



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẢNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0032643

(51)⁷ G06N 3/02; G06K 9/03; G06N 20/00;
G06F 17/21; G06K 9/72

(13) B

(21) 1-2019-07082

(22) 13/12/2019

(30) 10201904825X 28/05/2019 SG

(45) 25/07/2022 412

(43) 25/12/2020 393

(73) Advanced New Technologies Co., Ltd. (KY)

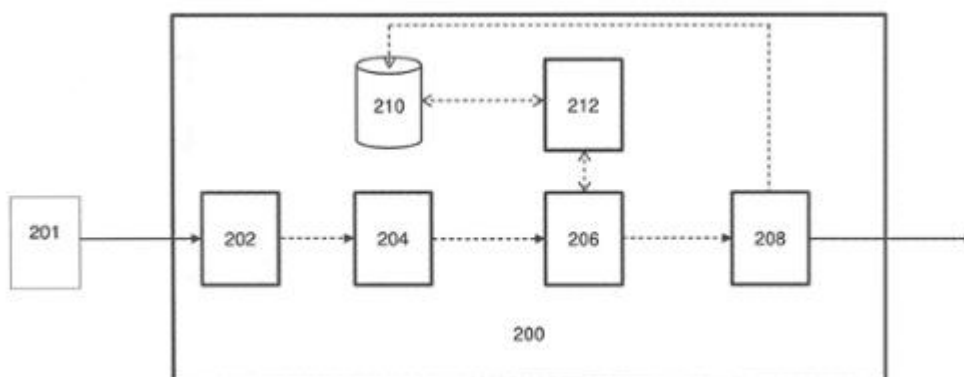
Cayman Corporate Centre, 27 Hospital Road, George Town, Grand Cayman KY1-9008, Cayman Islands

(72) Ruoyu Li (CN).

(74) Công ty TNHH Tầm nhìn và Liên danh (VISION & ASSOCIATES CO.LTD.)

(54) PHƯƠNG PHÁP, HỆ THỐNG HIỆU CHỈNH NHẬN DẠNG KÝ TỰ QUANG HỌC (OCR) TỰ ĐỘNG VÀ VẬT GHI ĐỌC ĐƯỢC BẰNG MÁY TÍNH, LÂU DÀI

(57) Sáng chế đề cập đến hệ thống nhận dạng ký tự quang học (OCR), bao gồm: thiết bị thu thập được tạo cấu hình để thu được ảnh kỹ thuật số của tài liệu vật lý; thiết bị chuyển đổi ảnh được tạo cấu hình để chuyển đổi ảnh kỹ thuật số của tài liệu vật lý này thành văn bản tương ứng được đọc được bằng máy; thiết bị hiệu chỉnh được tạo cấu hình để: đánh giá văn bản đọc được bằng máy này bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh văn bản đọc được bằng máy hay không; nếu cần phải hiệu chỉnh văn bản đọc được bằng máy đó, thì xác định văn bản tương tự nhất với văn bản đọc được bằng máy đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi; và hiệu chỉnh văn bản đọc được bằng máy này bằng văn bản tương tự nhất đã xác định được; và thiết bị đầu ra được tạo cấu hình để kết xuất văn bản đọc được bằng máy đã hiệu chỉnh.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập chung đến, nhưng không loại trừ, kỹ thuật hiệu chỉnh nhận dạng ký tự quang học (Optical Character Recognition - OCR) tự động.

Tình trạng kỹ thuật của sáng chế

Định danh khách hàng điện tử (electronic-Know Your Customer - eKYC) là quy trình thẩm tra kỹ thuật số được doanh nghiệp thực hiện nhằm xác minh danh tính của khách hàng và đánh giá các rủi ro tiềm ẩn về các ý đồ bất hợp pháp đối với quan hệ kinh doanh (ví dụ như rửa tiền).

Kỹ thuật nhận dạng ký tự quang học (OCR) là kỹ thuật nhận dạng các văn bản trong tài liệu vật lý và chuyển đổi chúng thành các văn bản được mã hoá bằng máy.

Khi thực hiện eKYC, kỹ thuật OCR có thể được sử dụng để nhận dạng các tài liệu định danh (ID) chính thức, ví dụ như thẻ căn cước, hộ chiếu, v.v.. Các kỹ thuật OCR hiện nay có thể đạt được tỷ lệ chính xác cao hợp lý; tuy nhiên đôi khi cũng xuất hiện lỗi. Do đó, bước xác định các kết quả OCR cần được hiệu chỉnh và tự động hiệu chỉnh các kết quả đó là rất hữu ích.

Quy trình hiệu chỉnh OCR tự động có thể bao gồm bước trực tiếp tìm kiếm kho ngữ liệu tên/địa chỉ để phát hiện tên/địa chỉ tương tự nhất để hiệu chỉnh. Khi tỷ lệ bao quát của kho ngữ liệu tên và địa chỉ không thể được đảm bảo, thì bước tìm kiếm trực tiếp có thể dẫn đến việc nhiều kết quả đúng bị hiệu chỉnh thành sai.

Bản chất kỹ thuật của sáng chế

Theo một phương án, mô hình ngôn ngữ được huấn luyện bởi kho ngữ liệu tên và địa chỉ đã xây dựng được sử dụng để xác định xem liệu kết quả OCR có cần được hiệu chỉnh hay không. Nếu cần phải kết quả OCR này, thì quy trình khoảng cách chỉnh sửa biến đổi được sử dụng để xác định độ tương tự theo trực quan của các văn bản. Các kết quả OCR (được hiệu chỉnh và không được hiệu chỉnh) được thêm vào kho ngữ liệu để làm giàu kho ngữ liệu này và nâng cao hiệu suất cho mô hình ngôn ngữ và quy trình khoảng cách chỉnh sửa biến đổi.

Theo một phương án khác, sáng chế đề xuất hệ thống nhận dạng ký tự quang học (OCR), hệ thống này bao gồm: thiết bị thu thập được tạo cấu hình để thu được ảnh kỹ thuật số của tài liệu vật lý; thiết bị chuyển đổi ảnh được tạo cấu hình để chuyển đổi ảnh kỹ thuật số của tài liệu vật lý đó thành văn bản tương ứng đọc được bằng máy; thiết bị hiệu chỉnh được tạo cấu hình để: đánh giá văn bản đọc được bằng máy bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (Long short-term memory - LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh văn bản đọc được bằng máy đó hay không; nếu cần phải hiệu chỉnh văn bản đọc được bằng máy đó, thì xác định văn bản tương tự nhất với văn bản đọc được bằng máy đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi; và hiệu chỉnh văn bản đọc được bằng máy đó bằng văn bản tương tự nhất đã xác định được; và thiết bị đầu ra được tạo cấu hình để kết xuất văn bản đọc được bằng máy đã hiệu chỉnh.

Hệ thống OCR này có thể còn có: thiết bị cơ sở dữ liệu có lưu trữ trong đó tập hợp tên và địa chỉ liên kết với miền đích, trong đó thiết bị cơ sở dữ liệu này được tạo cấu hình để xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ đã lưu trữ liên kết với miền đích; và thiết bị huấn luyện được tạo cấu hình để huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đó. Thiết bị đầu ra có thể còn được tạo cấu hình để thêm ít nhất là văn bản đọc được bằng máy đã hiệu chỉnh vào tập hợp tên và địa chỉ lưu trữ trong thiết bị cơ sở dữ liệu.

Thiết bị cơ sở dữ liệu có thể còn được tạo cấu hình để tái dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ chứa văn bản đọc được bằng máy đã hiệu chỉnh. Thiết bị huấn luyện có thể còn được tạo cấu hình để huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng. Thiết bị hiệu chỉnh có thể còn được tạo cấu hình để đánh giá log(tần suất cặp chỉnh sửa trong lịch sử) để xác định văn bản tương tự nhất với văn bản đọc được bằng máy cần hiệu chỉnh.

Theo một phương án khác, sáng chế đề xuất hệ thống hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động, hệ thống này bao gồm: thiết bị xử lý; và thiết bị ghi nhớ gồm mã chương trình máy tính; thiết bị ghi nhớ và mã chương trình máy tính được tạo cấu hình để, thông qua thiết bị xử lý, cho phép hệ thống thực hiện ít nhất là thao tác: đánh giá kết quả OCR bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh kết quả OCR

này hay không; nếu cần phải hiệu chỉnh kết quả OCR đó, thì xác định văn bản tương tự nhất với kết quả OCR đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi; và hiệu chỉnh kết quả OCR đó bằng văn bản tương tự nhất đã xác định được.

Theo một phương án khác, sáng chế đề xuất phương pháp hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động, bao gồm các bước: đánh giá kết quả OCR bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh kết quả OCR này hay không; nếu cần phải hiệu chỉnh kết quả OCR đó, thì xác định văn bản tương tự nhất với kết quả OCR đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi; và hiệu chỉnh kết quả OCR đó bằng văn bản tương tự nhất đã xác định được.

Mô tả vắn tắt các hình vẽ

Các phương án được sáng chế nêu ra chỉ nhằm lấy làm ví dụ và sẽ được hiểu sâu hơn và rõ ràng hơn đối với người có hiểu biết trung bình trong lĩnh vực này khi đọc phần mô tả chi tiết sau đây dựa trên các hình vẽ kèm theo, trong đó:

Fig.1 là lưu đồ minh họa phương pháp hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động theo một phương án của sáng chế.

Fig.2 là hình vẽ phối cảnh của hệ thống nhận dạng ký tự quang học (OCR) theo một phương án của sáng chế.

Fig.3 là lưu đồ minh họa phương pháp thực thi được bằng máy tính để hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động theo một phương án của sáng chế.

Fig.4 thể hiện sơ đồ khối của hệ thống máy tính thích hợp để sử dụng trong quá trình thực thi ít nhất một vài bước của phương pháp hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động.

Mô tả chi tiết sáng chế

Các phương án sẽ được mô tả, chỉ nhằm lấy làm ví dụ, dựa trên các hình vẽ. Các số chỉ dẫn và ký tự tương tự trên các hình vẽ chỉ các chi tiết tương tự hoặc tương đương.

Một vài đoạn trong phần mô tả dưới đây được trình bày một cách rõ ràng hoặc ngầm định theo các thuật toán và các dạng biểu diễn bằng hàm số hoặc ký hiệu về các thao tác trên dữ liệu trong bộ nhớ máy tính. Phần mô tả về thuật toán này và các dạng

biểu diễn bằng hàm số hoặc ký hiệu là phương tiện được người có hiểu biết trung bình trong lĩnh vực xử lý dữ liệu dùng để truyền tải hiệu quả nhất bản chất công việc của họ tới người có hiểu biết trung bình khác trong lĩnh vực liên quan. Thuật toán nêu trong bản mô tả này và nói chung, được hiểu là một trình tự các bước tự phù hợp giúp thu được kết quả mong muốn. Các bước này là các bước can thiệp vật lý của các đại lượng vật lý, như tín hiệu điện, tín hiệu từ, hoặc tín hiệu quang có thể được lưu trữ, truyền, kết hợp, so sánh và thao tác khác đi.

Trừ khi có quy định khác một cách cụ thể và như rõ ràng từ phần mô tả sau đây và cần hiểu rằng trong toàn bộ bản mô tả sáng chế, các câu chữ sử dụng các thuật ngữ như “tiếp nhận”, “quét”, “tính toán”, “xác định”, “thay thế”, “tạo ra”, “khởi tạo”, “kết xuất” hoặc thuật ngữ tương tự, dùng để chỉ hoạt động và các quy trình của hệ thống máy tính hoặc thiết bị điện tử tương tự, mà vận hành và biến đổi dữ liệu được biểu diễn dưới dạng các đại lượng vật lý trong hệ thống máy tính thành một dữ liệu khác cũng được biểu diễn một cách tương tự dưới dạng các đại lượng vật lý trong hệ thống máy tính hoặc các thiết bị lưu trữ, truyền phát hoặc hiển thị thông tin khác.

Sáng chế cũng bộc lộ thiết bị dùng để thực hiện các thao tác của phương pháp theo sáng chế. Thiết bị này có thể được tạo kết cấu riêng cho các mục đích được yêu cầu hoặc có thể bao gồm máy tính hoặc thiết bị khác được kích hoạt hoặc được tái tạo cấu hình một cách chọn lọc bởi chương trình máy tính lưu trữ trong máy tính. Các thuật toán và dạng hiển thị nêu trong bản mô tả này vốn không liên quan đến máy tính cụ thể bất kỳ hoặc thiết bị khác. Nhiều dạng máy khác nhau có thể được sử dụng với các chương trình theo những gì đề cập trong phần mô tả. Theo cách khác, kết cấu của thiết bị chuyên dụng hơn để thực hiện các bước của phương pháp được yêu cầu có thể là phù hợp. Kết cấu của máy tính thích hợp để thực thi các phương pháp/quy trình khác nhau nêu trong bản mô tả này sẽ được đưa ra từ phần mô tả dưới đây.

Ngoài ra, sáng chế cũng ngầm định bộc lộ chương trình máy tính, trong đó sẽ rõ ràng đối với người có hiểu biết trung bình trong lĩnh vực này rằng các bước riêng lẻ của phương pháp được mô tả trong bản mô tả có thể được thực hiện bằng mã máy tính. Chương trình máy tính không được dự định là bị giới hạn ở ngôn ngữ lập trình cụ thể bất kỳ và việc thực thi ngôn ngữ đó. Cần hiểu rằng nhiều ngôn ngữ lập trình và việc mã hóa chúng có thể được sử dụng để thực hiện nội dung trong phần mô tả. Ngoài ra, chương trình máy tính không được dự định là bị giới hạn ở tiến trình điều khiển cụ thể

bất kỳ. Có nhiều biến thể khác của chương trình máy tính đó mà có thể sử dụng các tiến trình điều khiển khác nhau mà không vượt ra khỏi ý tưởng hoặc phạm vi của sáng chế.

Hơn nữa, một hoặc nhiều bước của chương trình máy tính có thể được thực hiện song song thay vì tuần tự. Chương trình máy tính này có thể được lưu trữ trong phương tiện đọc được bằng máy tính bất kỳ. Phương tiện đọc được bằng máy tính này có thể gồm có bộ phận lưu trữ như đĩa từ hoặc đĩa quang, chip ghi nhớ hoặc các bộ phận lưu trữ khác thích hợp để giao tiếp với máy tính. Phương tiện đọc được bằng máy tính cũng có thể bao gồm phương tiện nối dây cứng như được lấy làm ví dụ trong hệ thống Internet hoặc phương tiện không dây như được lấy làm ví dụ trong hệ thống điện thoại di động GSM. Chương trình máy tính khi được tải và thực thi trên máy tính như vậy hiệu quả cho phép một thiết bị thực thi các bước của phương pháp được ưu tiên.

Các kỹ thuật nhận dạng ký tự quang học (OCR) hiện hành có thể đạt được tỷ lệ chính xác cao; tuy nhiên, đôi khi vẫn xuất hiện lỗi. Do đó, các phương án nhằm xác định các kết quả OCR cần hiệu chỉnh và cung cấp cơ chế hiệu chỉnh tự động các kết quả này.

Các tên và địa chỉ tạo nên phần chính của các kết quả eKYC OCR lần lượt hướng tới chia sẻ cùng một xu hướng (ví dụ các tên có thể chia sẻ cùng một xu hướng, và các địa chỉ có thể chia sẻ cùng một xu hướng). Ví dụ, công dân Malaysia là người Hồi giáo thường có tên là Mohammad hoặc Muhammad. Theo đó, theo các phương án, sáng chế đề cập đến phương pháp hiệu chỉnh OCR tự động dựa trên mô hình ngôn ngữ nơ-ron và khoảng cách chỉnh sửa biến đổi. Việc tạo mô hình ngôn ngữ nơ-ron được sử dụng để tạo mô hình cho các quy tắc nội bộ về tên và địa chỉ nhằm tìm ra các kết quả OCR cần hiệu chỉnh. Cụ thể, mô hình ngôn ngữ có thể, theo bối cảnh đã cho, dự đoán một từ cụ thể và xác suất tương ứng của nó; và tạo chuỗi, cho điểm số để đánh giá xem liệu đầu ra đã dự đoán có đáp ứng các quy tắc nội bộ hay không. Khoảng cách chỉnh sửa biến đổi được sử dụng để xác định câu trả lời đúng từ kho ngữ liệu tên và địa chỉ đã có.

Mô hình ngôn ngữ thống kê tạo ra sự phân bố theo xác suất trên chuỗi các từ. Theo đó, theo bối cảnh đưa ra, sự phân bố theo xác suất của một từ cụ thể có thể được dự đoán. Với trình tự như vậy, khi nói đến độ dài m thì có thể gán xác suất $P(w_1, w_2, \dots, w_m)$ cho toàn bộ chuỗi. Ví dụ, khi cho bối cảnh, “con mèo () ở trong sân”, thì có thể suy ra từ () là “chạy”, “đang ngủ” và v.v., và có thể xác định xác suất đối với mọi từ. Đồng thời, nhờ sử dụng mô hình ngôn ngữ, có thể là, câu “con mèo đang lướt trong

sân” được cho số điểm thấp, nghĩa là nó không phù hợp với quy tắc nội bộ của ngôn ngữ. Các tên và địa chỉ có thể được xem là chuỗi các ký tự. Ví dụ, các tác giả sáng chế biết “Jack” là một cái tên, và rất có khả năng sẽ có lỗi với tên “XXYYZZ” mà có thể được đánh giá bởi mô hình ngôn ngữ đã được kho ngữ liệu tên huấn luyện. Tóm lại, mô hình ngôn ngữ học xác suất xuất hiện từ dựa trên các ví dụ về văn bản từ kho ngữ liệu văn bản (ví dụ như tập văn bản lớn và được tạo cấu trúc). Khi cho một câu (chuỗi các từ) cụ thể, thì mô hình ngôn ngữ có thể dự đoán từ tiếp theo khi cho bối cảnh hoặc gán điểm số cho toàn bộ câu khi đánh giá xem liệu câu này có đáp ứng các quy tắc nội bộ của ngôn ngữ hay không.

Một ví dụ khác là, trong bối cảnh tên và địa chỉ, “câu” = “tên”, “từ” = “ký tự”, cho nên khi cho “Mari” thì có thể dự đoán ký tự tiếp theo là “một” với điểm số tin cậy cao. Khi cho “Maria”, có thể đánh giá xem liệu đó có phải là tên đúng hay không. Mặt khác, nếu ký tự được nhận dạng sai trong quy trình OCR, thì có thể sử dụng mô hình ngôn ngữ để xác định rằng ký tự đã nhận dạng không tuân thủ quy tắc của kho ngữ liệu, như tên “Maaia”.

Theo một phương án lấy là ví dụ, mô hình ngôn ngữ được huấn luyện bởi kho ngữ liệu tên và địa chỉ đã xây dựng được sử dụng để xác định xem liệu kết quả OCR có cần được hiệu chỉnh hay không. Nếu kết quả OCR cần được hiệu chỉnh, thì quy trình khoảng cách chỉnh sửa biến đổi được sử dụng để xác định độ tương tự theo trực quan của các văn bản. Quy trình khoảng cách chỉnh sửa biến đổi này sử dụng khoảng cách chỉnh sửa dưới dạng đường cơ sở, mà sử dụng các bước thao tác (thêm, xóa hoặc chuyển đổi ký tự) để đánh giá độ tương tự giữa hai chuỗi ký tự, và xem xét độ tương tự theo trực quan của các ký tự đó. Ví dụ, việc chuyển đổi từ “B” thành “R” có nhiều khả năng xảy ra hơn so với chuyển đổi từ “B” thành “Z”, vì chuyển đổi từ “B” thành “R” là tương đối tương tự về mặt trực quan hơn. Các kết quả OCR được thêm vào kho ngữ liệu để làm giàu kho ngữ liệu này và nâng cao hiệu suất cho mô hình ngôn ngữ và quy trình khoảng cách chỉnh sửa biến đổi.

Fig.1 là lưu đồ 100 minh họa phương pháp hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động theo một phương án.

Ở bước 102, dữ liệu lịch sử được thu thập từ miền đích (ví dụ, công dân từ một nước cụ thể hoặc khu vực địa lý đang quan tâm). Dữ liệu lịch sử này bao gồm các tên

và địa chỉ liên kết với công dân từ nước cụ thể hoặc khu vực địa lý đang quan tâm đó.

Ở bước 104, kho ngữ liệu (thư viện) tên và địa chỉ được xây dựng từ dữ liệu lịch sử thu thập được trong bước 102. Cụ thể, các trường tên và địa chỉ của dữ liệu lịch sử này được sử dụng để xây dựng kho ngữ liệu tên và địa chỉ.

Ở bước 106, mô hình ngôn ngữ được huấn luyện với kho ngữ liệu tên và địa chỉ đã xây dựng được trong bước 102 bằng cách sử dụng mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM). Mạng nơ-ron LSTM là kiến trúc mạng nơ-ron hồi quy (RNN) nhân tạo được sử dụng trong trường học sâu. Không giống như các mạng nơ-ron truyền thẳng tiêu chuẩn, các mạng nơ-ron LSTM có kết nối phản hồi để xử lý tốt hơn toàn bộ chuỗi dữ liệu, và do đó đặc biệt hữu dụng để huấn luyện mô hình ngôn ngữ. Đơn vị LSTM thường bao gồm tế bào, cổng vào, cổng ra và cổng quên. Tế bào ghi nhớ các giá trị trong các thời cách bất ý và ba cổng điều chỉnh luồng thông tin vào và ra khỏi tế bào.

Ở bước 108, mô hình ngôn ngữ huấn luyện được sử dụng để đánh giá các kết quả OCR để xác định điểm số đánh giá. Điểm số tương đối cao tương ứng với xác suất tương đối cao rằng từ đó đáp ứng quy tắc nội bộ (ngụ ý rằng có khả năng tương đối cao là từ đó được nhận dạng đúng). Các kết quả OCR cần được hiệu chỉnh nếu điểm số đánh giá thấp hơn ngưỡng định trước.

Ở bước 110, văn bản tương tự nhất được xác định từ kho ngữ liệu tên và địa chỉ đã xây dựng trong bước 102 bằng cách sử dụng khoảng cách chỉnh sửa biến đổi. Khoảng cách chỉnh sửa là cách định lượng mức độ không tương tự giữa hai chuỗi (ví dụ như các từ) với nhau bằng cách đếm số lượng thao tác tối thiểu cần để biến đổi một chuỗi thành chuỗi còn lại. Cụ thể, khoảng cách chỉnh sửa được sử dụng dưới dạng đường cơ sở. Từ khoảng cách chỉnh sửa, có thể chuyển đổi chuỗi A thành chuỗi B với số bước ít nhất, và dùng cho mọi thao tác của mọi bước, đây được gọi là cặp chỉnh sửa. Sau đó, độ tương tự theo trực quan được tạo mô hình cho mọi cặp chỉnh sửa. Việc tính độ tương tự theo trực quan là dựa trên mức độ thường xuyên mà người sử dụng hiệu chỉnh các kết quả OCR từ trước đó. Độ tương tự theo trực quan có thể được xác định bởi $\log(\text{tần suất cặp chỉnh sửa trong lịch sử})$. Đối với mỗi kết quả OCR cần được hiệu chỉnh, kho ngữ liệu tên và địa chỉ được truy cập và từ tương tự nhất được tìm kiếm dựa trên kết quả khoảng cách chỉnh sửa biến đổi.

Ví dụ như, đối với tên “MICHAEL”, kho ngữ liệu có “MICHAEL” và

“MICHAEM”. Tuy nhiên, MICHAEL được chọn, mặc dù cả hai có cùng khoảng cách chỉnh sửa ban đầu bằng 1, vì “L” và “I” tương tự về mặt trực quan hơn so với “T” và “M”, do đó nó được coi là tương tự hơn nhờ kỹ thuật khoảng cách chỉnh sửa biến đổi.

Ở bước 112, các kết quả OCR đã hiệu chỉnh được trả lại. Các kết quả OCR đã được nhận dạng chính xác (ví dụ như các OCR không được hiệu chỉnh) cũng được trả lại. Ở bước 114, các kết quả OCR đã được hiệu chỉnh (và, tùy ý, không được hiệu chỉnh) được thêm vào dữ liệu lịch sử để làm giàu kho ngữ liệu tên và địa chỉ.

Theo một phương án, sáng chế đề cập đến hệ thống OCR bao gồm tổ hợp của phần cứng và phần mềm được sử dụng để chuyển đổi các tài liệu vật lý thành mã/văn bản đọc được bằng máy. Phần cứng, bao gồm mô đun/bộ phận chụp ảnh/máy quét quang học và bảng mạch chuyên dụng được sử dụng để sao chép hoặc đọc văn bản trên các tài liệu vật lý. Phần mềm OCR chuyển đổi ảnh tài liệu đã quét thành mã/dữ liệu tương ứng đọc được bằng máy.

Fig.2 là hình vẽ dạng sơ đồ của hệ thống nhận dạng ký tự quang học (OCR) 200 theo một phương án. Hệ thống OCR 200 gồm có thiết bị thu thập 202, thiết bị chuyển đổi ảnh 204, thiết bị hiệu chỉnh 206, và thiết bị đầu ra 208.

Thiết bị thu thập 202, mà có thể có dạng máy quét quang học, camera, thiết bị chụp ảnh, v.v., thu được ảnh kỹ thuật số của tài liệu vật lý 201. Thiết bị chuyển đổi ảnh 204, cùng với phần mềm OCR thích hợp, chuyển đổi ít nhất một phần ảnh kỹ thuật số của tài liệu vật lý 201 thành văn bản (hoặc mã) tương ứng đọc được bằng máy.

Thiết bị hiệu chỉnh 206 đánh giá văn bản đọc được bằng máy bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh văn bản đọc được bằng máy này hay không. LSTM cũng được coi là mạng nơ-ron hồi quy hai hướng với bộ nhớ dài-ngắn hạn. Nếu cần phải hiệu chỉnh văn bản đọc được bằng máy đó, thì thiết bị hiệu chỉnh 206 sẽ xác định văn bản tương tự nhất với văn bản đọc được bằng máy đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi. Thiết bị hiệu chỉnh 206 này còn có thể đánh giá log(tần suất cặp chỉnh sửa trong lịch sử) để xác định văn bản tương tự nhất với văn bản đọc được bằng máy cần hiệu chỉnh. Mặt khác, nếu không cần phải hiệu chỉnh văn bản đọc được bằng máy, thì thiết bị hiệu chỉnh 206 chuyển văn bản đọc được bằng máy đó (không có bất kỳ biến đổi nào) vào thiết bị đầu ra 208.

Thiết bị hiệu chỉnh 206 còn có thể hiệu chỉnh văn bản đọc được bằng máy đó bằng văn bản tương tự nhất đã xác định được. Nói cách khác, văn bản đọc được bằng máy bị lỗi được thay thế bằng văn bản tương tự nhất đã xác định được đó.

Thiết bị đầu ra 208 kết xuất văn bản đọc được bằng máy đã hiệu chỉnh. Ví dụ, thiết bị đầu ra 208 này có thể kết xuất văn bản đọc được bằng máy đã hiệu chỉnh ra thiết bị hiển thị kỹ thuật số, là bộ xử lý (không được thể hiện trên hình vẽ) để xử lý tiếp hoặc ra thiết bị bên ngoài.

Hệ thống OCR 200 có thể còn bao gồm thiết bị cơ sở dữ liệu 210 có lưu trữ trong đó tập hợp tên và địa chỉ liên kết với miền đích (ví dụ như công dân từ một nước cụ thể hoặc khu vực địa lý đang quan tâm). Thiết bị cơ sở dữ liệu 210 xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ đã lưu trữ liên kết với miền đích.

Hệ thống OCR 200 có thể còn bao gồm thiết bị huấn luyện 212 được tạo cấu hình để huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ. Thiết bị huấn luyện 212 này cũng được tạo cấu hình để thực hiện các quy trình xử lý sơ bộ đối với khoảng cách chỉnh sửa biến đổi. Quy trình xử lý sơ bộ có thể bao gồm ít nhất là các bước sau: (i) tạo mô hình độ tương tự theo trực quan đối với mọi cặp chỉnh sửa và (ii) tính độ tương tự theo trực quan này dựa trên mức độ thường xuyên mà người sử dụng hiệu chỉnh các kết quả OCR của cặp chỉnh sửa từ trước đó, ví dụ, $\log(\text{tần suất cặp chỉnh sửa trong lịch sử})$.

Thiết bị đầu ra 208 có thể còn được tạo cấu hình để thêm văn bản đọc được bằng máy đã hiệu chỉnh vào tập hợp tên và địa chỉ lưu trữ trong thiết bị cơ sở dữ liệu 210. Văn bản đọc được bằng máy đã được nhận dạng chính xác cũng được thêm vào tập hợp tên và địa chỉ lưu trữ trong thiết bị cơ sở dữ liệu 210. Thiết bị cơ sở dữ liệu 210 này có thể còn được tạo cấu hình để tái dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ mà còn chứa cả văn bản đọc được bằng máy đã hiệu chỉnh. Thiết bị huấn luyện 212 có thể còn được tạo cấu hình để huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng. Thiết bị huấn luyện 212 cũng có thể còn được tạo cấu hình để cập nhật các kết quả tính được về khoảng cách chỉnh sửa biến đổi.

Fig.3 là lưu đồ minh họa phương pháp thực thi bằng máy tính 300 để hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động theo một phương án. Phương pháp 300 này

bao gồm bước 306 liên quan đến việc đánh giá kết quả OCR bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh kết quả OCR đó hay không.

Nếu cần phải hiệu chỉnh kết quả OCR, thì phương pháp 300 bao gồm bước 308 liên quan đến việc xác định văn bản tương tự nhất với kết quả OCR đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi. Kỹ thuật khoảng cách chỉnh sửa biến đổi gồm thao tác đánh giá $\log(\text{tần suất cập chỉnh sửa trong lịch sử})$ để xác định văn bản tương tự nhất với kết quả OCR.

Tiếp đến, bước 310 liên quan đến việc hiệu chỉnh kết quả OCR bằng văn bản tương tự nhất đã xác định được.

Trước các bước 306, 308 và 310, phương pháp 300 có thể còn bao gồm bước 302 nhằm xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ liên kết với miền đích (ví dụ, công dân từ một nước cụ thể hoặc khu vực địa lý đang quan tâm). Sau đó, phương pháp 300 này có thể còn bao gồm bước 304 nhằm huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã xây dựng được.

Sau bước 310, phương pháp 300 có thể còn bao gồm bước 312 nhằm thêm kết quả OCR đã hiệu chỉnh vào tập hợp tên và địa chỉ liên kết với miền đích. Các kết quả OCR đã được nhận dạng chính xác (ví dụ, các kết quả OCR không được hiệu chỉnh) cũng được thêm vào tập hợp tên và địa chỉ liên kết với miền đích này.

Sau đó, phương pháp 300 có thể còn bao gồm bước 302' nhằm tái dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ chứa kết quả OCR đã hiệu chỉnh. Nói cách khác, bước 302 này được lặp lại với tập hợp tên và địa chỉ đã cập nhật chứa kết quả OCR đã hiệu chỉnh.

Sau đó, phương pháp 300 có thể còn bao gồm bước 304' nhằm huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng. Nói cách khác, bước 304 được lặp lại với kho ngữ liệu tên và địa chỉ đã tái dựng.

Cơ chế phản hồi liên quan đến các bước 312, 302' và 304' giúp nâng cao hiệu suất hiệu chỉnh OCR tự động bằng cách bổ sung và làm giàu liên tục cho kho ngữ liệu tên và địa chỉ bằng các kết quả đã hiệu chỉnh, mà sau đó giúp cho mô hình ngôn ngữ

mạng nơ-ron LSTM được huấn luyện tốt hơn.

Fig.4 thể hiện sơ đồ khối của hệ thống máy tính thích hợp để sử dụng trong quá trình thực thi ít nhất một vài bước của phương pháp mô tả ở trên nhằm hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động.

Phân mô tả về hệ thống máy tính/thiết bị điện toán 400 dưới đây được nêu ra chỉ nhằm lấy làm ví dụ và không được dự định là sẽ bị giới hạn.

Như được thể hiện trên Fig.4, thiết bị điện toán 400 lấy là ví dụ gồm có bộ xử lý 404 để thực thi các thường trình phần mềm. Mặc dù hình vẽ chỉ thể hiện một bộ xử lý để cho rõ, thiết bị điện toán 400 cũng có thể bao gồm hệ thống nhiều bộ xử lý. Bộ xử lý 404 được kết nối với hạ tầng truyền thông 406 để truyền thông với các thành phần khác của thiết bị điện toán 400. Hạ tầng truyền thông 406 có thể bao gồm, ví dụ, bus truyền thông, thanh ngang, hoặc mạng.

Thiết bị điện toán 400 còn có bộ nhớ chính 408, chẳng hạn như bộ nhớ truy cập ngẫu nhiên (RAM) và bộ nhớ phụ 410. Bộ nhớ phụ 410 có thể bao gồm, ví dụ, ổ đĩa cứng 412 và/hoặc ổ đĩa lưu trữ tháo ra được 414, mà có thể bao gồm ổ băng từ, ổ đĩa quang, hoặc dạng ổ tương tự. Ổ đĩa lưu trữ tháo ra được 414 đọc từ và/hoặc ghi vào bộ phận lưu trữ tháo ra được 418 theo cách thông dụng. Bộ phận lưu trữ rời 418 này có thể bao gồm băng từ, đĩa quang, hoặc dạng tương tự, mà được đọc và ghi bởi ổ đĩa lưu trữ tháo ra được 414. Như sẽ được hiểu bởi người có hiểu biết trung bình trong (các) lĩnh vực liên quan, bộ phận lưu trữ tháo ra được 418 gồm có vật ghi đọc được bằng máy tính chứa các lệnh mã chương trình thực thi được bằng máy tính và/hoặc dữ liệu trong đó.

Theo một phương án thay thế, bộ nhớ phụ 410 ngoài ra hoặc theo cách khác có thể bao gồm các bộ phận tương tự khác để cho phép các chương trình máy tính hoặc các lệnh khác được tải vào trong thiết bị điện toán 400. Các bộ phận như vậy có thể bao gồm, ví dụ, bộ phận lưu trữ tháo ra được 422 và giao diện 420. Các ví dụ về bộ phận lưu trữ tháo ra được 422 và giao diện 420 bao gồm chip bộ nhớ tháo ra được (như EPROM hoặc PROM) và ổ cắm liên quan và các bộ phận lưu trữ tháo ra được 422 khác và các giao diện 420 cho phép phần mềm và dữ liệu được truyền từ bộ phận lưu trữ tháo ra được 422 tới hệ thống máy tính 400.

Thiết bị điện toán 400 cũng gồm có ít nhất một giao diện truyền thông 424. Giao diện truyền thông 424 này cho phép phần mềm và dữ liệu được truyền giữa thiết bị điện

toán 400 và các thiết bị bên ngoài qua đường truyền thông 426. Theo các phương án khác nhau, giao diện truyền thông 424 cho phép dữ liệu được truyền giữa thiết bị điện toán 400 và mạng truyền thông dữ liệu, như mạng truyền thông dữ liệu chung hoặc riêng. Giao diện truyền thông 424 có thể được sử dụng để trao đổi dữ liệu giữa các thiết bị điện toán 400 khác nhau mà các thiết bị điện toán 400 này tạo thành một phần của mạng máy tính liên kết. Các ví dụ về giao diện truyền thông 424 có thể bao gồm modem, giao diện mạng (như thẻ Ethernet), cổng truyền thông, anten có mạch kết hợp và dạng tương tự. Giao diện truyền thông 424 có thể là có dây hoặc không dây. Phần mềm và dữ liệu truyền qua giao diện truyền thông 424 có dạng tín hiệu mà có thể là các tín hiệu điện tử, tín hiệu điện từ, tín hiệu quang hoặc các tín hiệu khác có khả năng được giao diện truyền thông 424 tiếp nhận. Các tín hiệu này được đưa vào giao diện truyền thông qua đường truyền thông 426.

Tuỳ ý, thiết bị điện toán 400 còn có giao diện hiển thị 402 thực hiện các thao tác nhằm đưa ảnh đến màn hình hiển thị 430 đi kèm và giao diện audio 432 để thực hiện các thao tác nhằm phát nội dung audio qua (các) loa 434 đi kèm.

Như được sử dụng trong bản mô tả, thuật ngữ "sản phẩm chương trình máy tính" có thể phần nào dùng để chỉ bộ phận lưu trữ tháo ra được 418, bộ phận lưu trữ tháo ra được 422, đĩa cứng lắp trong ổ đĩa cứng 412, hoặc sóng mang mang phần mềm qua đường truyền thông 426 (liên kết không dây hoặc cáp) tới giao diện truyền thông 424. Phương tiện lưu trữ đọc được bằng máy tính dùng để chỉ phương tiện lưu trữ hữu hình không truyền sóng bất kỳ mà cung cấp các lệnh và/hoặc dữ liệu đã ghi cho thiết bị điện toán 400 để thực thi và/hoặc xử lý. Các ví dụ về phương tiện lưu trữ này bao gồm đĩa mềm, băng từ, CD-ROM, DVD, đĩa Blu-ray™, ổ đĩa cứng, ROM hoặc mạch tích hợp, bộ nhớ USB, đĩa quang từ, hoặc thẻ đọc được bằng máy tính như thẻ PCMCIA và dạng tương tự, bất kể các thiết bị này nằm bên trong hay bên ngoài thiết bị điện toán 400. Các ví dụ về phương tiện truyền phát đọc được bằng máy tính truyền sóng hoặc không hữu hình, mà có thể cũng tham gia vào quá trình cung cấp phần mềm, chương trình ứng dụng, lệnh và/hoặc dữ liệu cho thiết bị điện toán 400, bao gồm các kênh truyền phát radio hoặc hồng ngoại cũng như kết nối mạng với một máy tính hoặc thiết bị mạng khác và Internet hoặc Intranet gồm việc chuyển thư điện tử và thông tin được ghi trên trang web và dạng tương tự.

Các chương trình máy tính (còn được gọi là mã chương trình máy tính) được lưu

trữ trong bộ nhớ chính 408 và/hoặc bộ nhớ phụ 410. Các chương trình máy tính này cũng có thể cũng được tiếp nhận qua giao diện truyền thông 424. Các chương trình máy tính này, khi được thực thi, cho phép thiết bị điện toán 400 thực hiện một hoặc nhiều đặc điểm theo các phương án nêu trong bản mô tả. Theo các phương án khác nhau, các chương trình máy tính, khi được thực thi, cho phép bộ xử lý 404 thực hiện các đặc điểm theo các phương án mô tả ở trên. Theo đó, các chương trình máy tính này là các bộ điều khiển của hệ thống máy tính 400.

Phần mềm có thể được lưu trữ trong sản phẩm chương trình máy tính và được tải vào thiết bị điện toán 400 bằng cách sử dụng ổ đĩa lưu trữ tháo ra được 414, ổ đĩa cứng 412, hoặc giao diện 420. Theo cách khác, sản phẩm chương trình máy tính này có thể được tải xuống hệ thống máy tính 400 qua đường truyền thông 426. Phần mềm, khi được thực thi bởi bộ xử lý 404, khiến cho thiết bị điện toán 400 thực hiện các chức năng theo các phương án mô tả trong bản mô tả.

Cần hiểu rằng phương án trên Fig.4 chỉ được thể hiện nhằm lấy làm ví dụ. Do đó, theo một số phương án, một hoặc nhiều đặc điểm của thiết bị điện toán 400 có thể được lược bỏ. Ngoài ra, theo một số phương án, một hoặc nhiều đặc điểm của thiết bị điện toán 400 có thể được kết hợp với nhau. Ngoài ra, theo một số phương án, một hoặc nhiều đặc điểm của thiết bị điện toán 400 có thể được chia thành một hoặc nhiều thành phần.

Theo một phương án, sáng chế đề cập đến hệ thống hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động. Hệ thống này bao gồm thiết bị xử lý (như bộ xử lý 404) và thiết bị ghi nhớ (như bộ nhớ phụ 410) chứa mã chương trình máy tính. Thiết bị ghi nhớ và mã chương trình máy tính được tạo cấu hình để, thông qua thiết bị xử lý, cho phép hệ thống thực hiện ít nhất là thao tác đánh giá kết quả OCR bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) đã được huấn luyện để xác định xem có cần hiệu chỉnh kết quả OCR này hay không. Nếu cần phải hiệu chỉnh kết quả OCR đó, thì hệ thống này còn được cho phép xác định văn bản tương tự nhất với kết quả OCR đó từ kho ngữ liệu tên và địa chỉ bằng cách sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi.

Hệ thống này cũng còn được cho phép hiệu chỉnh kết quả OCR đó bằng văn bản tương tự nhất đã xác định được. Cụ thể, hệ thống này có thể đánh giá $\log(\text{tần suất cặp}$

chỉnh sửa trong lịch sử) để xác định văn bản tương tự nhất với kết quả OCR đó.

Hệ thống này còn có thể được cho phép: (i) xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ liên kết với miền đích; và (ii) trong pha huấn luyện, huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đó.

Quy trình xử lý sơ bộ khoảng cách chỉnh sửa biến đổi có thể được thực hiện trong pha huấn luyện. Khi thu được kết quả “tần suất cặp chỉnh sửa” từ dữ liệu lịch sử, kết quả này có thể được xác định và lưu trữ trước pha huấn luyện và được cập nhật khi kho ngữ liệu tên và địa chỉ mới đã sẵn sàng.

Để thực thi cơ chế phản hồi để nâng cao hiệu suất, hệ thống này có thể còn được cho phép: thêm kết quả OCR đã hiệu chỉnh vào tập hợp tên và địa chỉ liên kết với miền đích; (b) tái dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ chứa kết quả OCR đã hiệu chỉnh; và (c) huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng được.

Các kỹ thuật được mô tả trong phần mô tả này đem đến một hoặc nhiều hiệu quả kỹ thuật. Cụ thể, nhờ sử dụng mô hình ngôn ngữ được kho ngữ liệu tên và địa chỉ huấn luyện, có thể xác định tự động các văn bản có thể là không đúng, và bỏ qua các văn bản đúng. Các kết quả đã hiệu chỉnh bởi kho ngữ liệu thiết thực hơn, và sự gia tăng kích thước kho ngữ liệu này duy trì việc nâng cao hiệu suất hiệu chỉnh OCR tự động.

Người có hiểu biết trung bình trong lĩnh vực này sẽ hiểu rằng nhiều biến đổi và/hoặc cải biến khác có thể được thực hiện đối với sáng chế như được thể hiện theo các phương án cụ thể mà không nằm ngoài ý tưởng và phạm vi của sáng chế như được mô tả rộng rãi. Do đó, các phương án của sáng chế được xem xét theo tất cả các khía cạnh được minh họa và không bị giới hạn.

Ví dụ, các phương pháp khác có thể được áp dụng để xác định xem liệu các kết quả OCR có cần được hiệu chỉnh hay không. Ví dụ, các thuật toán phân loại văn bản có thể được sử dụng để phân loại các kết quả cần được hiệu chỉnh. Điểm số về độ tin cậy của hệ thống OCR cũng có thể được áp dụng để so sánh với ngưỡng định trước để xác định xem liệu các kết quả OCR có cần được hiệu chỉnh hay không. Hơn nữa, ngoài các khoảng cách chỉnh sửa biến đổi, các phương pháp khoảng cách khác cũng có thể được sử dụng để xác định văn bản tương tự nhất từ kho ngữ liệu tên và địa chỉ, như khoảng

cách Jaccard.

YÊU CẦU BẢO HỘ

1. Phương pháp được thực thi bằng máy tính để hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động, bao gồm các bước:

đánh giá, bởi một hay nhiều bộ xử lý, kết quả OCR bằng cách sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn (LSTM) để xác định xem có cần hiệu chỉnh kết quả OCR này hay không;

để đáp ứng với việc xác định rằng cần hiệu chỉnh kết quả OCR, thì xác định, bởi một hoặc nhiều bộ xử lý, miền đích của kết quả OCR;

tìm kiếm, bởi một hoặc nhiều bộ xử lý, dữ liệu lịch sử tương ứng với miền đích, trong đó dữ liệu lịch sử này bao gồm bao gồm các hiệu chỉnh trước đó của kho ngữ liệu tên và địa chỉ;

xác định, bởi một hoặc nhiều bộ xử lý, văn bản tương tự nhất so với kết quả OCR từ kho ngữ liệu tên và địa chỉ sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi, trong đó kỹ thuật khoảng cách chỉnh sửa biến đổi này bao gồm bước tính độ tương tự giữa văn bản tương tự nhất và kết quả OCR dựa trên tần suất của các hiệu chỉnh trước đó; và

hiệu chỉnh, bởi một hoặc nhiều bộ xử lý, kết quả OCR với văn bản tương tự nhất.

2. Phương pháp được thực thi bằng máy tính theo điểm 1, trong đó bước tính độ tương tự trực quan giữa văn bản tương tự nhất và kết quả OCR được thực hiện dựa trên bước đánh giá hàm số $\log(\text{tần suất của các cặp chỉnh sửa trong lịch sử})$, trong đó cặp chỉnh sửa là cặp của kết quả OCR và văn bản trung gian trong dữ liệu lịch sử để chuyển đổi thành văn bản tương tự nhất.

3. Phương pháp được thực thi bằng máy tính theo điểm 1, trong đó phương pháp này còn bao gồm bước:

xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ liên kết với miền đích.

4. Phương pháp được thực thi bằng máy tính theo điểm 3, trong đó phương pháp này còn bao gồm bước:

huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đó.

5. Phương pháp được thực thi bằng máy tính theo điểm 4, trong đó phương pháp này còn bao gồm bước:

thêm ít nhất kết quả OCR đã hiệu chỉnh vào tập hợp tên và địa chỉ liên kết với miền đích.

6. Phương pháp được thực thi bằng máy tính theo điểm 5, trong đó phương pháp này còn bao gồm bước:

tái dựng kho ngữ liệu tên và địa chỉ, dưới dạng tên đã tái dựng, dựa trên tập hợp tên và địa chỉ chứa kết quả OCR được hiệu chỉnh.

7. Phương pháp thực thi được bằng máy tính theo điểm 6, trong đó phương pháp này còn bao gồm bước:

huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM bằng cách sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng.

8. Vật ghi đọc được bằng máy tính, lâu dài lưu trữ một hoặc nhiều lệnh thực thi được bằng hệ thống máy tính để thực hiện một hoặc nhiều thao tác để hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động, bao gồm các bước:

đánh giá, bởi một hoặc nhiều bộ xử lý, kết quả OCR sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn để xác định xem có cần hiệu chỉnh kết quả OCR này hay không;

để đáp ứng với việc xác định rằng cần hiệu chỉnh kết quả OCR, thì xác định, bởi một hoặc nhiều bộ xử lý, miền đích của kết quả OCR này;

tìm kiếm, bởi một hoặc nhiều bộ xử lý, dữ liệu lịch sử tương ứng với miền đích, trong đó dữ liệu lịch sử bao gồm các hiệu chỉnh trước đó của kho ngữ liệu tên và địa chỉ;

xác định, bởi một hoặc nhiều bộ xử lý, văn bản tương tự nhất so với kết quả OCR từ kho ngữ liệu tên và địa chỉ sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi, trong đó kỹ thuật khoảng cách chỉnh sửa biến đổi này bao gồm bước tính độ tương tự giữa văn bản tương tự nhất và kết quả OCR dựa trên tần suất của các hiệu chỉnh trước đó; và

hiệu chỉnh, bởi một hoặc nhiều bộ xử lý, kết quả OCR với văn bản tương tự nhất.

9. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 8, trong đó bước tính độ tương tự

trực quan giữa văn bản tương tự nhất và kết quả OCR được thực hiện dựa trên bước đánh giá hàm số $\log(\text{tần suất của các cặp chỉnh sửa trong lịch sử})$, trong đó cặp chỉnh sửa là cặp của kết quả OCR và văn bản trung gian trong dữ liệu lịch sử để chuyển đổi thành văn bản tương tự nhất.

10. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 8, còn bao gồm một hoặc nhiều lệnh để:

xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ liên kết với miền đích.

11. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 10, còn bao gồm một hoặc nhiều lệnh để:

huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM sử dụng kho ngữ liệu tên và địa chỉ.

12. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 11, còn bao gồm một hoặc nhiều lệnh để:

thêm ít nhất kết quả OCR đã hiệu chỉnh vào tập hợp tên và địa chỉ liên kết với miền đích.

13. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 12, còn bao gồm một hoặc nhiều lệnh để:

tái dựng kho ngữ liệu tên và địa chỉ, dưới dạng tên được tái dựng, dựa trên tập hợp tên và địa chỉ mà chứa kết quả OCR đã hiệu chỉnh.

14. Vật ghi đọc được bằng máy tính, lâu dài theo điểm 13, còn bao gồm một hoặc nhiều lệnh để:

huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng.

15. Hệ thống được thực thi bằng máy tính, bao gồm:

một hoặc nhiều máy tính; và

một hoặc nhiều thiết bị bộ nhớ máy tính được nối với một hoặc nhiều máy tính theo kiểu có thể tương tác và có vật ghi đọc được bằng máy, lâu dài, hữu hình lưu trữ một hoặc nhiều lệnh mà, khi được thực thi bởi một hoặc nhiều máy tính, thực hiện một

hoặc nhiều thao tác để hiệu chỉnh nhận dạng ký tự quang học (OCR) tự động, bao gồm các bước:

đánh giá, bởi một hoặc nhiều bộ xử lý, kết quả OCR sử dụng mô hình ngôn ngữ mạng nơ-ron bộ nhớ dài-ngắn hạn để xác định xem có cần hiệu chỉnh kết quả OCR này hay không;

để đáp ứng với việc xác định rằng cần hiệu chỉnh với kết quả OCR, thì xác định, bởi một hoặc nhiều bộ xử lý, miền đích của kết quả OCR;

tìm kiếm, bởi một hoặc nhiều bộ xử lý, dữ liệu lịch sử tương ứng với miền đích, trong đó dữ liệu lịch sử bao gồm các hiệu chỉnh trước đó của kho ngữ liệu tên và địa chỉ;

xác định, bởi một hoặc nhiều bộ xử lý, văn bản tương tự nhất so với kết quả OCR từ kho ngữ liệu tên và địa chỉ sử dụng kỹ thuật khoảng cách chỉnh sửa biến đổi, trong đó kỹ thuật khoảng cách chỉnh sửa biến đổi này bao gồm bước tính độ tương tự giữa văn bản tương tự nhất và kết quả OCR dựa trên tần suất của các hiệu chỉnh trước đó; và

hiệu chỉnh, bởi một hoặc nhiều bộ xử lý, kết quả OCR với văn bản tương tự nhất.

16. Hệ thống được thực thi bằng máy tính theo điểm 15, trong đó bước tính độ tương tự trực quan giữa văn bản tương tự nhất và kết quả OCR được thực hiện dựa trên bước đánh giá hàm số $\log(\text{tần suất của các cặp chỉnh sửa trong lịch sử})$, trong đó cặp chỉnh sửa là cặp của kết quả OCR và văn bản trung gian trong dữ liệu lịch sử để chuyển đổi thành văn bản tương tự nhất.

17. Hệ thống được thực thi bằng máy tính theo điểm 15, còn thực hiện một hoặc nhiều thao tác bao gồm:

xây dựng kho ngữ liệu tên và địa chỉ dựa trên tập hợp tên và địa chỉ liên kết với miền đích.

18. Hệ thống được thực thi bằng máy tính theo điểm 17, còn thực hiện một hoặc nhiều thao tác bao gồm:

huấn luyện mô hình ngôn ngữ mạng nơ-ron LSTM sử dụng kho ngữ liệu tên và địa chỉ.

19. Hệ thống được thực thi bằng máy tính theo điểm 18, còn thực hiện một hoặc nhiều

thao tác bao gồm:

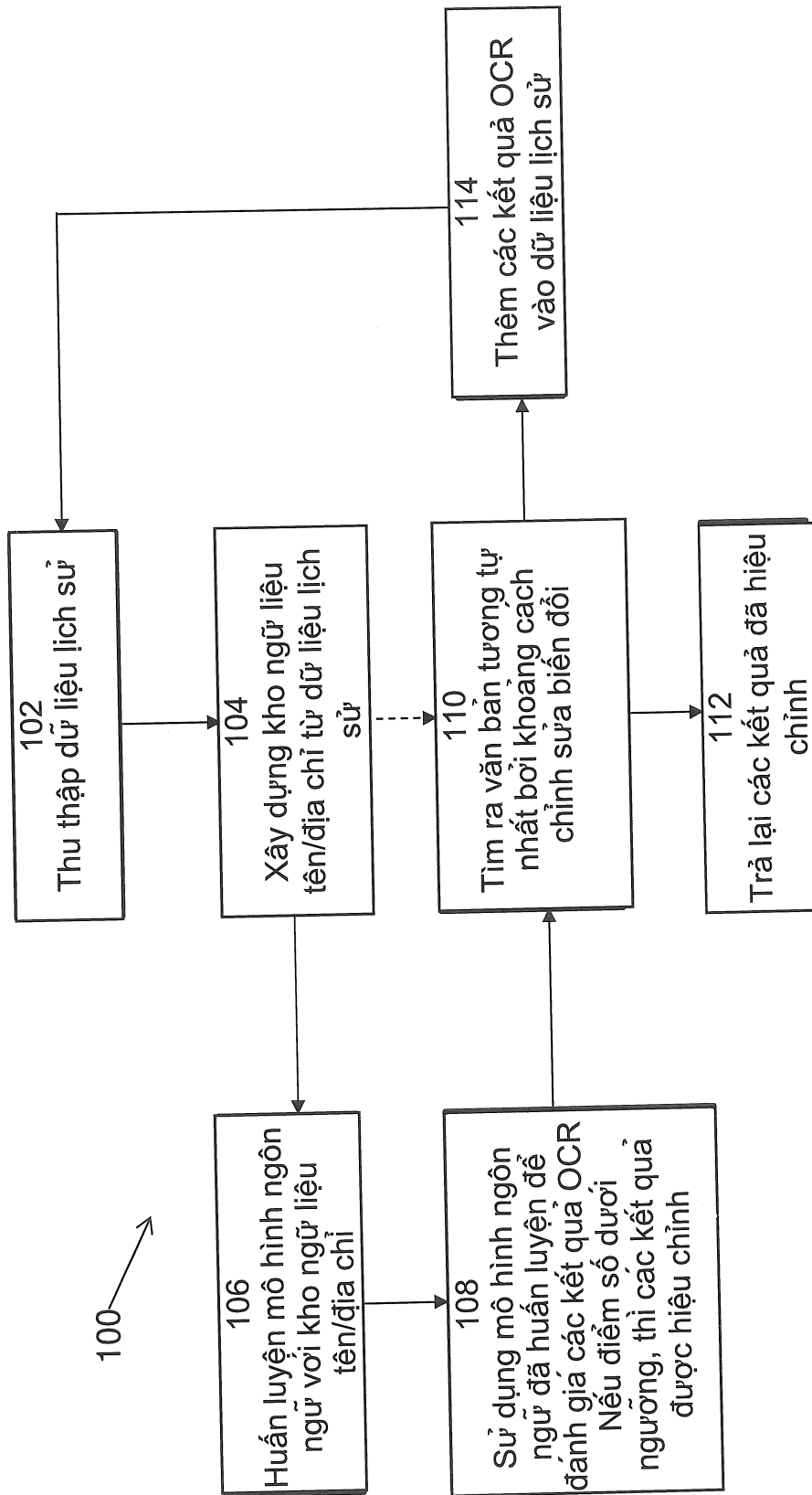
thêm ít nhất một kết quả OCR đã hiệu chỉnh vào tập hợp tên và địa chỉ liên kết với miền đích.

20. Hệ thống được thực thi bằng máy tính theo điểm 19, còn thực hiện một hoặc nhiều thao tác bao gồm:

tái dựng kho ngữ liệu tên và địa chỉ, dưới dạng tên được tái dựng, dựa trên tập hợp tên và địa chỉ mà chứa kết quả OCR đã hiệu chỉnh.

21. Hệ thống được thực thi bằng máy tính theo điểm 20, còn thực hiện một hoặc nhiều thao tác bao gồm:

huấn luyện lại mô hình ngôn ngữ mạng nơ-ron LSTM sử dụng kho ngữ liệu tên và địa chỉ đã tái dựng.

**Fig.1**

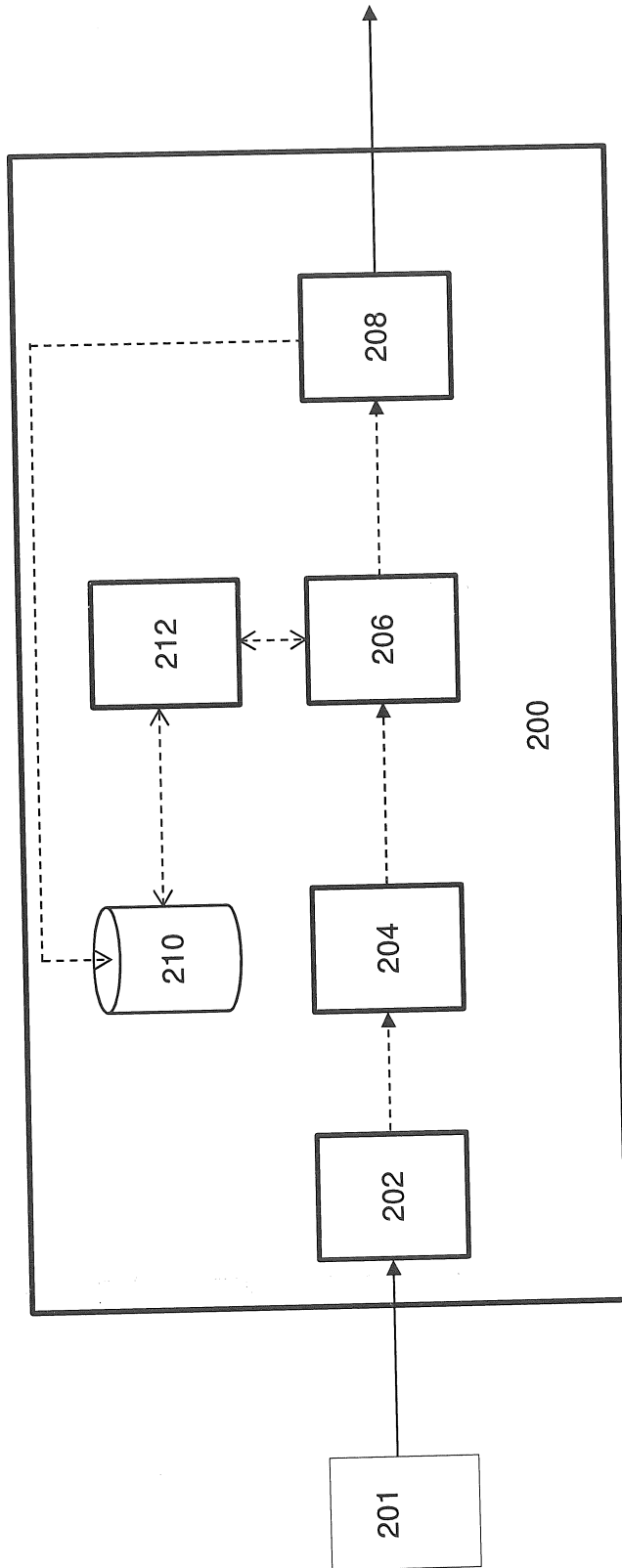


Fig.2

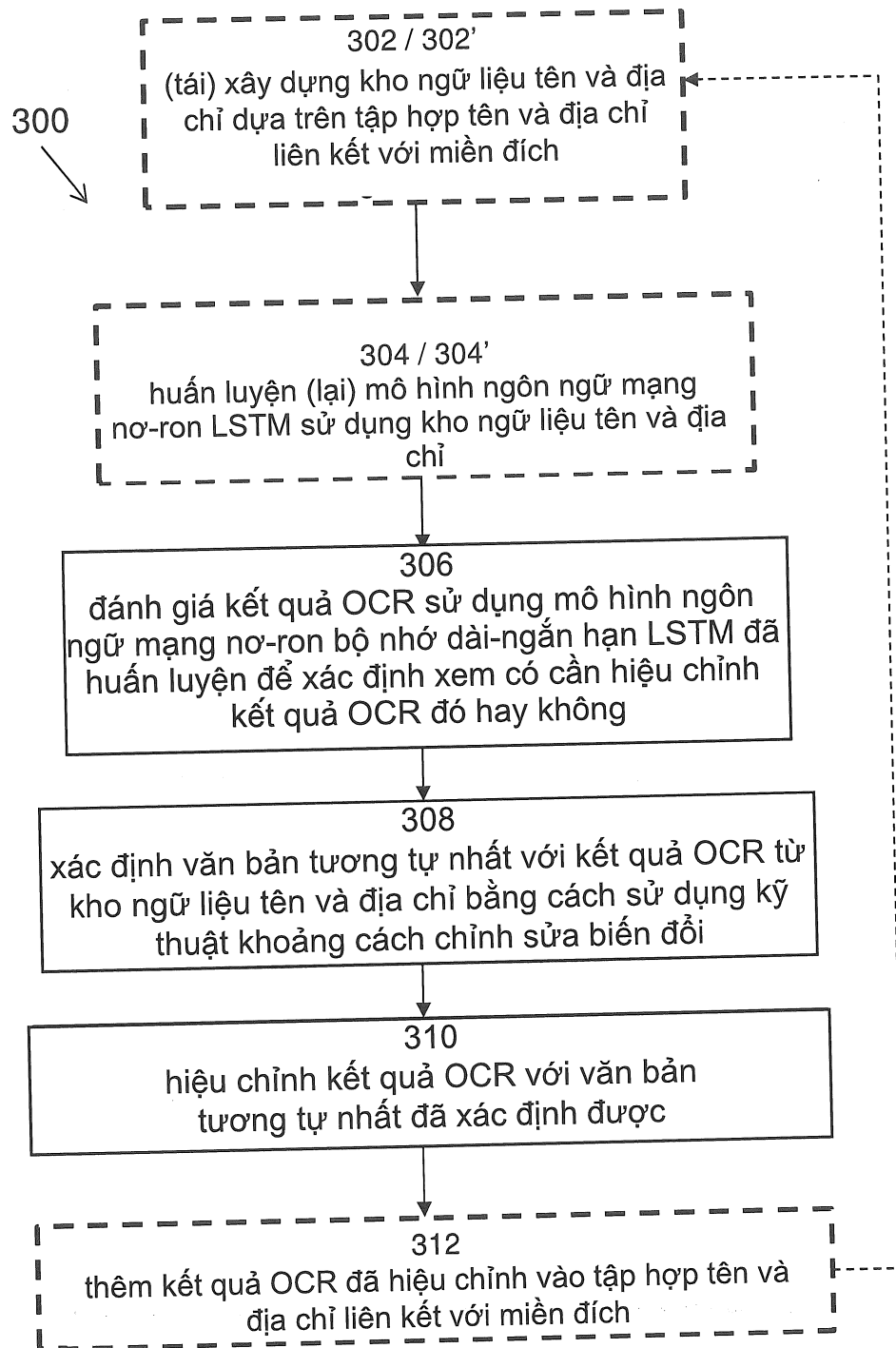


Fig.3

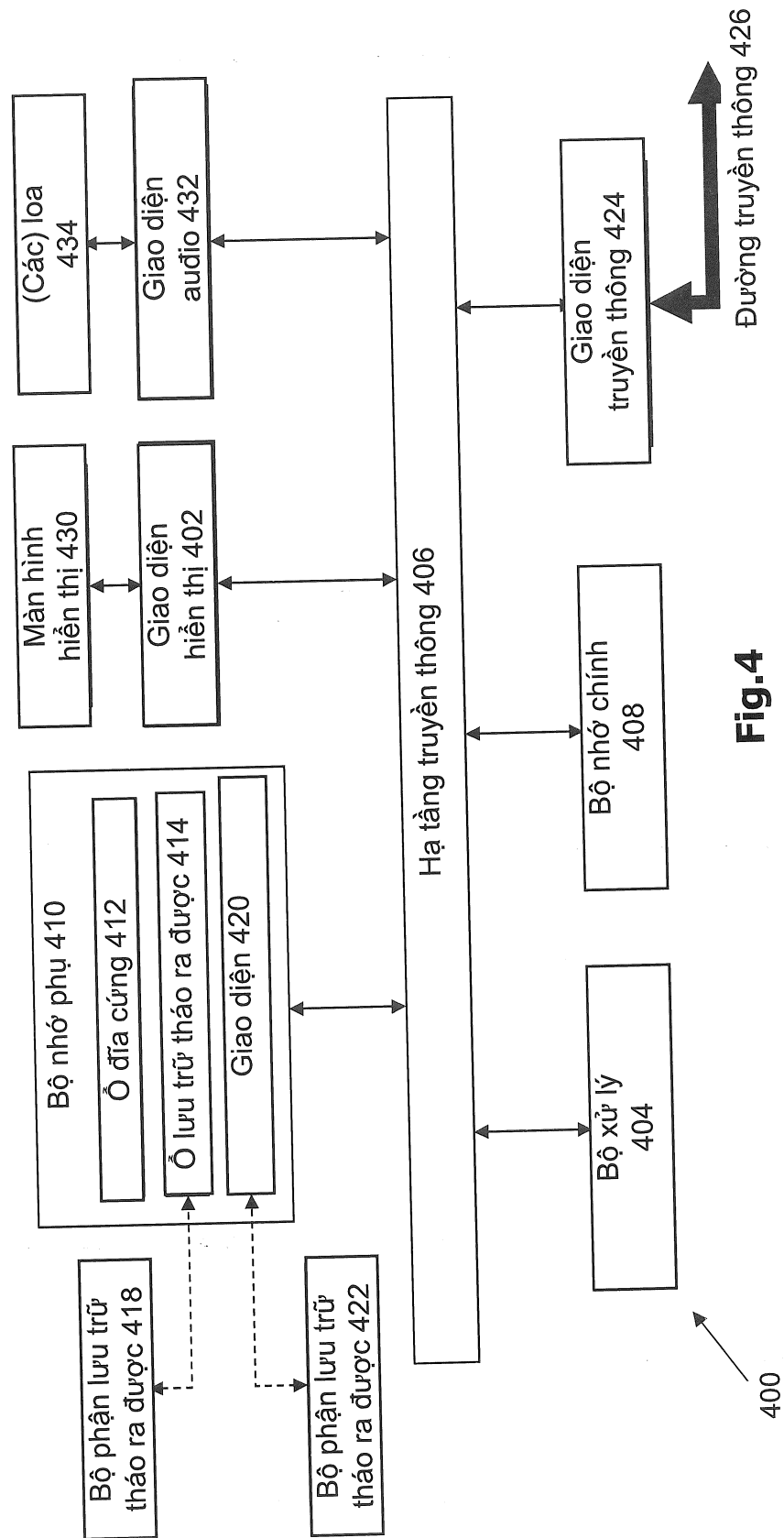


Fig.4