UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

JOÃO PAULO ABDALA BOHACZK

RECONHECIMENTO ÓPTICO DE CARACTERES PARA LEITURA E TRADUÇÃO DE DOCUMENTOS EM FORMATO PDF

GUARAPUAVA

2025

JOÃO PAULO ABDALA BOHACZK

RECONHECIMENTO ÓPTICO DE CARACTERES PARA LEITURA E TRADUÇÃO DE DOCUMENTOS EM FORMATO PDF

Optical Character Recognition for Reading and Translating PDF Documents

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Kelly Lais Wiggers

GUARAPUAVA

2025

RESUMO

Diante da inacessibilidade de determinados livros e outros textos publicados antes da era digital, bem como do tratamento secundário da comunidade de OCR para com o tratamento dos documentos analisados em texto digital próprio para a leitura casual, determina-se a necessidade de uma ferramenta que, trabalhando com OCR, dedique-se na extração de textos de imagens ou arquivos PDF, visando facilitar a leitura e disseminação desses textos. Para atingir esse objetivo será realizada uma pesquisa comparativa entre ferramentas OCR. Será construída uma API em Python para realizar o trabalho de processamento do OCR e a formatação do texto. Isso será integrado por uma interface web desenvolvida com o framework PHP Laravel, com o objetivo de disponibilizar o projeto para a maior quantidade de pessoas possível. O presente trabalho possui tanto uma via científica e experimental como prática, buscando comparar as ferramentas OCR e solucionar algo que é até então ignorado pelas ferramentas, a devida formatação do texto para leitura, também tendo como objetivo disponibilizar os resultados dessa pesquisa como uma ferramenta de fácil uso e acesso para a comunidade.

Palavras-chave: ocr; pdf; revistas.

ABSTRACT

Given the inaccessibility of certain books and other texts published before the digital era, as well as the secondary treatment by the OCR community toward the processing of analyzed documents into digital text suitable for casual reading, there is a need for a tool that, working with OCR, focuses on extracting text from images or PDF files, aiming to facilitate the reading and dissemination of these texts. To achieve this objective, comparative research between OCR tools will be conducted. A Python API will be built to perform the OCR processing work and text formatting. This will be integrated through a web interface developed with the PHP Laravel framework, with the goal of making the project available to as many people as possible. The present work has both a scientific and experimental aspect as well as a practical one, seeking to compare OCR tools and solve something that has been ignored by this tools until now: the proper formatting of text for reading, also aiming to make the results of this research available as an easy-to-use and accessible tool for the community.

Keywords: ocr; pdf; magazines.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais

A escrita é uma grande invenção da humanidade, sendo ela primariamente a representação dos fonemas sonoros, isto é, representação da fala através de símbolos (SCHMANDT-BESSERAT; ERARD, 2009, p. 7). É o salto tecnológico que permitiu que a humanidade se desenvolvesse intelectualmente através do compartilhamento de informações. Durante longo tempo a escrita foi empregada em diferentes materiais e superfícies com o objetivo de transmitir informações, passando por pedras e papiros até finalmente chegar ao papel que é conhecido atualmente. (GABRIAL, 2009)

Grande parte dos escritos da humanidade estão presos em mídia física, como livros e revistas, o que dificulta a sua transmissão por meios digitais. Uma das formas de compartilhamento mais eficientes é através de fotografias que podem ser juntadas em, por exemplo, um arquivo PDF. Isso preserva a sua estrutura original, mas impede a análise de dados, pesquisa e capacidade de compartilhamento que poderia ser obtida se o texto fosse devidamente convertido em formato digital.

A humanidade conta com séculos de história escrita inacessíveis, ou com acesso dificultado e inapropriado, para o formato digital. É de grande interesse obter esses textos digitalizados propriamente para o seu compartilhamento, de forma que se preserve ao máximo a sua estrutura original, facilitando a sua leitura e disseminação nos meios digitais.

Algumas tecnologias atuais se destacam na extração de textos em imagens para o formato digital, como por exemplo, Optical Character Recognition (OCR), que é muito popular nos dias de hoje, estando presente na maioria dos smartphones na função da câmera ou na galeria de fotos, fazendo com que seja fácil extrair textos de fotos. Pesquisas recentes tem utilizado OCR para análise de documentos históricos (BOENIG *et al.*, 2019), análise de jornais antigos (DROBAC, 2019) e extração de texto de fotografias de forma geral (LI *et al.*, 2022, p. 9).

Uma das dificuldades encontradas na extração de texto via ferramentas de OCR é a falta de formatação dos textos extraídos. O objetivo do OCR é a extração do texto puro para processamento computacional, contudo, a sua reconstrução para a leitura humana é algo secundário e que muitas vezes deixa a desejar. Outro problema é que essas ferramentas tem de ser treinadas usando Ground Truth (GT) para se tornarem mais eficientes (BOENIG *et al.*, 2019).

Visto que o acesso a essas ferramentas e sua otimização exige um olhar especializado e dedicado, esse trabalho busca solucionar esse problema, focando na extração e compartilhamento de textos. Além disso, visa obter uma solução genérica o suficiente para ser usada em diferentes situações.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo deste trabalho é extrair e transcrever textos completos provenientes de arquivos, os quais podem ser imagens ou arquivos em formato PDF, mediante técnicas de processamento de imagens e OCR.

1.2.2 Objetivos específicos

- Definir a base de dados para realização dos experimentos.
- Efetuar pré-processamento na base de dados, como realce, redimensionamento, recortes, dentre outros.
- Implementar experimentos comparativos com as tecnologias de OCR disponíveis.
- Construir uma API para permitir a inserção e formatação de arquivos
- Criar um protocolo experimental robusto para validação dos resultados.

1.3 Justificativa

Boenig *et al.* (2019), ao comentar a digitalização de textos por bibliotecas, mostra a distinção entre "digitalização de imagem"(*image digitalization*) e o processamento dessas imagens por OCR. Os avanços na tecnologia de OCR permitiram que a maioria desses textos fossem processados, resultando em *dirty full text*, que tem validade de processamento para alguns casos científicos, mas que pecam em sua precisão e falta de estrutura.

Isso faz com que a conversão de uma página de, por exemplo, uma revista ou um livro, seja uma fotografia ou um arquivo PDF composto por essas fotografias, seja frustrante, caso o objetivo seja a leitura por humanos, devido as imprecisões no resultado final. Por vezes o texto digitalizado final acaba sendo ilegível.

Portanto, ao avaliar-se que a tecnologia de OCR está mais direcionada para a análise de dados do que necessariamente para o consumo humano desses dados, o desenvolvimento de tecnologias que auxiliem na leitura e transcrição do conteúdo para versões em formato digital pode ser visto como complemento e optimização da tecnologia, focando no consumo humano desses textos.

Como contribuição tem-se o desenvolvimento de uma ferramenta que visa facilitar a formatação de textos digitalizados com OCR, preservando sua estrutura original, o que torna sua leitura mais agradável e aumenta seu potencial de difusão por meios digitais, como redes sociais e sites dedicados a disseminação cultural.

A principal motivação para o desenvolvimento deste projeto é a dificuldade que se encontra no compartilhamento, anotação e localização de textos que tiveram seu principal período de publicação anterior à popularização da internet. E, consequentemente, esses textos podem não ter sua republicação disponível em formato digital.

Isso é um caso comum em, por exemplo: cursos de faculdades da área de humanas, em que há um desinteresse por parte das editoras na republicação de determinados livros que ainda são muito lidos internamente na academia, fazendo com que um dos parcos meios de acesso desse material seja por escaneamento (VARELLA; FERREIRA, 2012)

O processamento desses textos é custoso e tecnicamente desafiador, mas os avanços recentes de reconhecimento textual utilizando redes neurais oferece maior precisão e eficácia para a área. Isso possibilita visualizar o avanço do uso de redes neurais na tecnologia de OCR, bem como suas possíveis limitações e avaliar sua eficiência. Existem várias ferramentas de OCR disponíveis no mercado que vão desde o nível industrial (PaddleOCR) até o nível acadêmico (OCR-D), o que faz disso um bom momento para testes e comparações entre elas.

Foram selecionadas cinco ferramentas de OCR para testes. PaddleOCR (PaddlePaddle Community, 2020) é uma ferramenta chinesa que tem foco em nível industrial de escalonamento; OCR-D (OCR-D Project, 2020) faz parte de um programa alemão de preservação histórica, tendo foco em documentos históricos dos séculos XVI ao XVIII; Docling (Docling Team, 2024) tem como foco a comunidade de IA, especializando-se em extrair texto em arquivos *json* e *markdown*; Calimari-OCR (WICK; REUL; PUPPE, 2020) tem foco tanto em fontes históricas como modernas e busca ser mais simples e modular; Tesseract (TEAM, 2024) é a ferramenta mais tradicional da comunidade, contando com boa documentação e vastos exemplos de implementação devido a sua popularidade.

Como resultado do projeto desenvolvido será possível disponibilizar uma ferramenta que digitalize texto a partir de imagens ou arquivos PDF facilmente, bem como contribuir para uma pesquisa comparativa de ferramentas OCR disponíveis no mercado.

2 CONTEXTUALIZAÇÃO

Após a *Prensa de tipos móveis de Gutenberg*, que permitiu a impressão em massa de livros (GABRIAL, 2009) a internet se tornou o maior meio de disseminação de escritos, principalmente por textos em formato digital. Contudo, um texto digital não possui presença física, diferenciando a experiência de se ler algo reproduzido no papel. Além disso, a sua durabilidade em relação ao livro impresso também é questionável, visto que os HDs possuem durabilidade limitada, enquanto existem exemplares de livros com séculos de idade (GABRIAL, 2009, p.30).

Um texto no formato digital possui capacidades de compartilhamento que permite muitas vantagens em seu uso. Além disso, a capacidade de anotações e de análise textual utilizando do suporte digital é facilitada por ferramentas de pesquisa e análise de dados. Isso permite verificar, com facilidade, a decorrência das palavras mais usadas por um autor ao longo de sua obra, ou então, de reencontrar aquele trecho do texto que só se lembra de memória e esqueceu-se de anotar.

Existem diversas ferramentas que permitem realizar análise textual e extrair informações do material. Essas ferramentas utilizam diferentes técnicas para tal objetivo, como a técnica de Deep Neural Network (DNN) ou o OCR (KUMAR; TANWAR; TIWARI, 2025). O DNN baseia-se em uma rede neural e um aprendizado por padrões, necessitando de grande volume de treinamento. Já o OCR é uma tecnologia que permite reconhecer e converter texto impresso ou manuscrito em formato digital. As duas ferramentas podem trabalhar em conjunto, utilizando o DNN como parte do pós-processamento do OCR. Para o presente trabalho serão explorados os algoritmos de OCR.

Apesar de a tecnologia de OCR ser encontrada em vários meios, como por exemplo para detecção de informações em documentos, não foi encontrada uma forma ou ferramenta de acesso fácil para a digitalização de livros. Além disso, foram observados alguns experimentos realizados para a digitalização de páginas de uma revista, e que resultaram em baixo desempenho. Outras limitações do uso do OCR são relacionadas à (PACKER, 2011):

- Precisão: problemas ao processar conteúdo manuscrito, layouts complexos, texto distorcido ou rotacionado, tamanhos de fonte muito pequenos e imagens de baixa qualidade ou borradas.
- Fonte e idioma: funciona bem com fontes bem definidas, por exemplo, Arial ou Times New Roman. E tem limitações com alfabetos não latinos.
- Formatação e layout: ocorre falha ao preservar a formatação original do documento
- Dependência da qualidade da imagem: tem baixo desempenho com iluminação inadequada, baixa resolução ou mesmo distorções.

Como pesquisa prévia foram realizados testes com duas ferramentas de OCR, Docling e PaddleOCR, para verificar qual delas melhor se adapta ao nosso objetivo.

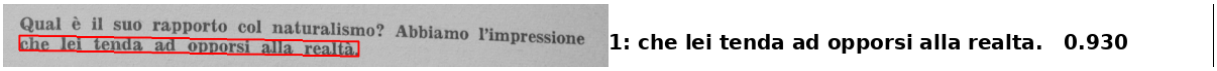

Figura 1 – Parágrafo de um texto juntamente com o seu processamento feito pela ferramenta PaddleOCR

Fonte: O autor (2025)

O Docling (Docling Team, 2024) é uma ferramenta focada em tratamento de textos para consumo de IA, contando com opções de output para, por exemplo: *markdown* e *json*. Já o PaddleOCR (PaddlePaddle Community, 2020) é mais abrangente, de uso industrial, além de textos puros pode ser utilizado em fotografias feitas ao ar livre para identificar texto. Por ser uma ferramenta de origem chinesa presumimos que seu reconhecimento de caracteres seja mais eficiente, visto que os ideogramas chineses tem maior complexidade visual do que os do alfabeto romano utilizados em línguas ocidentais.

A Figura 1 apresenta um exemplo para um parágrafo simples, contendo duas linhas de um texto. A ferramenta utilizada para a extração do texto usando OCR e também para gerar a imagem foi o PaddleOCR (PaddlePaddle Community, 2020). Nela pode-se observar que, apesar de capturar fielmente a segunda linha de texto, a primeira linha não foi detectada.

O processamento da mesma imagem pelo Docling resultou em uma leitura de imagem, ou seja, não foi possível obter um resultado textual, a ferramenta, ao não conseguir extrair nenhum texto da imagem, tratou-o como uma figura na página.

Portanto são encontrados desafios a serem superados tanto na etapa de pré-processamento, treinamento e pós-processamento dos textos. A comunidade em torno do OCR tem diversos experimentos em torno de textos históricos como é o caso de Drobac (2019), que faz a análise de jornais publicados entre 1771 e 1939, escritos em finlandês e sueco. Para esse experimento, as dificuldades encontradas estão nas duas línguas presentes nos jornais, o que fez com que a autora realizasse experimentos com modelos treinados individualmente para cada língua e também em um modelo treinado nas duas línguas ao mesmo tempo; Drobac (2019) menciona uma dificuldade comum encontrada no OCR de jornais, a sua formatação em várias colunas.

Já Hildebrand *et al.* (2019), realiza uma pesquisa com quatro historiadores que fizeram uso de ferramentas OCR para realizar alguma parte de suas pesquisas. Os principais usos desses pesquisadores são relacionados a localizar menções de palavras ou temas, bem como sua distribuição pelo tempo, como filtro para selecionar documentos. Portanto, pensa-se de forma generalista, uma coleção de documentos a serem levantas, para então ser feito o *close reading* dos que foram selecionados.

É possível observar que os autores citados dedicaram-se ao processamento computacional e não ao retorno desses textos para a leitura por humanos. Mesmo em uma pesquisa mais relacionada ao campo das humanidades como a de Hildebrand *et al.* (2019), onde o interesse dos entrevistados é pela leitura do texto em si, a formatação do texto extraído com a ferramenta

OCR não é mencionado, sendo algo secundário ou ignorado pela comunidade, o OCR é usado primariamente para pesquisas e processamento de dados.

Diante desse contexto, com o desenvolvimento da ferramenta proposta, espera-se que, ao integrar algoritmos de pré-processamento de imagens, bem como melhorias em algoritmos de OCR, seja possível obter uma ferramenta de uso prático para extração desses textos com destino a leitura em si, pense-se como modelo algo como as mudanças realizadas na edição física de um livro e a sua diagramação para e-book como objetivo para o consumo por um leitor final. Como resultado, permitir a disseminação dos textos digitalizados para uso geral da comunidade, resolvendo, assim, as lacunas no uso de OCR encontradas.

3 PROPOSTA

- Descrição do Trabalho: Para o desenvolvimento deste projeto, serão realizados experimentos comparativos das tecnologias OCR disponíveis. Após, os resultados serão comparados no processo de digitalização da base de dados escolhida. A partir dos resultados obtidos, será escolhida uma ferramenta com o objetivo de otimizar a fase de digitalização. Na sequencia, será construída uma API em Python que realize a digitalização e formatação do texto extraído de imagens e arquivos PDF, visando solucionar os erros cometidos na fase de digitalização, bem como no pós-processamento de texto encontrados no processamento de ferramentas de OCR. A API será integrada a um projeto web desenvolvido com o framework Laravel, visando facilitar a disponibilização do projeto a um público abrangente.
- Abordagem ou Solução Proposta: para o desenvolvimento da solução proposta, tem-se as seguintes etapas:
  - Definição da base de dados: O objeto escolhido para realizar essas conversões é a primeira edição da revista italiana de cinema, *Cinema & Film*, que teve suas atividades entre 1967-1970. Essa revista foi escolhida por ser um material de difícil acesso, disponível somente em bibliotecas especializadas de cinematecas, fazendo com que a maior parte do público atual da revista tenha acesso somente através de fotografias que foram transformadas em arquivos PDF. A formatação da revista em texto digital visa tanto facilitar a leitura da revista como a maior autonomia do usuário para ferramentas de tradução.
  - Implementação de pré-processamento da base: caso o arquivo processado seja um PDF, ele deve ser transformado em formato de imagem. A partir de cada imagem gerada, será realizada a sua binarização, redimensionamento e localização de regiões de interesse (LI *et al.*, 2022).
  - Implementação de OCR: serão utilizadas e comparadas diversas ferramentas OCR disponíveis, PaddleOCR (PaddlePaddle Community, 2020), Docling (Docling Team, 2024), OCR-D (OCR-D Project, 2020), Calamari-OCR (WICK; REUL; PUPPE, 2020) e Tesseract (TEAM, 2024).
  - A avaliação das ferramentas terá como critério a maior quantidade de acertos de palavras em conjunto com a facilidade que consegue-se extrair a formatação da página original, isto é, a detecção de linhas, parágrafos e imagens do texto original e a maleabilidade do *output* gerado por ela dessas informações para que haja a reformatação do texto.
  - Desenvolvimento da API em Python que irá processar as imagens enviadas e retornar o texto formatado em markdown ou PDF.

- Público-alvo ou Beneficiários: qualquer pessoa que deseje digitalizar um texto, mais potencialmente um público acadêmico.
- Resultados Esperados: treinamento de uma ferramenta que possa digitalizar e formatar o texto digitalizado de forma eficiente, baseado em sua formatação original. A ferramente irá possibilitar a inserção de um documento, seja em PDF ou imagem, e trará como resultado o texto digitalizado. Além disso, permitir a disponibilização dessa ferramenta ao público.

4 CONCLUSÃO

Diante da inacessibilidade de determinados livros (VARELLA; FERREIRA, 2012) e outros textos publicados antes da era digital, bem como do tratamento secundário da comunidade de OCR para com o uso dos documentos analisados em texto digital próprio para a leitura casual, determinamos a necessidade de uma ferramenta que, trabalhando com OCR, dedique-se na extração de textos de imagens ou arquivos PDF, visando facilitar a leitura e disseminação desses textos.

Para atingir esse objetivo será realizada uma pesquisa comparativa entre ferramentas OCR, sendo selecionada aquela que mais se destacar no acerto de palavras e no retorno da formatação original do texto. A ferramenta selecionada será treinada para atingir maior nível de precisão.

Será construída uma API em Python para realizar o trabalho de processamento do OCR e a formatação do texto. Isso será integrado por uma interface web desenvolvida com o framework PHP Laravel, com o objetivo de disponibilizar o projeto para a maior quantidade de pessoas possível.

Portanto, o presente trabalho possui tanto uma via científica e experimental como prática, buscando comparar as ferramentas OCR e solucionar algo que é até então ignorado pelas ferramentas, a devida formatação do texto para leitura, também tendo como objetivo disponibilizar os resultados dessa pesquisa como uma ferramenta de fácil uso e acesso para a comunidade.

REFERÊNCIAS

BOENIG, M. *et al.* Labelling ocr ground truth for usage in repositories. *In*: **Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019)**. New York, NY, USA: ACM, 2019. p. 3–8. Disponível em: https://doi.org/10.1145/3322905.3322916.

Docling Team. **Docling**. 2024. https://github.com/docling-project/docling. Acesso em: 17 abr. 2025. Veja também: https://arxiv.org/abs/2408.09869. Disponível em: https://github.com/docling-project/docling.

DROBAC, S. **DATeCH2019 - Session 4. Senka Drobac**. 2019. Apresentação do artigo "Improving OCR of Historical Newspapers and Journals Published in Finland". Disponível em: https://www.youtube.com/watch?v=3ZhHPx00wjg.

GABRIAL, B. History of writing technologies. *In*: BAZERMAN, C. (Ed.). **Handbook of Research on Writing: History, Society, School, Individual, Text**. [*S.l.*]: Taylor & Francis Group, 2009. p. 27–39.

HILDEBRAND, M. *et al.* Impact analysis of ocr quality on research tasks in digital archives. **CORE**, 2019. Disponível em: https://core.ac.uk/download/pdf/301647312.pdf.

KUMAR, B.; TANWAR, S.; TIWARI, S. Ocr using traditional and dnn approach. *In*: ____. **Mechatronics**. CRC Press, 2025. p. 196–210. ISBN 9781003494478. Disponível em: http://dx.doi.org/10.1201/9781003494478-11.

LI, C. *et al.* **Dive into OCR**. PaddleOCR Community, 2022. Disponível em: https://github.com/PaddleOCR-Community/Dive-into-OCR.

OCR-D Project. **OCR-D: Integrated Workflow for OCR in Historical Documents**. 2020. https://ocr-d.de/en. Acesso em: 17 abr. 2025. Disponível em: https://ocr-d.de/en.

PACKER, T. L. Performing information extraction to improve ocr error detection in semi-structured historical documents. *In*: **Proceedings of the 2011 Workshop on Historical Document Imaging and Processing**. ACM, 2011. (HIP '11), p. 67–74. Disponível em: http://dx.doi.org/10.1145/2037342.2037354.

PaddlePaddle Community. **PaddleOCR**. 2020. https://github.com/PaddlePaddle/PaddleOCR. Acesso em: 17 abr. 2025. Disponível em: https://github.com/PaddlePaddle/PaddleOCR.

SCHMANDT-BESSERAT, D.; ERARD, M. Origins and forms of writing. *In*: BAZERMAN, C. (Ed.). **Handbook of Research on Writing: History, Society, School, Individual, Text**. [*S.l.*]: Taylor & Francis Group, 2009. p. 7–24.

TEAM, T. T. O. **Tesseract OCR**. 2024. https://github.com/tesseract-ocr/tesseract. Acesso em: maio de 2025.

VARELLA, G.; FERREIRA, I. M. **Livros inacessíveis, lei antiquada**. 2012. Acesso em: 28 abr. 2025. Disponível em: https://idec.org.br/em-acao/artigo/livros-inacessiveis-lei-antiquada.

WICK, C.; REUL, C.; PUPPE, F. Calamari - a high-performance tensorflow-based deep learning package for optical character recognition. **Digital Humanities Quarterly**, v. 14, n. 2, 2020. Disponível em: https://digitalhumanities.org/dhq/vol/14/2/000451/000451.html.