Sumário

[**Introdução** 1](#_Toc69565665)

[**Referencial Teórico** 1](#_Toc69565666)

**Resumo:**

A cada ano que se passa os avanços tecnológicos substituem mais e mais a necessidade de mão de obra humana. Tarefas que outrora nos pareciam impossíveis de se ter um computador executando, hoje são uma realidade em diversos setores da indústria.

Nosso objetivo de pesquisa é extrair o texto escrito a mão em português de imagens, utilizando para isso um OCR (Optical Character Recognition) e o conceito de aprendizado de máquina para treinar o OCR e reconhecer com precisão a escrita dos integrantes do grupo.

# **Introdução**

# **Referencial Teórico**

**Tesseract**

O Tesseract é uma ferramenta de código aberto desenvolvido pela HP, Hewllet Packard, entre 1984 e 1994. Começou como um projeto de pesquisa e ganhou força como um possível complemento ao hardware e software da linha de scanners, visto que a maioria das ferramentas OCR da época estavam apenas iniciando e apresentavam muitas falhas. Nesse projeto, foram priorizadas suas capacidades de rejeição ao invés da melhora de sua precisão. E em 2005, a HP tornou o Tesseract em código aberto.(SMITH, Ray, 2007)

**Arquitetura do Tesseract**

O Tesseract utiliza, como entrada, imagens binárias com regiões definidas de textos poligonais opcionais. O processamento segue o padrão passo a passo, embora alguma das etapas tenham sido incomuns para época. O primeiro passo é a análise de componentes conectados em que esboços do mesmo serão guardados, assim é possível detectar textos invertidos e reconhece-los facilmente como um texto preto no branco. E, foi a primeira ferramenta OCR a tratar de textos branco no preto. Nessa fase, os esboços são agrupados, por enquadramento, em *Blobs.*(SMITH, Ray, 2007)

Os Blobs são organizados em linhas de texto, e as linhas e regiões são analisados para um argumento definido ou um texto proporcional. As linhas de texto são quebradas em palavras através das diferenças de espaçamento de caracteres. Textos de argumentos definidos são então separados em caracteres. O texto proporcional é separado em palavras utilizando espaços definidos e espaços incertos. O reconhecimento passa por dois passos, no primeiro as palavras são reconhecidas em cada turno. Cada palavra, considerada satisfatória, é então passada para um classificador adaptativo como dados treinados. Assim, o classificador adaptativo tem maiores chances de reconhecer melhor textos ao decorrer da página. Como essa melhora no classificador adaptativo pode ocorrer no final do texto, uma segunda checagem no texto é feita, onde palavras não bem reconhecidas são trabalhadas novamente. A fase final toma conta de espaços incertos e checa hipóteses alternativas para localizar textos em minúsculo.(SMITH, Ray, 2007)

Para trabalhar com as linhas, as partes principais são filtragens dos blobs e a construção de linhas. Após as linhas serem encontradas, as bases das linhas serão definidas para que cada uma das linhas do texto sejam separadas. Para textos que não possuem um espaçamento bem definidos, o Tesseract soluciona esses problemas medindo o espaço em um alcance vertical limitado entre a base da linha e a média da linha. Espaços que estão pertos do limite se tornam difusos, para que uma decisão final seja feita após o reconhecimento da palavra. Se o resultado do reconhecimento de uma palavra não é satisfatório, as palavras são cortadas através de pontos em vértices côncavos. Se mesmo, depois de vários cortes, uma palavra ainda não ser satisfatória, ela é passada para o associador. O associador começa uma procura de combinações dos blobs, do melhor para a pior, com possíveis caracteres. Assim, esse esquema de corte e associação, simplifica a estrutura de dados que são necessárias para manter um gráfico de segmentação completa. (SMITH, Ray, 2007)

O classificador do Tesseract utiliza de dados treinados para que consiga fazer a classificação e possíveis correspondências de caracteres. A utilização de um classificador adaptativo é recomendada para identificar diferentes tipos de fonte. A normalização de base de linha e altura x faz com que a distinção de caracteres em caixa alta e baixa sejam facilitadas. (SMITH, Ray, 2007)

Com esses atributos, o Tesseract se tornou líder do mercado em termos de precisão, tendo as suas escolhas incomuns de atributos como uma vantagem. (SMITH, Ray, 2007)

**Reconhecimento de línguas pelo Tesseract**

O Tesseract tinha sido desenvolvido para reconhecer apenas a língua inglesa, mas passaram a suportar outras línguas a partir de 2007.(SMITH, Ray W., 2013)

O Sistema de OCR para uma língua especifica não contribui muito para utilizar o OCR em outras línguas do mundo, já que há muitas em uso antes mesmo de incluir variantes históricas. (SMITH, Ray W., 2013)

O primeiro passo de decisão de um projeto para desenvolver um sistema OCR multilinguístico é a representação interna de Recognition Units (RU),ou, unidades de reconhecimentos. O RU é um formato individual que um mecanismo de OCR reconhece, é utilizado para generalização do alfabeto, conjunto de caracteres e grafemas, em que cada um já tem um significado específico. (SMITH, Ray W., 2013)

O Tesseract é regularmente testado em 30 línguas, utilizando um conjunto de testes criado de vários modos, como no caso das línguas baseadas no latim e o russo que utilizaram dados criados a partir da digitalização de livros e camadas de texto de PDF. (SMITH, Ray W., 2013)

SMITH, Ray. An Overview of the Tesseract OCR Engine. 2., 2007. **Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)** [...]. Curitiba, Brasil: IEEE, 2007. v. 2, p. 629–633. DOI 10.1109/ICDAR.2007.4376991. Available at: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf.

SMITH, Ray W. History of the Tesseract OCR engine: what worked and what didn’t. **Document Recognition and Retrieval XX**, v. 8658, n. February 2013, p. 865802, 2013. https://doi.org/10.1117/12.2010051.