

**PROCESSAMENTO DE SINAIS**

**Alan Silva Furquim**

**Vitor Augusto de Lima Soares**

**Integrando Python com o Matlab para detectar e localizar um texto dentro de um arquivo pdf**

**Sorocaba, SP**

**2023**

**SUMÁRIO**

[1 INTRODUÇÃO 3](#_Toc149654908)

[2 DESENVOLVIMENTO 3](#_Toc149654909)

[3 CONCLUSÃO 3](#_Toc149654910)

[REFERÊNCIAS 4](#_Toc149654911)

# 1 INTRODUÇÃO

O exemplo escolhido, "How to detect & localize a text in pdf using OCR in MATLAB", do File Exchange, foi selecionado para resolver um desafio específico: a detecção e localização de texto em documentos PDF usando OCR. Esta escolha foi motivada pela limitação do MATLAB em lidar diretamente com arquivos PDF para esse propósito.

Este projeto é de suma importância em contextos que demandam a análise textual de documentos PDF, como em processamento de documentos e automação de escritórios. A colaboração entre MATLAB e Python neste projeto não apenas resolve essa limitação, mas também destaca a relevância da interoperabilidade entre plataformas, um aspecto crucial na atual paisagem tecnológica. Ao integrar a funcionalidade do Python (pdf2image ou fitz) com o ambiente MATLAB, este projeto não só supera obstáculos técnicos, mas também abre portas para futuras colaborações interplataforma, oferecendo uma solução técnica inovadora para um problema comum no processamento de documentos PDF.

# 2 DESENVOLVIMENTO

**2.1 OCR no MATLAB: Teoria e Funcionamento**

A biblioteca OCR (Reconhecimento Óptico de Caracteres) do MATLAB é uma poderosa ferramenta que se baseia em algoritmos sofisticados para transformar texto contido em imagens em caracteres legíveis computacionalmente. O processo começa com a conversão da imagem contendo texto em escala de cinza para realçar as características dos caracteres. Em seguida, técnicas avançadas de segmentação de texto são aplicadas para identificar regiões de interesse que contêm palavras ou frases.

O reconhecimento dos caracteres ocorre por meio de algoritmos de aprendizado de máquina, como redes neurais convolucionais (CNNs), que são treinadas para reconhecer padrões específicos nos caracteres. Após a identificação dos caracteres, o OCR do MATLAB realiza a correspondência com seu banco de dados de caracteres conhecidos, fornecendo assim o texto reconhecido. Esse processo complexo e altamente eficiente torna a biblioteca OCR do MATLAB uma ferramenta essencial para inúmeras aplicações, desde extração de texto em documentos até automação de tarefas que envolvem dados textuais contidos em imagens.

**2.2 pdf2image ou fitz: Teoria e Funcionamento**

A biblioteca pdf2image em Python desempenha um papel fundamental na conversão de páginas de documentos PDF em imagens. Internamente, ela aproveita bibliotecas Python como PyMuPDF (também conhecida como PyMuPDF), proporcionando acesso ao conteúdo dos arquivos PDF. Ao receber um arquivo PDF como entrada, a pdf2image utiliza o PyMuPDF para processar o documento, extraindo as páginas em formato de imagem e organizando-as em uma estrutura de dados adequada. Durante esse processo, cada página do PDF é transformada em uma imagem rasterizada, comumente em formato PNG ou JPEG, facilitando sua manipulação e processamento no ambiente Python.

Vale destacar que a pdf2image é especialmente popular em sistemas Linux, onde sua utilização é mais comum. Em contrapartida, no ambiente Windows, a biblioteca fitz é amplamente empregada para tarefas semelhantes.

**2.3 Funcionamento e execução**

Utilizar OCR para detectar e localizar textos no MATLAB é simples. Entretanto só funciona se o arquivo de entrada for uma imagem (jpg, png). Então para contornar esse problema utilizamos a biblioteca Python pdf2image para converter o arquivo de extensão .pdf para .jpg, e assim utilizar o OCR do MATLAB para extrair dados de um arquivo pdf.

**Texto

Descrição gerada automaticamente** Figura1 - Workflow

Para converter o arquivo .pdf em uma foto, foi desenvolvida uma função Python que recebe o caminho do arquivo a ser convertido, a página que deve ser convertida e o nome do arquivo final que será gerado.

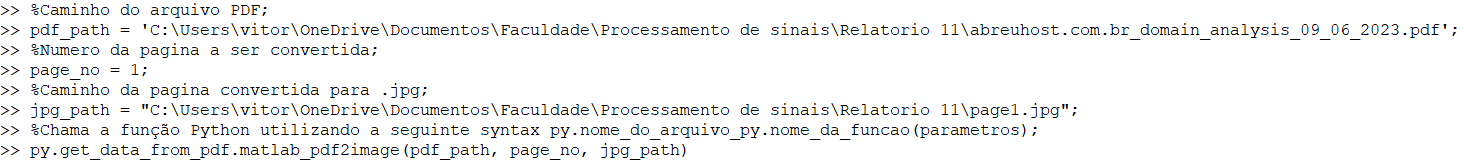
Figura2 – Função matlab\_pdf2image

Texto

Descrição gerada automaticamente

Com a função criada, é necessário integrar esse script com o MATLAB utilizando o MATLAB Api.

Figura3 – Chamada da função Python no matlab



Como o arquivo .pdf já foi convertido para .jpg, é possível utilizar o OCR do MATLAB para pesquisar palavras contidas na foto.

# 3 CONCLUSÃO

No decorrer deste projeto, foi possível alcançar significativos avanços na detecção e localização de texto em documentos PDF, superando a limitação inicial do MATLAB ao utilizar a biblioteca pdf2image em Python. A colaboração interplataforma revelou-se eficaz, permitindo a integração harmoniosa entre as funcionalidades do MATLAB e Python. Os resultados obtidos demonstraram a aplicabilidade prática dessa abordagem, evidenciando a precisão e eficiência na extração de texto de arquivos PDF.

A experiência adquirida ao superar esses obstáculos contribuiu significativamente para a formação, aprimorando habilidades em solucionar problemas complexos e reforçando a importância da colaboração entre linguagens de programação distintas.REFERÊNCIAS

Kevin Chng (2023). How to detect & localize a text in pdf using OCR in MATLAB (https://www.mathworks.com/matlabcentral/fileexchange/72156-how-to-detect-localize-a-text-in-pdf-using-ocr-in-matlab), MATLAB Central File Exchange. Retrieved November 6, 2023.

MathWorks. "OCR (Text Detection)." MathWorks, 2023. (<https://www.mathworks.com/help/vision/ref/ocr.html>), MathWorks. Retrieved November 6, 2023

PyPI. "pdf2image." PyPI, 2023. Web. [06/11/2023]. (<https://pypi.org/project/pdf2image/>), PyPI. Retrieved November 4, 2023