Report Đồ án

Duc Ho V

July 2023

MỤC LỤC

MỞ ĐẦU

CHƯƠNG 1: GIỚI THIỆU

1.1 Lý do chọn đề tài

Cuộc sống hiện nay việc mua bán trao đổi hàng hóa được diễn ra thường xuyên giữa người mua và người bán. Ban đầu hóa đơn có giá trị làm bằng chứng chứng nhận cho việc chuyển nhượng hàng hóa giữa hai bên, có giá trị làm bằng chứng chứng nhận cho việc chuyển nhượng hàng hoá giữa hai bên. Mọi việc tranh chấp trong mua bán hàng hoá hai bên tự giải quyết.

Trong quá trình phát triển xã hội, hoá đơn được phổ biến dần trong một cộng đồng khi được cộng đồng chấp nhận một cách tự nguyện. Các cộng đồng có thể là các Phường hội hoặc các định chế làng, xã. Những tranh chấp trong việc mua bán hàng hoá được các cộng đồng xử lý trên cơ sở dân sự. Khi nhà nước tham dự vào quản lý mua bán hàng hoá và xử lý những tranh chấp về hàng hoá dựa trên pháp luật dân sự và hình sự thì hoá đơn được nhà nước quy định để làm căn cứ pháp lý chứng minh cho việc chuyển nhượng hàng hoá giữa các bên và làm căn cứ để xác nhận quyền sở hữu hợp pháp của người có hàng hoá. Do đó hóa đơn là một loại tài liệu quan trọng trong các giao dịch. Nó được sử dụng để ghi lại các giao dịch mua bán hàng hóa và dịch vụ. Thông tin trên hóa đơn bao gồm tên của người bán, tên của người mua, ngày lập hóa đơn, số lượng hàng hóa hoặc dịch vụ, giá cả, tổng số tiền phải thanh toán.v.v...

Hiện nay, hóa đơn thông thường được lập dưới dạng tài liệu giấy, có thể là hóa đơn giá trị gia tăng, hóa đơn bán hàng, tem, vé, thẻ, phiếu thu tiền bảo hiểm... Hóa đơn giấy có thể được phát hành theo các hình thức như hóa đơn đặt in, hóa đơn tự in, hóa đơn mua của cơ quan thuế. Điều này gây ra một số khó khăn trong việc quản lý hóa đơn, chẳng hạn như:

- Quá nhiều hóa đơn: Các doanh nghiệp có thể phát sinh một số lượng lớn hóa đơn, từ các nhà cung cấp, khách hàng và các bên liên quan khác. Việc quản lý nhiều hóa đơn có thể là một thách thức, đặc biệt nếu chúng không được tổ chức và lưu trữ một cách hiệu quả.
- Sai sót và mất mát: Quản lý hóa đơn thủ công có thể gặp phải sai sót và mất mát hóa đơn, đặc biệt khi các hóa đơn được lưu trữ và xử lý bằng tay. Điều này có thể dẫn đến việc đòi tiền sai, không thu được tiền đúng lúc hoặc mất cơ hội thu hồi tiền nợ.
- Tìm kiếm hóa đơn: Khi cần tìm một hóa đơn cụ thể, việc tìm kiếm nó có thể là một thách thức nếu nó không được tổ chức và lưu trữ một cách hiệu quả. Điều này có thể dẫn đến chậm trễ trong quá trình thanh toán hóa đơn hoặc thậm chí mất hóa đơn.
- Lưu trữ hóa đơn. Các hóa đơn phải được lưu trữ trong một thời gian nhất định theo quy định của pháp luật. Điều này có thể là một thách thức nếu không có một quy trình lưu trữ hóa đơn hiệu quả.

- Tuân thủ luật pháp: Việc tuân thủ các quy định và luật pháp về hóa đơn là rất quan trọng. Nếu không tuân thủ đúng, doanh nghiệp có thể phải đối mặt với các vấn đề pháp lý và hậu quả tài chính nghiêm trọng.
- Thay đổi trong quy định thuế: Thay đổi trong quy định thuế và các quy tắc về hóa đơn có thể làm cho việc quản lý hóa đơn trở nên phức tạp hơn, đòi hỏi doanh nghiệp phải cập nhật và điều chỉnh quy trình của mình thường xuyên.

Để giải quyết các vấn đề này ta có thể cân nhắc sử dụng các phần mềm quản lý hóa đơn hiện đại. OCR có thể giải quyết vấn đề này bằng cách tự động hóa quá trình nhập liệu của hóa đơn một cách đơn giản và dễ dàng.

1.2 Giải pháp

OCR là một công nghệ có thể giải quyết các vấn đề trên bằng cách tự động trích xuất thông tin từ hóa đơn. Đây là một công nghệ cho phép máy tính nhận dạng và chuyển đổi văn bản từ hình ảnh chứa văn bản thành dạng văn bản có thể chỉnh sửa, tìm kiếm và lưu trữ. Áp dụng OCR trong việc quản lý hóa đơn có thể giúp giải quyết một số vấn đề như sau:

- **Tiết kiệm thời gian**: OCR giúp doanh nghiệp tiết kiệm lượng lớn thời gian so với quá trình nhập dữ liệu thủ công. Với công cụ OCR, thông tin có thể dễ dàng được trích xuất sang các định dạng kỹ thuật số theo nhu cầu chỉ bằng việc chụp và tải ảnh lên. Không chỉ vậy, dữ liệu khi được trích xuất có thể dễ dàng được tìm kiếm, chỉnh sửa và thực hiện nhiều tác vụ khác, hỗ trợ quy trình xử lý tài liệu dễ dàng và thuận tiện hơn. Trên thực tế, nghiên cứu đã phát hiện ra rằng lượng thời gian dành cho công việc giấy tờ có thể giảm 75% khi sử dụng OCR. Trung bình, thời gian để trích xuất một tài liệu sang dạng số chỉ từ 0.5 – 2 giây với công cụ OCR, một sự tối ưu đáng kể so với thời giang trung bình 1– 5 phút khi sử dụng phương pháp nhập liệu truyền thống. [1]
- **Cải thiện độ chính xác**: Việc nhập liệu bằng tay không chỉ tốn nhiều thời gian, nguồn lực mà còn có mức độ rủi ro cao trong sai sót nhập. Nhất là với các loại tài liệu bao gồm nhiều trường thông tin bằng số, địa chỉ email, địa chỉ nhà,... việc nhập tay thủ công khó có thể chính xác 100%. Những lỗi sai thông tin ngay từ bước đầu sẽ khiến kho dữ liệu doanh nghiệp không được "sạch" và chính xác.
- **Hỗ trợ tuân thủ luật pháp**: Sử dụng OCR giúp đảm bảo tính chính xác và toàn vẹn của dữ liệu trên hóa đơn, từ đó đảm bảo tuân thủ các quy định về hóa đơn và thuế.
- **Quản lý hóa đơn điện tử**: Kết hợp OCR với hóa đơn điện tử giúp tự động tạo và lưu trữ các hóa đơn điện tử, giảm thiểu việc sử dụng giấy tờ truyền thống và tiết kiệm không gian lưu trữ.

Nhìn chung, OCR là một công nghệ có nhiều tiềm năng ứng dụng trong lĩnh vực kế toán và tài chính. OCR có thể giúp các doanh nghiệp tiết kiệm thời gian, tăng cường độ chính xác và cải thiện khả năng truy xuất thông tin hóa đơn. Tuy nhiên, để OCR hiệu quả, ta cần đảm bảo rằng hóa đơn được quét và lưu trữ ở định dạng tốt, đủ để đảm bảo hiệu suất nhận dạng của OCR cao nhất.

1.3 Mục Tiêu

Dựa vào những vấn đề của hóa đơn và các giải pháp của OCR ở mục 1.1 và 1.2, mục tiêu của đề tài "**Nghiên cứu ứng dụng công nghệ OCR nhận dạng hóa đơn**" là tìm hiểu, đánh giá khả năng ứng dụng của công nghệ OCR hiện nay trong việc quản lý hóa đơn. Cụ thể đề tài tập trung vào các mục tiêu sau:

- Tìm hiểu về công nghệ OCR: Nghiên cứu các nguyên lý hoạt động của OCR, các phương pháp và thuật toán phổ biến trong việc nhận dạng văn bản từ hình ảnh.
- Phân tích hiệu quả và lợi ích của ứng dụng OCR trong quản lý hóa đơn: So sánh các phương pháp truyền thống và ứng dụng OCR trong việc quản lý hóa đơn, đánh giá hiệu quả và lợi ích mà OCR mang lại, bao gồm tối ưu hóa thời gian, giảm thiểu sai sót, tiết kiệm chi phí và tăng cường khả năng xử lý lượng hóa đơn lớn.
- Đề xuất giải pháp và quy trình triển khai OCR: Dựa trên kết quả nghiên cứu, đề xuất các giải pháp và quy trình triển khai OCR trong việc quản lý hóa đơn, bao gồm lựa chọn phần mềm OCR phù hợp, quy trình xử lý hóa đơn, quản lý dữ liệu và bảo đảm tính an toàn thông tin.
- Đánh giá hiệu quả thực tế: Tiến hành thử nghiệm ứng dụng OCR trong môi trường thực tế của doanh nghiệp hoặc tổ chức để đánh giá hiệu quả, tính ổn định và khả năng mở rộng của giải pháp OCR.

Dựa trên kết quả đánh giá, đề xuất các cải tiến và phát triển tương lai của công nghệ OCR trong việc quản lý hóa đơn, nhằm nâng cao hiệu quả và khả năng ứng dụng của nó trong thực tế

1.4 Phạm vi đề tài

1.5 Bố cục chương

CHƯƠNG 2: CÁC CÔNG TRÌNH LIÊN QUAN

2.1 Cơ sở lý thuyết

Trong môi trường kinh doanh ngày nay, việc quản lý và xử lý thông tin từ các tài liệu văn bản, như hóa đơn, là một phần quan trọng đối với sự hiệu quả và tính minh bạch của hoạt động doanh nghiệp. Tuy nhiên, việc thực hiện thủ công nhận dạng và nhập liệu từ các hóa đơn có thể gây tốn thời gian, công sức và dễ dẫn đến sai sót. Để giải quyết vấn đề này, công nghệ OCR (Optical Character Recognition) đã trở thành một công cụ mạnh mẽ, hứa hẹn mang lại sự tự động hóa và cải thiện đáng kể quá trình xử lý thông tin.

Chương này tập trung vào việc trình bày những nguyên tắc cơ bản và cơ sở lý thuyết liên quan đến ứng dụng công nghệ OCR trong việc nhận dạng hóa đơn. Đi sâu vào hiểu biết về cách công nghệ OCR hoạt động, các bước tiền xử lý hình ảnh cần thiết để tối ưu hóa quá trình nhận dạng, và cách mạng nơ-ron học sâu đã thúc đẩy sự phát triển của công nghệ OCR.

2.1.1 OCR là gì?

Nhận dạng ký tự quang học đây là quá trình chuyển đổi một hình ảnh văn bản viết tay, đánh máy hoặc in thành định dạng văn bản mà máy có thể hiểu được. Nó được sử dụng rộng rãi để nhận dạng và tìm kiếm văn bản từ các tài liệu điện tử hoặc để xuất bản văn bản trên một trang web. [2], [3]

OCR được sử dụng rộng rãi như một hình thức nhập dữ liệu từ các bản ghi dữ liệu giấy in - cho dù đó là tài liệu hộ chiếu, hóa đơn, sao kê ngân hàng, biên lai vi tính hóa, danh thiếp, thư, dữ liệu in hoặc bất kỳ tài liệu phù hợp nào - đó là một phương pháp phổ biến để số hóa các văn bản in sao cho chúng có thể được chỉnh sửa, tìm kiếm, lưu trữ bằng điện tử, hiển thị trực tuyến và được sử dụng trong các quy trình máy như điện toán nhận thức, dịch máy, chuyển văn bản thành giọng nói (trích xuất), dữ liệu chính và khai thác văn bản. OCR là một lĩnh vực nghiên cứu về nhận dạng mẫu, trí tuệ nhân tạo và thị giác máy tính.[4]

Nhận dạng ký tự quang học đã được áp dụng vào nhiều ứng dụng khác nhau. Dưới đây là một số ứng dụng của OCR: [3]

- **Nhận dạng chữ viết tay:** Máy tính để nhận và diễn dịch thông tin viết tay rõ ràng từ các nguồn như tài liệu giấy, ảnh, màn hình cảm ứng và các thiết bị khác. Hình ảnh văn bản viết có thể được cảm nhận "ngoại tuyến" từ tờ giấy thông qua quét quang học hoặc nhận dạng từ thông minh. Một cách khác, các chuyển động của đầu bút viết có thể được cảm nhận "trực tuyến", ví dụ như bề mặt màn hình máy tính dựa trên bút viết.
- **Ngân hàng:** Được sử dụng để xử lý séc mà không cần sự tham gia của con người. Một tờ séc có thể được đặt vào máy, trong đó hệ thống quét số tiền cần phát hành và số tiền chính xác sẽ được chuyển khoản. Công

nghệ này đã gần như được hoàn thiện cho các séc được in ấn và cũng khá chính xác đối với các séc viết tay, giảm thiểu thời gian chờ đợi tại ngân hàng.

- **Chăm sóc sức khỏe:** Các chuyên gia y tế luôn phải đối mặt với số lượng lớn các biểu mẫu cho mỗi bệnh nhân, bao gồm cả biểu mẫu bảo hiểm cũng như các biểu mẫu sức khỏe chung. Để theo kịp với tất cả thông tin này, việc nhập dữ liệu liên quan vào một cơ sở dữ liệu điện tử có thể được truy cập khi cần thiết. Các công cụ xử lý biểu mẫu, được cung cấp bởi công nghệ OCR, có khả năng trích xuất thông tin từ các biểu mẫu và đưa vào cơ sở dữ liệu, để mỗi dữ liệu bệnh nhân được ghi lại đúng thời điểm.
- **Captcha:** Trong CAPTCHA, một hình ảnh gồm các ký tự hoặc số được tạo ra, bị mờ đi bằng các kỹ thuật biến dạng hình ảnh, biến đổi kích thước và phông chữ, phông nền gây xao lãng, đoạn ngẫu nhiên, đánh dấu và nhiễu trong hình ảnh. Hệ thống này có thể được sử dụng để loại bỏ nhiễu và phân đoạn hình ảnh để làm cho hình ảnh dễ xử lý cho các hệ thống OCR
- **Ảnh hóa đơn:** Được sử dụng rộng rãi trong nhiều ứng dụng kinh doanh để theo dõi hồ sơ tài chính và ngăn chặn việc tích lũy các khoản thanh toán chồng chất.
- **Nhận dạng biển số xe**: Sử dụng để tự động nhận dạng và ghi nhận biển số xe trên các hình ảnh hoặc video.
- ...

Từ những ứng dụng trên ta có thể thấy rằng OCR đang được sử dụng rộng rãi trong cuộc sống hàng ngày, nó đang đóng vai trò quan trọng trong việc chuyển đổi số hiện nay. Điều này rất quan trọng để tối ưu hóa quá trình làm việc với thông tin trong thời đại công nghệ thông tin.

2.1.2 Lịch sử của OCR

OCR được ra đời và cuối thế kỉ 19, được cấp bằng sáng chế tại Mỹ vào ngày 31 tháng 12 năm 1935 của Gustav Tauschek đến từ Viên, Áo, đây là một trong những phát minh sớm nhất liên quan đến OCR. OCR ban đầu được sử dụng để số hóa các văn bản in và cho phép chúng có thể đọc được bằng máy. Khi công nghệ OCR tiếp tục phát triển, nó đã được sử dụng rộng rãi trong các ngành công nghiệp khác nhau.

Sự khởi đầu thực sự của những hệ thống OCR ban đầu thực sự bắt đầu vào những năm 1960 và 1970. Các hệ thống này được thiết kế cho các trường hợp sử dụng cụ thể, chẳng hạn như phân loại thư dựa trên mã zip hoặc đọc số viết tay. Phông chữ có thể đọc bằng máy quang học đầu tiên OCR-A được phát triển vào năm 1968 bởi nhà thiết kế kiểu chữ người Thụy Sĩ Adrian Frutiger.

Trong suốt những năm 1980, công nghệ OCR đã đạt được những bước tiến đáng kể với sự phát triển của các thuật toán mới và các máy tính mạnh hơn. Các hệ thống OCR có thể nhận dạng nhiều loại phông chữ hơn và có thể xử lý các hình ảnh phức tạp hơn, khiến chúng trở nên chính xác và hữu ích hơn cho nhiều ứng dụng hơn.

Vào những năm 1990, việc sử dụng rộng rãi máy tính cá nhân và internet đã dẫn đến sự gia tăng đáng kể trong việc sử dụng công nghệ OCR. Các hệ thống OCR được sử dụng để số hóa sách, tạp chí và các tài liệu in khác, giúp tìm kiếm và truy cập thông tin dễ dàng hơn. Công nghệ này cũng được sử dụng để tự động hóa các quy trình nhập dữ liệu trong các ngành như tài chính, chăm sóc sức khỏe và chính phủ.

Vào đầu những năm 2000, lịch sử của công nghệ OCR đã phát triển với việc giới thiệu các thuật toán mới và phần cứng được cải tiến. Các hệ thống OCR trở nên chính xác hơn và có thể nhận dạng nhiều loại ký tự và ngôn ngữ hơn. Điều này đã mở đường cho việc áp dụng rộng rãi công nghệ OCR trong nhiều ngành và ứng dụng khác nhau, chẳng hạn như quản lý tài liệu và xử lý hóa đơn. Trong khung thời gian này, Google cũng nổi tiếng (và gây tranh cãi) đã ra mắt Google Sách, có tên mã là Dự án Đại dương, sử dụng OCR để số hóa hàng chục triệu cuốn sách và làm cho văn bản của chúng có thể tìm kiếm được.

Ngày nay, công nghệ OCR tiên tiến và phức tạp hơn bao giờ hết. Các hệ thống OCR có thể nhận dạng nhiều loại ký tự và ngôn ngữ, chữ viết tay và các hình ảnh phức tạp khác. Công nghệ OCR đang tiếp tục phát triển và những tiến bộ mới nhất về trí tuệ nhân tạo và máy học đang dẫn đến các hệ thống thậm chí còn phức tạp và chính xác hơn.

Lịch sử OCR bắt đầu với những phát minh mang tính cách mạng được thiết kế để cải thiện chất lượng cuộc sống cho nhân loại. Nhiều thập kỷ sau, công nghệ này vẫn đang trải qua quá trình phát triển và cải tiến liên tục, đồng thời là một yếu tố quyết định quan trọng của thời đại kỹ thuật số. OCR đã trải qua một chặng đường dài và đang thực sự cải thiện chất lượng cuộc sống của phần lớn nhân loại. Ngày nay, nhiều ngành công nghiệp và ứng dụng sử dụng OCR. Trong những thập kỷ tới, nó sẽ đóng một vai trò quan trọng trong quá trình chuyển đổi kỹ thuật số toàn cầu.[5]

2.1.3 Mạng Nơ-ron học sâu và ứng dụng trong OCR

Cùng với sự phát triển ngày càng rộng rãi và phổ biến của phần cứng máy tính cũng như sự bùng nổ của các phương pháp Học sâu. Qua đó, đã tác động mạnh mẽ đến OCR. Trước khi Học sâu trở thành một phương pháp phổ biến, OCR sử dụng các phương pháp truyền thống dựa trên việc xây dựng các quy tắc phức tạp và thủ công để nhận dạng các ký tự.

2.2 Nguyên tắc hoạt động của OCR

2.2.1 Phát hiện văn bản

2.2.1.1 DBNet

2.2.1.2 SAST

2.2.2 Nhận dạng văn bản

2.2.3 Trích xuất thông tin chính

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

REFERENCES

[1] FPT.AI. "Nâng cấp quy trình số hóa tài liệu của doanh nghiệp với công nghệ ocr." (2020), [Online]. Available: https://fpt.ai/vi/nang-cap-quy-trinh-so-hoa-tai-lieu-cua-doanh-nghiep-voi-cong-nghe-ocr (visited on 07/08/2023).

[2] AWS. "Ocr (nhận dạng ký tự quang học) là gì?" (), [Online]. Available: https://aws.amazon.com/vi/what-is/ocr/ (visited on 07/08/2023).

[3] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of ocr applications," *International Journal of Machine Learning and Computing (IJMLC)*, Jan. 2012. DOI: 10.7763/IJMLC.2012.V2.137.

[4] Wikipedia. "Optical character recognition." (2002), [Online]. Available: https://en.wikipedia.org/wiki/Optical_character_recognition (visited on 07/08/2023).

[5] D. V. Everen. "The history of ocr." (2023), [Online]. Available: https://www.veryfi.com/ocr-api-platform/history-of-ocr/ (visited on 07/08/2023).