đầu vào cần có độ tương phản cao, nên là chữ đen nền trắng và giảm thiểu các điểm nhiễu trên hình.

Luận văn này sẽ tiếp tục theo những hướng nghiên cứu trên bằng cách huấn luyện mô hình máy học để đọc Tiếng Việt và kết hợp mô hình đọc Tiếng Anh có sẵn để phù hợp với đọc tên thuốc. Ngoài ra vẫn sẽ có các công việc tiền xử lý hình ảnh và xác định vùng thông tin cần đọc. Thông tin sau khi được đọc sẽ so sánh với ngân hàng dữ liệu nhằm đưa ra được kết quả chính xác nhất.

## Hướng tiếp cận và giải quyết vấn đề

Đọc thông tin đơn thuốc qua hình chụp được chia thành ba phần:

1. Tiền xử lý hình ảnh đưa vào. Hình đưa vào sẽ đc xử lý và cắt giữ đơn thuốc.
2. Xác định vùng chứa văn bản và tìm vùng thông tin chứa tên tên thuốc.
3. Nhận diện ký tự trong vùng chứa văn bản và nâng cao tỉ lệ chính xác của tên thuốc.

## Mục tiêu luận văn

Các mục tiêu sinh viên tự đề ra để hoàn thành tốt đề tài luận văn:

* Xây dựng model nhận dạng đơn thuốc thông qua hình chụp. Hình ảnh đầu vào phải được tiền xử lý và bao quát được đươc đơn thuốc.
* Khoanh vùng các ký tự cần đọc trên hình ảnh.
* Ngân hàng dữ liệu phải luôn được cập nhật để thuận tiện cho việc so sánh.
* Đạt dược độ chính xác cao nhất với tên thuốc.
* Giảm thiểu thời gian đọc dữ liệu.
* Hỗ trợ vừa đọc được tên thuốc Tiếng Việt và Tiếng Anh.

## Phạm vi đề tài

Luận văn của nhóm được giới hạn trong phạm vi đọc các nội dung của đơn thuốc thông qua ảnh được chụp và đưa ra được tên thuốc chính xác nhất trên đơn thuốc. Nhóm sẽ tạo ra một ứng dụng có thể đáp ứng được yêu cầu trên gồm chụp ảnh hoặc tải ảnh lên để xử lý và trả về kết quả là thông tin của tên loại thuốc. Nội dung chính được sử dụng gồm các chữ số, ngôn ngữ Tiếng Việt và Tiếng Anh. Phạm vi trên không bao gồm các đơn thuốc viết tay mà là sử dụng các đơn thuốc được in thành văn bản.

## Cấu trúc của luận văn

# **CHƯƠNG 2: LÝ THUYẾT NỀN TẢNG**

## Giới thiệu về nhận dạng kí tự quang học(OCR)

Nhận dạng kí tự quang học là ứng dụng công nghệ để đọc các hình ảnh của chữ viết tay hoặc chữ đánh máy có định dạng: PDF, JPG, JPEg, PMB... thành các văn bản tài liệu. Phần mềm nhận diện quang học thực hiện được điều này nhờ công nghệ xử lý ngôn ngữ tự nhiên (NLP), giúp nhận dạng các chữ cái và ký tự và sau đó sao chép chúng theo định dạng và thứ tự được viết.

Nhận diện kí tự quang học chuyên dùng để nhập dữ liệu và truyền tải dữ liệu. Nhờ phần mềm nhận dạng chữ OCR, những văn bản số hóa, tìm kiếm và chỉnh sửa sẽ được thực hiện điện tử. Đồng thời, chúng giúp tiết kiệm không gian lưu trữ tài liệu bằng việc hiển thị trên trực tiếp.

Công nghệ nhận diện ký tự quang học(OCR) được hình thành từ lĩnh nghiên cứu về nhận dạng mẫu, trí tuệ nhân tạo và thị giác máy tính. Trước khi nhận diện ký tự quang học(OCR) và trí tuệ nhân tạo(AI) kết hợp với nhau, nhận dạng quang học đã được sử dụng rộng rãi.

Ngày nay nhận diện ký tự quang học (OCR) được sử dụng rộng rãi văn bản viết tay hoặc đánh máy thành các văn bản mềm có thể chỉnh sửa được. Ngoài ra, chất lượng của nhận diện ký tự quang học (OCR) càng ngày được cải thiện đáng kể dẫn đến độ chính xác của kết quả nhận diện cũng được tăng lên. Các doanh nghiệp, công ty lớn nhỏ hiện đang rất ưa chuộng phần mềm nhận diện quang học để tăng hiệu quả trích xuất văn bản.

Trong nhận dạng kí tự, các văn bản được coi là đã được phát hiện, định vị và các khung hình chữ nhật chứa văn bản có sẵn.

Bất kỳ hệ thống nhận diện kí tự quang học (OCR) nào cũng tuân theo các bước sau:

A picture containing diagram

Description automatically generated

**Image Preprocessing( Tiền sử lý)**

1. Loại bỏ nhiễu khỏi hình ảnh
2. Xóa nền phức tạp khỏi hình ảnh
3. Xử lý các điều kiện khác nhau trong ảnh

**Text Detection (Phát hiện văn bản )**

Phát hiện văn bản là quá trình phát hiện các đoạn văn bản có trong hình ảnh, trả về kết quả bao quanh nó với một khung hình chữ nhật.

A picture containing text, indoor

Description automatically generated

***Text Recognition ( Nhận dạng văn bản)***

Khi chúng ta đã phát hiện các hộp giới hạn có văn bản, bước tiếp theo là nhận dạng văn bản.Quá trình nhận dạng là quét các biểu tượng gặp và phân chúng thành kí tự cơ bản.

A picture containing logo

Description automatically generated

## Các giải pháp hiện nay.

Việc nhận dạng kí tứ quang học gồm 2 phần: phát hiện văn bản và nhận diện văn bản

### Phát hiện văn bản (Text detection)

#### Mạng văn bản kết nối(CTPN)

Mạng văn bản kết nối(CTPN) phát hiện một dòng văn bản trong 1 chuỗi các đề xuất văn bản trong features map. Các đề xuất tuần tự được kết nối tự nhiên bởi một mạng nơ-ron tuần hoàn, được kết hợp liền mạch vào mạng phức hợp, dẫn đến một mô hình có thể đào tạo từ đầu đến cuối. Điều này cho phép CTPN khám phá thông tin ngữ cảnh phong phú của hình ảnh, làm cho nó trở nên mạnh mẽ để phát hiện văn bản cực kỳ mơ hồ. CTPN hoạt động đáng tin cậy trên văn bản đa quy mô và đa ngôn ngữ mà không cần xử lý hậu kỳ thêm, bắt đầu từ các phương pháp từ dưới lên trước nhiều bước lọc.

Text, letter

Description automatically generated

Calendar

Description automatically generatedƯu điểm của CTPN là phát hiện được nguyên một hàng chứa văn bản, giúp cho văn bản được nhận dạng không bị mất tính liên kết. Bounding box của CTPN khá lơn, điều này tạo thuận lợi cho việc phát hiện các từ Tiếng Việt vì Tiếng Việt bao gồm dấu. Nhưng CTPN có một nhược điểm khá lớn là với những tấm hình không trực diện hoặc bị nghiêng thì CTPN hoạt động không tốt.

#### Craft

CRAFT phát hiện hiệu quả vùng văn bản bằng cách khám phá từng vùng ký tự và mối quan hệ giữa các ký tự. Hộp văn bản giới hạn có được bằng cách đơn giản tìm các hình chữ nhật giới hạn tối thiểu trên bản đồ nhị phân sau khi ngưỡng vùng ký tự và điểm số sở thích

Text, letter

Description automatically generated

Ưu điểm của Craft là có độ chính xác lớn, có thể phát hiện hầu hết các từ trong văn bản lên đến 98%. Craft có một điểm mạnh rất lớn là bất biến với các phép xoay. Với những hình không trực diện hoặc bị nghiêng thì Craft vẫn pháp hiện rất tốt. Điểm yếu của Craft thường gặp là vì nó là mô hình phức tạp nên tốc độ phát hiện còn chậm, các bounding box sẽ bao quanh từ thay vì thành một hang. Ngoài ra, trong một vài trường hợp Craft sẽ không phát hiện được chữ số.

Text

Description automatically generated

### Nhận diện văn bản (Text regconition)

#### Tesseract OCR

Tesseract là ứng dụng nhận diện ký tự đa ngôn ngữ, trích xuất nội dung văn bản từ các bản scan, hình chụp được phát triển bởi Google. Ứng dụng này cho kết quả chính xác cao khi nhận diện hình có chữ màu đen trên nền trắng. Bằng các thuật toán xử lý ảnh, ảnh đơn thuốc được biến đổi thành hình đen trắng với hai kênh màu. Sau đó đưa vào mô hình Tesseract để nhận diện

Mô hình này có thể cho kết quả chính xác với những hình được chụp đúng góc, chất lượng hình ảnh tốt. Nhưng không thể nhận diện với các hình chụp nghiêng, ngược. Đồng thời mô hình này chỉ trích xuất được toàn nội dung, toàn bộ chữ trong hình. Không trích xuất được theo mẫu

#### Easy OCR

## Lựa chọn giải pháp phù hợp để giải quyết vấn đề

#### Như đã biết trong các bài toán nhận dạng đối tượng hay văn bản, sẽ quy bài toán về 2 phần: phát hiện và nhận dang. Cách hoạt động của mỗi phần thường chọn một mô hình tốt để tối ưu quá trình phát hiện hoặc nhận dạng nhằm cho ra kết quả tốt nhất.

#### Về phần phát hiện, nhóm chúng em sẽ sử dụng mô hình CRAFT detection để phát hiện vùng chứa kí tự văn bản. So với các thuật toán phát hiện còn lại thì CRAFT cho độ chính xác cao hơn và mô hình này còn bất biến với phép xoay. Giúp cho việc giải quyết bài toán không cần thông qua công việc xử lý ảnh. Những dòng chữ trên đơn thuốc được phát hiện đầy đủ và chính xác nên ở phần này không cần phải huấn luyện lại, nhằm đẩy nhanh quá trình xuất ra kết quả cho quá trình nhận dạng.

#### Về phần nhận dang, chúng em đã sử dụng mô hình nhận dạng Tesseract OCR để nhận dạng và cho ra kết quả với độ chính xác hơn 90%. Ngoài ra để có được tên thuốc đúng nhất, bọn em sử dụng thêm khoảng cách Levenshtein nhằm so sánh và chỉnh sửa tên thuốc vừa nhận dạng được.

# **CHƯƠNG 3: TESSERACT**

## Giới thiệu tổng quan về Tesseract OCR

Tesseract là một công cụ mã nguồn mở để nhận diện ký tự quang học. Tesseract được phát triển bởi HP trong khoảng từ năm 1984 đến năm 1994 và đã được sửa đổi và cải tiến vào năm 1995 với độ chính xác cao hơn. Cuối năm 2005, HP phát hành Tesseract cho mã nguồn mở và tập trung hơn vào việc cải thiện độ chính xác. Bây giờ Tesseract đang được phát triển và duy trì bởi Google. Tesseract là một công cụ OCR có hỗ trợ unicode và khả năng nhận dạng hơn 100 ngôn ngữ. Ngoài ra, Tesseract có thể được huấn luyện để nhận ra các ngôn ngữ khác.

## Kiến trúc Tesseract OCR

# Tesseract OCR hoạt động tuần tự theo các bước như trong sơ đồ khối. Bước đầu tiên là ngưỡng thích ứng, chuyển đổi hình ảnh thành hình ảnh nhị phân. Bước tiếp theo là phân tích thành phần được kết nối, được sử dụng để trích xuất phác thảo ký tự. Phương pháp này rất hữu ích vì nó thực hiện OCR của hình ảnh với văn bản màu trắng và nền đen.

# Tesseract là chương trình đầu tiên cung cấp kiểu xử lý này. Sau đó, các phác thảo được chuyển đổi thành các blob. Các blob được tổ chức thành các dòng văn bản, các dòng và vùng được phân tích cho một số vùng cố định hoặc kích thước văn bản tương đương. Văn bản được chia thành các từ sử dụng các khoảng trắng xác định và các khoảng mờ. Việc nhận dạng văn bản sau đó được bắt đầu như một quá trình hai lần như thể hiện trong hình.

Diagram

Description automatically generated

# Trong lần chuyển đầu tiên, tesseract sẽ cố gắng để nhận dạng từng từ trong văn bản. Mỗi từ đạt yêu cầu được chuyển tới bộ phân loại thích ứng dưới dạng dữ liệu huấn luyện. Bộ phân loại thích ứng cố gắng nhận dạng văn bản theo cách chính xác hơn. Vì bộ phân loại thích ứng đã nhận được một số dữ liệu huấn luyện, nó đã học được điều gì mới nên giai đoạn cuối cùng được sử dụng để giải quyết các vấn đề khác nhau và trích xuất văn bản từ hình ảnh.

## Thuật toán chuyển đổi ảnh thành gray scale.

Một hình ảnh kỹ thuật số có chiều rộng là M và chiều cao là N được biểu diễn dưới dạng bởi hàm như sau:

Ở đây các cặp được gọi là các pixel. Cặp (0,0) là pixel đầu tiên và cặp ( là pixel cuối cùng trong ảnh. Mỗi pixel đều có giá trị màu RGB riêng. Nếu pixel có cùng giá trị RGB thì nó sẽ được xếp vào họ màu xám (từ đen sang trắng). Vì vậy, dựa trên quan sát này, thuật toán chuyển đổi hình ảnh màu sang thang màu xám được phát triển, được trình bày dưới đây:

Với lần lượt là các giá trị màu đỏ, màu lục, màu lam của pixel . là giá trị trung bình của các pixel này luôn nhỏ hơn 256. Vì vậy, giá trị của nằm trong khoảng từ 0 đến 255, được gán cho màu đỏ, xanh lục và xanh lam pixel của pixel (x, y). Quá trình này được mô tả như sau:

Áp dụng thuật toán trên, ảnh màu sẽ trở thành hình ảnh gray scale.

## Từ điển tên thuốc

Bộ dữ liệu của nhóm sử dụng làm từ điển thuộc ngân hàng thuốc Việt Nam. Trong từ điển bao gồm tên thuốc, thành phần thuốc và thông số(định lượng của thuốc). Với mỗi tên thuốc và thành phần thuốc có thể có nhiều thông số khác nhau. Nếu 2 tên thuốc giống nhau mà có 2 thành phần thuốc khác nhau thì ta sẽ biết nó là 2 thuốc hoàn toàn khác nhau.