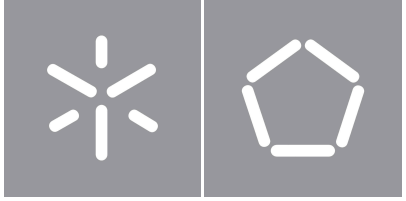
Universidade do Minho
Escola de Engenharia

Gonçalo Braz Afonso

OCR para documentos estruturados antigos
Old structured documents OCR

janeiro 2024



Universidade do Minho
Escola de Engenharia

Gonçalo Braz Afonso

OCR para documentos estruturados antigos
Old structured documents OCR

Dissertação de Mestrado
Mestrado em Engenharia Informática

Trabalho efetuado sob a orientação de
José João Antunes Guimarães Dias Almeida

janeiro 2024

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho:

[Caso o autor pretenda usar uma das licenças Creative Commons, deve escolher e deixar apenas um dos seguintes ícones e respetivo lettering e URL, eliminando o texto em itálico que se lhe segue. Contudo, é possível optar por outro tipo de licença, devendo, nesse caso, ser incluída a informação necessária adaptando devidamente esta minuta]

CC BY

https://creativecommons.org/licenses/by/4.0/ *[Esta licença permite que outros distribuam, remixem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito pela criação original. É a licença mais flexível de todas as licenças disponíveis. É recomendada para maximizar a disseminação e uso dos materiais licenciados.]*

CC BY-SA

https://creativecommons.org/licenses/by-sa/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito e que licenciem as novas criações ao abrigo de termos idênticos. Esta licença costuma ser comparada com as licenças de software livre e de código aberto «copyleft». Todos os trabalhos novos baseados no seu terão a mesma licença, portanto quaisquer trabalhos derivados também permitirão o uso comercial. Esta é a licença usada pela Wikipédia e é recomendada para materiais que seriam beneficiados com a incorporação de conteúdos da Wikipédia e de outros projetos com licenciamento semelhante.]*

CC BY-ND

https://creativecommons.org/licenses/by-nd/4.0/ *[Esta licença permite que outras pessoas usem o seu trabalho para qualquer fim, incluindo para fins comerciais. Contudo, o trabalho, na forma adaptada, não poderá ser partilhado com outras pessoas e têm que lhe ser atribuídos os devidos créditos.]*

CC BY-NC

https://creativecommons.org/licenses/by-nc/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho para fins não comerciais, e embora os novos trabalhos tenham de lhe atribuir o devido crédito e não possam ser usados para fins comerciais, eles não têm de licenciar esses trabalhos derivados ao abrigo dos mesmos termos.]*

CC BY-NC-SA

https://creativecommons.org/licenses/by-nc-sa/4.0/ *[Esta licença permite que outros remisturem, adaptem e criem a partir do seu trabalho para fins não comerciais, desde que lhe atribuam a si o devido crédito e que licenciem as novas criações ao abrigo de termos idênticos.]*

CC BY-NC-ND

https://creativecommons.org/licenses/by-nc-nd/4.0/ *[Esta é a mais restritiva das nossas seis licenças principais, só permitindo que outros façam download dos seus trabalhos e os comparti-*

Agradecimentos

Escreva aqui os seus agradecimentos. Não se esqueça de mencionar, caso seja esse o caso, os projetos e bolsas dos quais se beneficiou enquanto fazia a sua investigação. Pergunte ao seu orientador sobre o formato específico a ser usado. (As agências de financiamento são bastante rigorosas quanto a isso.)

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, Braga, janeiro 2024

Gonçalo Braz Afonso

Resumo

A digitalização de documentos permitiu uma nova forma de salvaguardar informação para a prosperidade, evitando a sua perda para o deterioramento físico destes. De forma a posteriormente transcrever estes documentos para permitir uma consulta, procura, processamento e manipulação mais simples destes o uso de software de **OCR** é essencial. Esta tecnologia é no entanto dependente, em diferentes níveis, das características do seu alvo, nomeadamente: qualidade da imagem, complexidade da estrutura do documento, linguagem do texto. Documentos mais antigos, em especial jornais por apresentarem estruturas mais complexas, apresentam por este motivo resultados que diferem bastante do seu conteúdo original; tanto a nível do texto reconhecido, como da sua organização para os diferentes outputs disponíveis (ex.: txt simples).

A tarefa de extrair informação destes documentos, como por exemplo o isolamento e extração de artigos, torna-se numa tarefa complexa e propícia a erros. Este trabalho pretende então a criação de uma ferramenta ou um conjunto de ferramentas que permitam auxiliar o processo de extração de conteúdo de documentos, primeiramente mas não exclusivamente, mais antigos e de estruturados, com especial foco em jornais. A pipeline do projeto pretende então ser capaz de detetar e lidar com os diferentes pontos de risco nestes documentos: qualidade da imagem, erros nos resultados de **OCR**, segmentação e organização do documento, criação do output organizado.

Diferentes alternativas para **OCR** assim como métodos de tratamento destes problemas serão estudados, comparados e implementados de forma a encontrar a melhor solução para a resolução deste problema. O produto final implementado será composto por uma ferramenta utilizável num formato de um **GUI** simples ou comando de consola.

Para documentos antigos a linguagem, como mencionado, pode afetar os resultados de **OCR**. Deste modo, como objetivo secundário, propõe-se a criação de uma ferramenta que facilite a criação de um dicionário para diferentes iterações de uma linguagem para este ser posteriormente fornecido ao motor **OCR**.

Palavras-chave OCR, Digitalização, Documentos estruturados, Documentos antigos, Segmentação de documentos, Tratamento de imagem, Modernização de texto

Abstract

Write abstract here (en)

Keywords OCR, Digitalization, Structured documents, Old documents, Document segmentation, Image treatment, Text modernization

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

EA Estado da Arte.

GUI graphic user interface.

OCR reconhecimento óptico de caracteres.

RPD Relatório de Pré-Dissertação.

Parte I

Material Introdutório

Capítulo 1

Introdução

1.1 Enquadramento e motivação

A digitalização tem um papel fundamental na conservação, disponibilização e proliferação de documentos físicos, não só contemporâneas, como de eras anteriores à revolução da informação. Esta tecnologia, acoplada a ferramentas de **OCR**, veio trazer uma facilidade de navegação, consulta e manipulação destes documentos que anteriormente não era possível.

A eficácia de **OCR** é no entanto dependente de vários fatores nas imagens ou ficheiros alvo: a qualidade das imagens, como a resolução, estado do documento, coloração, qualidade/tipo de escrita; a estrutura dos documentos - quanto mais complexo, mais difícil é obter a informação de forma automática mantendo a congruência original; linguagem do texto, sendo que por vezes diferentes tecnologias, como por exemplo **Tesseract**, procuram verificar a sua confiança na deteção com o vocabulário conhecido, o qual pode no entanto não coincidir com a época de produção do documento; entre outras.

Estas dependências são especialmente notórias quando se envolvem documentos mais antigos, os quais podem, além de apresentar envelhecimento causado pelo tempo e danos pelas condições de armazenamento, devido às limitações tecnológicas assim como por vezes à falta de convenções de formatação dos documentos, não dispor de uma consistência no formato e texto (estrutura, alinhamento, dimensões dos caracteres, fonte de texto consistente, etc.) usual nos documentos atuais. Estes fatores, resultam então num reconhecimento de texto não tão satisfatórios como se esperaria.

Estes documentos antigos são mais comummente, mas não exclusivamente, reconhecidos como anteriores à era da digitalização, sendo que o foco de trabalho será maioritariamente dirigido a documentos desta época, como jornais, revistas e outros, do século passado ou anteriores.

Em especial documentos com estruturas complexas, como é o caso de jornais, onde é possível a segmentação em diferentes partes com conteúdo e propósito distinto e ao mesmo tempo uma ordem de leitura complexa i.e., não segue apenas regras simples de posição do conteúdo (texto da esquerda antes

do texto da direita e cima antes de baixo) mas que exige também noção das características e relação do conteúdo.

Mesmo para ficheiros do tipo **hOCR** ou **PDF**, que já passaram por um processo de reconhecimento de texto, a complexidade da estrutura dos documentos originais ou problemas nos elementos que contém o texto (como por exemplo elementos sobrepostos ou que se intersetam) dificultam a extração e interpretação do seu conteúdo, podendo ser facilmente perdida a lógica original.

Por estas razões, seria útil uma ferramenta que permita uma deteção e tratamento destes documentos de forma automática e de uso simples, permitindo um certo nível de configuração para adaptação entre tipos de documentos com características bem definidas e distintas.

O presente documento pretende então servir como um estudo dos desafios apresentados por estes tipos de documentos perante **OCR**, assim como a procura de soluções para a melhoria dos resultados na deteção e extração de texto e assim criar uma ferramenta que torne o processo de extração de informação destes tipos de documentos mais simples e fiável.

Como trabalho complementar, é proposta a implementação de um método de modernização do conteúdo extraído, envolvendo a criação de uma ferramenta capaz de criar dicionários entre diferentes iterações de uma mesma linguagem.

1.2 Objetivos

O principal objetivo deste trabalho é a realização de um estudo sobre os problemas apresentados à extração de conteúdo de documentos de estrutura complexa - mantendo a sua lógica original -, assim como a implementação de uma solução para resolver ou mitigar estes desafios, aumentando a confiança na informação extraída.

Especificando, os objetivos são:

- Estudo sobre os diferentes softwares de **OCR** disponíveis e as diferenças entre estes.
- Estudo as dificuldades que documentos podem apresentar no processo de reconhecimento de texto.
- Estudar o trabalho desenvolvido sobre a área de tratamento de imagem, identificação de tipo de documento, segmentação de documentos, algoritmos de cálculo da ordem de leitura, melhoramento de resultados de OCR e métricas de validação de resultado OCR.
- Estudar trabalhos com âmbito similar ou relacionado ao presente.

- Implementação de um conjunto de ferramentas dirigidas à solução dos problemas propostos.
- Implementação de uma ferramenta em formato **GUI** e comando de consola que aplique uma pipeline cujo input seria um ficheiro - imagem, pdf, hOCR -, identifique e trate de problemas deste se necessário para melhorar os resultados de OCR, e por fim devolva um output que mantenha a lógica e conteúdo do documento original.
- Secundário : ferramenta para criação de dicionário de linguagem para modernização de documentos. Ferramenta tem como input duas versões de um documento na mesma linguagem mas iterações diferentes e dá como output um dicionário entre as versões.

  Estudo sobre alinhamento de documentos.

1.3 Estrutura da dissertação

Esta dissertação segue a seguinte estrutura:

- Capítulo 1: Breve contextualização sobre o tema proposto, as dificuldade impostas por documentos estruturados e com digitalizações ou condições físicas degradadas nos resultados **OCR** e a utilidade de uma ferramenta para o tratamento destas. Além disso foram listados os objetivos do trabalho.
- Capítulo 2: Estudo sobre o estado da arte nos tópicos relacionados ao tema da dissertação, as suas dificuldades e soluções destas; estudo de trabalho anteriormente realizado com âmbito similar ao atual.
- Capítulo 3: Listagem dos diferentes problemas que a solução irá abranger e os desafios que estes apresentam.
- Capítulo 4: Descrição da solução implementada, a sua arquitetura, componentes e características.
- Capítulo 5: Apresentação e estudo dos resultados da solução implementada.
- Capítulo 6: Reflexão sobre o trabalho realizado, os resultados e a experiência obtida, assim como uma breve exploração de caminhos para trabalho futuro do projeto.
- Capítulo 7: No último capítulo é explicado o plano de desenvolvimento da dissertação.

Capítulo 2

Estado da arte

Estado da arte revisto; trabalho relacionado.

2.1 Reconhecimento ótico de caracteres

2.1.1 Introdução

2.1.2 Breve história

2.1.3 Evolução até aos dias de hoje

2.1.4 Processo OCR

2.1.5 Problemas

2.1.6 Tecnologias OCR

2.2 Digitalização de documentos

2.3 Trabalho relacionado

2.3.1 Tratamento para OCR

2.3.2 Identificação de imagens

2.3.3 Segmentação de documentos

2.3.4 Ordem de leitura

Capítulo 3

O problema e os seus desafios

O problema e os seus desafios

3.1 Imagens

Exemplo de inserção de uma imagem como texto exibido,

— dentro no texto, bla-bla

— ou em formato flutuante

Figura 1: Legenda

Parte II

Core da Dissertação

Capítulo 4

Contribuição

Principais resultados e as suas evidências cientificas.

4.1 Introdução

4.2 Sumário

Capítulo 5

Aplicações

Aplicação do resultado principal (exemplos e casos de estudo)

5.1 Introdução

5.2 Sumário

Capítulo 6

Conclusões e trabalho futuro

Conclusões e trabalho futuro.

6.1 Conclusões

6.2 Perspetiva de trabalho futuro

Capítulo 7

Planeamento

7.1 Atividades

Tarefa	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Background e **EA**	•	•	•							
Preparação do **RPD**		•	•	•						
Contribuição				•	•	•	•	•	•	
Escrita							•	•	•	•

Tabela 1: Plano de atividades.

Bibliografia

Parte III
Apêndices

Apêndice A
Trabalho de apoio

Resultados auxiliares.

Apêndice B
Detalhes dos resultados

Detalhes de resultados cuja extensão comprometeria a legibilidade do texto principal.

Apêndice C
Listings

Se for o caso.

Apêndice D
Ferramentas

(Se for o caso)

Utilizadores de LaTeX devem consultar TUG , o grupo de utilizadores TeX .

Coloque aqui informação sobre financiamento, projeto FCT, etc. em que o trabalho se enquadra. Deixe em branco caso contrário.