Sumário

[**Introdução** 1](#_Toc69565665)

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

**Resumo:**

# **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, 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, 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, 2007)

SMITH, R. 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.