UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

WALISSON NASCIMENTO DE FARIAS

APERFEIÇOANDO O RECONHECIMENTO ÓPTICO DE CARACTERES EM IMAGENS DE DOCUMENTOS PESSOAIS

CAMPINA GRANDE - PB

2023

WALISSON NASCIMENTO DE FARIAS

APERFEIÇOANDO O RECONHECIMENTO ÓPTICO DE CARACTERES EM IMAGENS DE DOCUMENTOS PESSOAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Herman Martins Gomes

CAMPINA GRANDE - PB

2023

WALISSON NASCIMENTO DE FARIAS

APERFEIÇOANDO O RECONHECIMENTO ÓPTICO DE CARACTERES EM IMAGENS DE DOCUMENTOS PESSOAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Herman Martins Gomes

Orientador – UASC/CEEI/UFCG

Eanes Torres Pereira

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 17 de NOVEMBRO de 2023.

CAMPINA GRANDE - PB

RESUMO

O reconhecimento óptico de caracteres (OCR) desempenha um papel fundamental na digitalização e processamento de documentos pessoais, no entanto, enfrenta desafios significativos de precisão e eficiência, visto que as ferramentas que realizam OCR ainda dependem muito da qualidade da entrada de dados e das condições em que os documentos são escaneados ou fotografados. Para aperfeiçoar o reconhecimento óptico de caracteres (OCR), propõe-se a utilização da combinação de técnicas de pré-processamento e pós-processamento a fim de melhorar a qualidade do OCR.

O processo inicia-se através da coleta de um conjunto de dados representativo de imagens de documentos pessoais. Após a coleta, realiza-se o pré-processamento e pós-processamento das imagens, seguindo então do OCR e a utilização de uma métrica que avalia o OCR obtido.

As técnicas de pré-processamento incluíram modificação do DPI das imagens, suavização da imagem e conversão para escala de cinza, seguida pela aplicação do OCR. Além disso, houve um pós-processamento para remover a acentuação do texto extraído e convertê-lo em letras maiúsculas. Os resultados indicaram que o pré-processamento melhorou significativamente a precisão do OCR para documentos de identidade (RG), aumentando o F1-Score de 0.33 (sem pré-processamento) para 0.53 (com pré-processamento). Para imagens de CPF, o pré-processamento resultou em uma precisão de 73.48% e uma taxa de erro de 26.52%, enquanto o OCR sem pré-processamento teve uma precisão de 36.46% e uma taxa de erro de 63.54%.

Este estudo visa investigar técnicas com o propósito de melhorar o reconhecimento óptico de caracteres em documentos pessoais, contribuindo para maior precisão do OCR, com potenciais benefícios para aplicações que realizam a extração de conteúdo de imagens de documentos pessoais.

IMPROVING OPTICAL CHARACTER RECOGNITION IN PERSONAL DOCUMENT IMAGES

ABSTRACT

Optical character recognition (OCR) plays a key role in the digitization and processing of personal documents, however, it faces accuracy and efficiency challenges, since the tools that perform OCR still depend heavily on the quality of the input data and the conditions in which the documents are scanned or photographed. To improve optical character recognition, it is proposed a combination of pre-processing and post-processing techniques to improve OCR quality.

The process begins by collecting a representative dataset of images of personal documents. After that, the images are pre-processed and post-processed, followed by OCR and the use of a metric that evaluates the OCR obtained.

Pre-processing techniques included modifying the DPI of the images, smoothing the image and converting it to grayscale, followed by the application of OCR. In addition, post-processing was carried out to remove accents marks from the extracted text and convert it into capital letters. The results indicated that pre-processing method significantly improved OCR accuracy for identity documents (ID), increasing the F1-Score from 0.33 (without pre-processing) to 0.53 (with pre-processing). For CPF images, pre-processing procedure resulted in an accuracy of 73.48% and an error rate of 26.52%, while OCR without pre-processing had an accuracy of 36.46% and an error rate of 63.54%.

This study aims to investigate techniques for improving optical character recognition in personal documents, contributing to greater OCR accuracy, with potential benefits for applications that extract content from personal document images.

Aperfeiçoando o Reconhecimento Óptico de Caracteres em Imagens de Documentos Pessoais

Walisson Nascimento de Farias
Universidade Federal de Campina Grande
Campina Grande
walisson.farias@ccc.ufcg.edu.br

Herman Martins Gomes
Universidade Federal de Campina Grande
Campina Grande
hmg@computacao.edu.br

RESUMO

O reconhecimento óptico de caracteres (OCR) desempenha um papel fundamental na digitalização e processamento de documentos pessoais, no entanto, enfrenta desafios significativos de precisão e eficiência, visto que as ferramentas que realizam OCR ainda dependem muito da qualidade da entrada de dados e das condições em que os documentos são escaneados ou fotografados. Para aperfeiçoar o reconhecimento óptico de caracteres (OCR), propõe-se a utilização da combinação de técnicas de pré-processamento e pós-processamento a fim de melhorar a qualidade do OCR. O processo inicia-se através da coleta de um conjunto de dados representativo de imagens de documentos pessoais. Após a coleta, realiza-se o pré-processamento e pós-processamento das imagens, seguindo então do OCR e a utilização de uma métrica que avalia o OCR obtido. As técnicas de pré-processamento incluíram modificação do DPI das imagens, suavização da imagem e conversão para escala de cinza, seguida pela aplicação do OCR. Além disso, houve um pós-processamento para remover a acentuação do texto extraído e convertê-lo em letras maiúsculas. Os resultados indicaram que o pré-processamento melhorou significativamente a precisão do OCR para documentos de identidade (RG), aumentando o F1-Score de 0.33 (sem pré-processamento) para 0.53 (com pré-processamento). Para imagens de CPF, o pré-processamento resultou em uma precisão de 73.48% e uma taxa de erro de 26.52%, enquanto o OCR sem pré-processamento teve uma precisão de 36.46% e uma taxa de erro de 63.54%. Este estudo visa investigar técnicas com o propósito de melhorar o reconhecimento óptico de caracteres em documentos pessoais, contribuindo para maior precisão do OCR, com potenciais benefícios para aplicações que realizam a extração de conteúdo de imagens de documentos pessoais.

PALAVRAS-CHAVE

OCR, documentos pessoais, pré-processamento.

1. INTRODUÇÃO

No cenário atual da transformação digital, a digitalização e o processamento de documentos pessoais contribuem para a simplificação de processos e agilização de operações em diversas áreas. O Reconhecimento Óptico de Caracteres (OCR) é fundamental nesse processo, permitindo a conversão de informações contidas em imagens de documentos em texto digital de forma automatizada. No entanto, apesar dos avanços tecnológicos, tecnicas de OCR ainda enfrentam desafios e limitações no tocante à qualidade das imagens de entrada, as quais podem conter ruídos de aquisição, degradações por uso ou por idade dos documentos, entre outros.,

Atualmente, as ferramentas de OCR disponíveis são frequentemente ineficazes e incompletas na tarefa de extrair informações precisas de documentos pessoais, como CPFs, documentos de identidade, carteiras de motorista, etc. A precisão e a eficiência do OCR são importantes para garantir a integridade dos dados e para evitar erros. Portanto, a busca por técnicas que aprimorem o reconhecimento óptico de caracteres torna-se uma necessidade.

Este trabalho de conclusão de curso tem como objetivo principal abordar o problema de precisão e eficiência no reconhecimento óptico de caracteres em imagens de documentos pessoais. Para alcançar esse objetivo, é proposta a utilização da combinação de técnicas de processamento de imagem e aprendizado de máquina, abrangendo todo o ciclo de processamento, desde a coleta de um conjunto de dados representativo de imagens de documentos pessoais até o pré-processamento e pós-processamento das imagens, para aplicação do OCR e na utilização de uma métrica especializada para avaliar a qualidade do OCR obtido.

O restante deste documento está organizado como segue. Na próxima seção (2), será apresentada uma fundamentação teórica sobre OCR e métricas utilizadas para avaliação dos resultados. Em seguida, na seção (3), os materiais e métodos utilizados na pesquisa, incluindo o pré-processamento das imagens, a aplicação do OCR e das métricas de avaliação. Os resultados obtidos serão discutidos na seção subsequente (4), destacando a importância do pré-processamento e pós-processamento nas taxas de precisão e eficiência do OCR. Finalmente na seção de conclusão (5) são tecidas as considerações finais e possíveis trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção de fundamentação teórica, abordam-se os conceitos essenciais relacionados ao Reconhecimento Óptico de Caracteres (OCR) e às métricas de avaliação, incluindo as métricas F1-Score, Precisão e Taxa de Erro.

2.1. OCR (Reconhecimento Óptico de Caracteres)

O Reconhecimento Óptico de Caracteres (OCR) é uma tecnologia que visa a conversão de texto contido em imagens ou documentos digitalizados em texto editável. As técnicas de OCR têm uma ampla gama de aplicações, desde a digitalização de documentos históricos até a automação de processos de negócios. Trata-se de uma tecnologia que evoluiu significativamente nas últimas décadas, desempenhando um papel essencial na transformação digital de documentos. Os algoritmos e modelos de

OCR estão se tornando cada vez mais precisos, graças aos avanços em técnicas de aprendizado profundo e redes neurais convolucionais. Isso torna possível a extração precisa de texto mesmo em condições desafiadoras, como documentos antigos ou imagens de baixa qualidade.

O pytesseract é uma biblioteca Python que atua como uma interface para a biblioteca tradicional de OCR chamada Tesseract. Tesseract é um software de código aberto amplamente reconhecido por sua capacidade de reconhecimento óptico de caracteres (OCR) de alta precisão. O pytesseract permite que os desenvolvedores utilizem a funcionalidade do Tesseract OCR em seus programas Python. Esta biblioteca fornece uma maneira simples e conveniente de extrair texto de imagens ou documentos digitalizados.

2.3. Avaliação de desempenho de OCR

No contexto de OCR, o termo "*Ground Truth*" se refere à versão correta e exata do texto que se espera ser extraído de uma imagem ou documento. O *Ground Truth* é frequentemente utilizado como um padrão de comparação para avaliar o desempenho de um sistema de OCR. Para criar um conjunto de dados de treinamento e teste eficaz, é essencial ter um *Ground Truth* confiável que represente com precisão o conteúdo dos documentos pessoais. Este elemento é crucial para a avaliação de métricas de desempenho, como o F1-Score, a Precisão e a Taxa de Erro, as quais são discutidas a seguir.

2.3.1. F1-Score (RGs)

O F1-Score é uma métrica de avaliação amplamente utilizada na área de processamento de linguagem natural como também, no contexto de OCR. Esta métrica visa medir a precisão e a capacidade de um sistema de OCR em equilibrar a taxa de verdadeiros positivos, falsos positivos e falsos negativos.

O F1-Score combina duas métricas importantes. A precisão mede a proporção de verdadeiros positivos em relação ao total de positivos previstos pelo sistema de OCR. Em outras palavras, representa a capacidade do sistema de não classificar incorretamente os caracteres corretos como incorretos. Já a revocação mede a proporção de verdadeiros positivos em relação ao total de positivos reais presentes no documento. Isso demonstra a capacidade do sistema de capturar todos os caracteres corretos na imagem.

A Precisão e Revocação são calculadas a partir das seguintes fórmulas:

$$Precisão = \frac{Número\ de\ palavras\ corretamente\ identificadas\ pelo\ OCR}{Total\ de\ palavras\ identificadas\ pelo\ OCR}$$

$$Revocação = \frac{Número\ de\ palavras\ corretamente\ identificadas\ pelo\ OCR}{Total\ de\ palavras\ do\ texto\ de\ referência}$$

O F1-Score é calculado usando a seguinte fórmula:

$$F1\ Score = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação}$$

2.3.2. Taxa de Erro

A Taxa de Erro, por outro lado, avalia a proporção de itens que o sistema OCR não conseguiu extrair corretamente em relação ao total de itens presentes no documento original. Isso inclui os falsos positivos e falsos negativos, ou seja, os caracteres incorretamente identificados e os caracteres que não foram identificados, respectivamente. A Taxa de Erro é uma medida importante para avaliar a eficácia do OCR na identificação de caracteres.

$$Taxa\ de\ Erro = \frac{Número\ de\ dados\ ausentes\ no\ OCR}{Total\ de\ dados\ no\ texto\ de\ referência} \times 100\%$$

3. MATERIAIS E MÉTODOS

Nesta seção são descritas as etapas principais deste estudo, que abrangem a implementação e aprimoramento do Reconhecimento Óptico de Caracteres (OCR) para imagens de documentos de identidade (RG) e para imagens de CPFs. A seção está dividida em três partes principais: Detalhes de Implementação, Aperfeiçoamento do OCR para imagens de documentos de identidade (RG) e Aperfeiçoamento do OCR para imagens de CPFs.

3.1. Detalhes de implementação

O dataset escolhido para estudo está disponível no link: https://github.com/ricardobnjunior/Brazilian-Identity-Document-Dataset. A escolha desta fonte de dados para o estudo foi motivada pela diversidade das imagens, com contrastes, rotações e cores diversas. O dataset é composto por um total de 28.800 imagens de documentos de vários tipos, totalizando 7GB de dados. Essas imagens estão distribuídas em oito categorias distintas, sendo cada categoria composta por 3.600 imagens, todas no formato JPG. As categorias incluem: Frente e Verso da Carteira Nacional de Habilitação (CNH), Frente da Carteira Nacional de Habilitação (CNH), Verso da Carteira Nacional de Habilitação (CNH), Frente do Cadastro de Pessoas Físicas (CPF), Verso do Cadastro de Pessoas Físicas (CPF), Frente da Carteira de Identidade (RG), Verso da Carteira de Identidade (RG), e Frente e Verso da Carteira de Identidade (RG). Todos os dados do dataset são fictícios e foram gerados para fins de pesquisa e todas as informações pessoais contidas nele são simuladas e não representam informações reais de cidadãos brasileiros.

As seguintes bibliotecas foram empregadas neste trabalho:

- Opencv-python: O OpenCV (Open Source Computer Vision Library) é uma biblioteca de código aberto utilizada para processamento de imagem e visão computacional. Foi utilizada uma interface Python para a biblioteca.

- Pillow: O Pillow é uma biblioteca Python de processamento de imagens que oferece funcionalidades para abrir, manipular e salvar imagens em vários formatos.

- Unidecode: A biblioteca Unidecode é usada para normalizar e converter caracteres Unicode em representações ASCII aproximadas.

O código fonte contendo a implementação das técnicas investigadas neste trabalho está disponível no seguinte repositório gitbub: https://github.com/alissonfariias/OCR-improvement/tree/main.

3.2. Aperfeiçoamento do OCR para imagens de documentos de identidade (RG):

3.2.1. Pré-processamento dos arquivos 'ground truth':

Inicialmente os arquivos "ground truth" passaram por uma operação de pré-processamento, onde foi extraído apenas o conteúdo da coluna 'transcription' dos arquivos, e, após a extração, foram removidas a acentuação e os caracteres especiais das transcrições para depois armazenar as transcrições processadas em novos arquivos. As Figuras 1 e 2 contém

```
x, y, width, height, transcription
82, 378, 31, 468, REPÚBLICA FEDERATIVA DO BRASIL
141, 377, 28, 454, ESTADO DE MINAS GERAIS
[177, 172, 202, 205], [314, 890, 890, 315], -1, -1, POLÍCIA CIVIL DO ESTADO DE MINAS GERAIS
[238, 206, 204, 234], [401, 401, 805, 805], -1, -1, INSTITUTO DE IDENTIFICAÇÃO
[343, 530, 530, 342], [679, 681, 707, 705], -1, -1, POLEGAR DIREITO
[712, 712, 732, 735], [462, 741, 742, 461], -1, -1, ASSINATURA DO TITULAR
[789, 787, 810, 814], [443, 786, 786, 443], -1, -1, CARTEIRA DE IDENTIDADE
163, 170, 35, 146, 8280-0
```

Figura 1: Exemplo de arquivo 'ground truth' antes do pré-processamento

```
REPUBLICA FEDERATIVA DO BRASIL
ESTADO DE MINAS GERAIS
POLICIA CIVIL DO ESTADO DE MINAS GERAIS
INSTITUTO DE IDENTIFICACAO
POLEGAR DIREITO
ASSINATURA DO TITULAR
CARTEIRA DE IDENTIDADE
8280-0
```

Figura 2: Exemplo de arquivo 'ground truth' após pré-processamento

3.2.2. Pré-processamento dos arquivos de imagens dos documentos de identidade (RG):

Inicialmente as imagens passaram por uma série de etapas de pré-processamento com o objetivo de tornar a detecção de texto realizada pelo OCR mais precisa e eficaz. As etapas de pré-processamento estão listadas a seguir:

Modificação do DPI da imagem:

A primeira etapa é modificar o DPI (*dots per inch*) da imagem original. O DPI é um valor que determina a densidade de pontos por polegada na imagem, e imagens de alta resolução têm mais detalhes e mais legibilidade, tornando mais fácil para o OCR identificar os caracteres com precisão. Se uma imagem tiver baixa resolução, pode haver perda de detalhes, favorecendo erros no OCR. Nesse caso, calculam-se as novas dimensões da imagem com base no DPI especificado (neste caso, 500) e a imagem passa a ter novas dimensões.

Suavização da imagem:

As imagens de documentos de identidade podem conter ruído ou granulação, como pequenas variações de intensidade de cor ou pequenos pontos que não fazem parte do texto. Foi aplicada uma operação de suavização, utilizando o filtro Gaussiano, que ajuda a reduzir esses artefatos, melhorando a legibilidade do texto e a precisão do OCR, como também suaviza as transições entre áreas de intensidade de cor, sendo útil para suavizar as bordas dos caracteres, tornando os caracteres mais nítidos e facilitando a segmentação dos caracteres durante o processo de OCR.

O tamanho do *kernel* do filtro (5x5) é especificado para controlar o grau de suavização, e neste caso, (5, 5) indica que o kernel é uma matriz 5x5, e quanto maior o tamanho do kernel, maior será o efeito de suavização aplicado à imagem. No entanto, aumentar muito o tamanho do kernel pode resultar em uma suavização excessiva e perda de detalhes na imagem. Na Figura 3 é apresentada uma imagem de documento resultante da suavização aplicada.

Figura 3: Exemplo de imagem de documento de identidade (RG) com filtro Gaussiano

Conversão para escala de cinza:

A imagem suavizada é convertida para escala de cinza usando a função cv2.cvtColor da biblioteca OpenCV. A conversão para escala de cinza é comum em tarefas de OCR, pois ao reduzir a imagem a tons de cinza torna-se mais fácil a detecção de texto pois este é geralmente mais escuro ou mais claro em relação ao fundo da imagem. Consequentemente, essa etapa contribui para outras etapas de pré-processamento subsequentes.

Figura 4: Exemplo de imagem de documento de identidade (RG) com escala de cinza

3.2.3. Aplicação do OCR

O OCR é aplicado nas imagens tanto na sua forma original quanto após as etapas pré-processamento que inclui redimensionamento, suavização e conversão para escala de cinza. O texto extraído é salvo em arquivos separados para ambas as versões da imagem, sendo útil para comparar os resultados do OCR antes e depois do pré-processamento. Deste modo é possível verificar qual abordagem fornece melhores resultados para o conjunto de imagens.

3.2.4. Pós-processamento

Após a extração do texto, a acentuação do texto extraído é removida, com objetivo de tornar o texto mais consistente e evitar problemas de codificação de caracteres especiais, convertendo assim os caracteres acentuados em seus equivalentes não acentuados. Por exemplo, "á" se tornaria "a", "ç" se tornaria "c", etc, e após remover a acentuação, todo o texto modificado é convertido em letras maiúsculas, garantindo assim a consistência da extração e a independência da formatação original do texto na imagem. Por fim, o texto pós-processado é salvo em um arquivo de texto para uso posterior e comparação entre os OCRs gerados.

3.3. Aperfeiçoamento do OCR para imagens de CPFs

3.3.1. Pré-processamento dos arquivos 'ground truth':

Inicialmente todos os arquivos GT foram processados, e através de uma expressão regular (cpf_pattern), apenas os CPFs foram identificados no formato XXX.XXX.XXX-XX, onde X é um dígito, e, após identificados, todos os CPFs foram salvos em um único arquivo de texto para posterior análise comparativa com o OCR.

```
x, y, width, height, transcription
[380, 392, 407, 417, 431, 425, 419, 405, 380, 363, 380, 401, 409, 401, 373, 365],
[368, 354, 328, 298, 242, 196, 143, 120, 78, 88, 120, 164, 227, 282, 352, 355], -1,
-1, REPÚBLICA FEDERATIVA DO BRASIL
[426, 447, 439, 429, 415, 401, 391, 419], [142, 138, 116, 100, 68, 54, 70, 109],
-1, -1, DE 1889
[371, 385, 404, 425, 453, 458, 438, 436, 418, 403], [402, 415, 394, 363, 304, 260,
258, 282, 330, 365], -1, -1, 15 DE NOVEMBRO
32, 430, 25, 257, MINISTÉRIO DA FAZENDA
63, 406, 25, 283, SECRETARIA DA RECEITA FEDERAL
100, 444, 84, 201, CPF
201, 407, 24, 374, CADASTRO DE PESSOAS FÍSICAS
233, 573, 28, 210, NÚMERO DE INSCRIÇÃO
264, 495, 38, 285, 029.232.665-36
325, 716, 24, 63, NOME
345, 374, 30, 406, ARDUINO AUZIER IAMAGUTE
414, 652, 27, 126, NASCIMENTO
438, 643, 28, 132, 18/06/1976
```

Figura 5: Exemplo de arquivo textual (ground truth) de um CPF

```
024.661.386-62
027.566.460-08
323.660.704-11
992.740.997-46
289.508.573-08
052.694.286-01
219.767.364-55
563.314.219-55
466.333.798-89
…
```

Figura 6: Exemplo de arquivo textual criado que armazena apenas os CPFs extraídos dos arquivos ground truth

3.3.2. Pré-processamento dos arquivos das imagens dos CPFs:

O objetivo dessa etapa é garantir uma extração precisa dos números de CPFs através de estratégias de pré-processamento e reconhecimento de texto, a fim de melhorar a taxa de sucesso na extração dos CPFs. Os CPFs extraídos são registrados em um único arquivo de saída. As etapas de pré-processamento são semelhantes às etapas utilizadas para pré-processamento das imagens de documentos de identidade, utilizando a conversão para escala de cinza e a etapa de modificação do DPI da imagem, porém com adição de uma nova etapa de limiarização, onde a imagem em escala de cinza é convertida em uma imagem binária. A limiarização divide a imagem em pixels pretos e brancos, com base em um valor de limiar, e no contexto das imagens dos CPFs, os valores de limiar são configurados para binarizar a imagem.

Figura7: Exemplo de imagem do CPF convertida para escala de cinza

Figura 8: Exemplo de imagem do CPF com Limiarização

4. RESULTADOS

Nesta seção, apresentam-se os resultados obtidos do estudo sobre o aprimoramento do Reconhecimento Óptico de Caracteres (OCR) para documentos pessoais, com foco em documentos de identidade (RG) e Cadastro de Pessoas Físicas (CPF), onde se buscou avaliar o impacto das técnicas de pré-processamento e pós-processamento na precisão do OCR, bem como comparar o desempenho do OCR com e sem essas técnicas.

4.1. Resultado do OCR para as imagens de documentos de identidade (RG)

Foram realizadas avaliações comparativas entre a performance de um OCR com e sem pré-processamento+ pós-processamento das imagens das amostras dos documentos de identidade (RG). As métricas de avaliação utilizadas foram F1-Score, que leva em consideração tanto a precisão quanto a revocação do OCR. A média do F1-Score para o OCR com pré-processamento e pós-processamento foi de 0.53. Em contrapartida, o OCR sem pré-processamento das imagens obteve uma média de F1-Score de 0.33.

4.2. Resultado do OCR para as imagens de CPFs

Ao avaliar o desempenho do OCR com o uso de pré-processamento para a extração de CPFs, foram utilizadas duas métricas: Precisão e Taxa de Erro, e elas produziram os seguintes resultados:

Precisão: 73.48%
Taxa de Erro: 26.52%

Essas métricas foram obtidas por meio de uma análise entre os CPFs extraídos pelo sistema OCR e os CPFs do "ground truth". Isso permitiu determinar quantos CPFs eram comuns a ambos os conjuntos. A precisão foi calculada como a porcentagem de CPFs corretamente extraídos pelo OCR em relação ao total de CPFs no "ground truth". A Taxa de Erro refere-se aos CPFs presentes no conjunto do "ground truth", mas ausentes no conjunto do OCR, representando os CPFs que o sistema OCR não conseguiu extrair corretamente.

De maneira análoga, a avaliação do desempenho do OCR sem o uso de pré-processamento para a extração de CPF produziu os seguintes resultados:

Precisão: 36.46%
Taxa de Erro: 63.54%

No contexto de imagens de CPFs, a utilização da métrica Taxa de Erro é mais adequada, pois em muitos cenários, por se tratar de uma informação sensível, a prioridade deve ser evitar a utilização de informações incorretas e neste sentido a revocação pode ser menos crítica.

As métricas demonstram que a utilização dessas técnicas aumentou significativamente a precisão do OCR, tornando-o mais confiável na extração de informações precisas. Esses resultados são fundamentais para aprimorar a qualidade e confiabilidade do OCR em aplicações que envolvem documentos pessoais.

5. CONCLUSÃO

O uso de técnicas de pré-processamento e pós-processamento no OCR em imagens de CPFs melhorou significativamente o desempenho do OCR na extração de CPFs. A precisão aumentou de 36.46% para 73.48% e a taxa de erro diminuiu (de 63.54% para 26.52%) quando o pré-processamento foi aplicado . Para as imagens de documentos de identidade, o OCR com uso das técnicas de pré-processamento e pós-processamento resultou em um desempenho superior na detecção e reconhecimento de caracteres quando comparado sem o uso das técnicas, atingindo um F1-Score de 0.53. Isso ressalta o impacto significativo dessas etapas no aprimoramento da precisão do OCR em relação ao texto de referência. Por outro lado, o OCR sem pré-processamento e pós processamento obteve um F1-Score de 0.33, demonstrando um desempenho inferior na correspondência com o texto de referência, ou seja, o OCR sem tratamento anterior ou posterior é menos preciso na identificação e reconhecimento de caracteres.

Esses resultados enfatizam a importância do pré-processamento e pós-processamento de imagens para otimizar a performance do OCR. Ao abordar tarefas de reconhecimento de caracteres, a aplicação de técnicas de pré-processamento pode contribuir para aprimorar a precisão do OCR. Essas melhorias têm implicações relevantes em várias aplicações que dependem do OCR, como digitalização de documentos e reconhecimento de texto em imagens, bem como em diversas outras tarefas relacionadas ao processamento de texto.

Alguns possíveis trabalhos futuros incluem a incorporação de novas técnicas de processamento, dada a constante evolução das técnicas de processamento de imagem. Há um vasto território para exploração, sendo essencial considerar a adoção de técnicas de pré-processamento e pós-processamento emergentes. Pesquisas futuras podem focar na adaptação e integração de inovações tecnológicas, como a aplicação de redes neurais adversariais (GANs) para aprimorar a qualidade das imagens, ou a exploração de métodos avançados de filtragem e correção de distorções. Além disso, é importante considerar a ampliação das técnicas de pré-processamento para outros tipos de documentos, pois embora este estudo tenha se concentrado em

técnicas aplicadas a RGs e CPFs, a melhoria das técnicas de pré-processamento e pós-processamento pode ser estendida a outros tipos de documentos, como certidões de nascimento, passaportes, carteiras de motorista, e assim por diante.

6. REFERÊNCIAS

[1] Farias, A. (2018, 17 de maio). OCR: conheça a tecnologia que reconhece e auxilia a estruturação de dados. Disponível em: https://www.neomind.com.br/blog/ocr-estruturacao-de-dados/

[2] Javaid, S. (2023, 3 de fevereiro). 5 Steps to Prepare OCR Training Data in 2023. Disponível em: https://research.aimultiple.com/ocr-training-data/

[3] Lim, J. (2021, 28 de fevereiro). Targeted OCR on Documents with OpenCV and PyTesseract. Disponível em: https://medium.com/analytics-vidhya/targeted-ocr-on-documents-with-opencv-and-pytesseract-edc10b5ecb62

[4] Sharma, R. (2022, 26 de abril). How to Increase Accuracy of Tesseract. Disponível em: https://aurigait.com/blog/how-to-increase-accuracy-of-tesseract/

[5] Uezima, L. K., Nishimoto, H. Y. Y., Barbosa, R., & Basile, A. L. (s.d.). Utilizando o Tesseract OCR para reconhecimento de textos a mão. Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie. São Paulo, SP, Brasil. Disponível em: :https://adelpha-api.mackenzie.br/server/api/core/bitstreams/e9fcf867-97ef-453c-967a-b0d98483b632/content.

[6] Floriano, D. (2020, 12 de fevereiro). Deep Learning e OCR — Reconhecimento de Documentos. Disponível em: https://medium.com/senior/deep-learning-e-ocr-reconhecimento-de-documentos-76c580ca93b1.

[7] SOARES, Álysson de Sá; DAS NEVES JUNIOR, Ricardo Batista ; BEZERRA, Byron Leite Dantas. BID Dataset: a challenge dataset for document processing tasks. In: WORKSHOP DE TRABALHOS EM ANDAMENTO - CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), 33. , 2020, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 143-146. DOI: https://doi.org/10.5753/sibgrapi.est.2020.12997.

[8] Kuguoglu, B. K. (2018, 6 de junho). How to Use Image Preprocessing to Improve the Accuracy of Tesseract. Disponível em: https://www.freecodecamp.org/news/getting-started-with-tesseract-part-ii-f7f9a0899b3f/.

[9] Purohit, K. (2019, 9 de maio). Tutorial: Building a Custom OCR Using YOLO and Tesseract. Disponível em: https://medium.com/saarthi-ai/how-to-build-your-own-ocr-a5bb91b622ba.

[10] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of Post-OCR Processing Approaches. ACM Comput. Surv. 54, 6, Article 124 (July 2021), 37 pages. https://doi.org/10.1145/3453476.

[11] Fonseca Cacho, Jorge Ramon, "Improving OCR Post Processing with Machine Learning Tools" (2019). UNLV Theses, Dissertations, Professional Papers, and Capstones. 3722. http://dx.doi.org/10.34917/16076262.

[12] Gitansh Khirbat. 2017. OCR Post-Processing Text Correction using Simulated Annealing (OPTeCA). In Proceedings of the Australasian Language Technology Association Workshop 2017, pages 119–123, Brisbane, Australia.

[13] Parande, A. (2019, 11 de outubro). A Hitchhiker's Guide to OCR. Disponível em: https://medium.com/analytics-vidhya/a-hitchhikers-guide-to-ocr-8b869f4e3743

Sobre o autor:

Walisson Nascimento de Farias é aluno do curso de Ciência da Computação na Universidade Federal de Campina Grande.