Integration of Computer Vision Models for Document Interpretation and Anonymisation

Graduate

Marc Havrilla

Introduction: Visual Document Understanding (VDU) models, combined with Optical Character Recognition (OCR) or OCR-free, offer businesses and institutions a great opportunity to digitalise their processes and improve workflows. The digitalisation is progressing. However, challenges like sufficient know-how to integrate VDU models, compliance with data protection regulations and identifying the processes, where VDU models offer the most significant benefit, have to be resolved.

The main goal of the work is to analyse and evaluate the practicality and appropriateness of available VDU models for processing of documents (e.g. PDF of scanned documents) and to demonstrate these in a Proof-of-Concept application. Even though some regulatory aspects, especially regarding anonymisation, are discussed in the work, the developed application does not aspire to be regulatory compliant.

During this work, two areas have been identified, where a tool to extract text from an image, identify relevant entities of personal information and anonymise these, is beneficial. First, the anonymisation of medical documents makes more data available for research and educational purposes. A second application is data leakage prevention, where detecting client data from screenshots would lower the risk of data breaches.

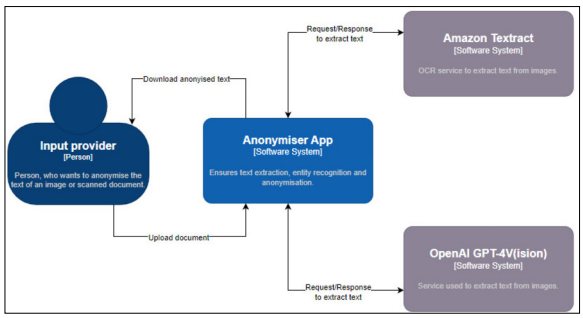
Approach / Technology: Various tools exist to extract text from an image. In the scope of this project, three tools have been integrated i.e., Tesseract, Amazon Textract and OpenAI GPT-4V(ison). The application extracts the text of uploaded documents or images and provides the user with the resulting text from all three tools. The user will be able to select the text with the best quality. Afterwards, a Named Entity Recognition (NER) Transformer model is used to identify the names of persons in the extracted text. The last step is the pseudonymisation of the entities: a randomly generated unique string replaces the entities in the text, so that a person cannot be identified based on the name in the text.

Another feature of the application is the evaluation of the OCR accuracy. The user is able to upload an additional ground truth file, which will then be compared with the output of the uploaded images. The OCR accuracy is implemented with string comparison algorithms. Furthermore, the NER model can also be tested by uploading the expected entities of the document in a separate file.

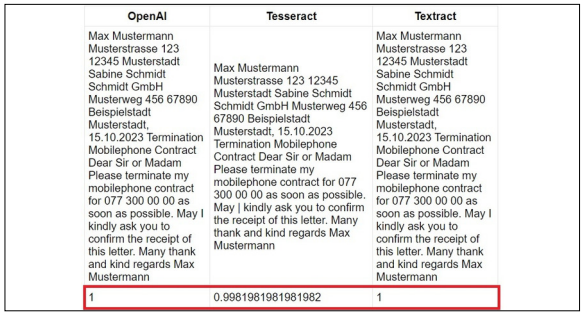
Conclusion: It is impressive how powerful today's text extraction and NER models have become. However, during the work, it was recognised that they are not yet off-the-shelf and just ready to use. Neither works each tool perfectly, so errors are propagated to subsequent processes, nor are the outputs of each tool standardised. To overcome such limitations, the process of text extraction and entity recognition should be executed by one model, which is also fine-tuned on the specific document types.

Example of an anonymised text.
Own presentment

C4 Context diagram of the Anonymiser App.
Own presentment



The screenshot shows the OCR accuracy of the three text extration tool for a simple testcase.
Own presentment



OpenAI	Tesseract	Textract
Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann	Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann	Max Mustermann Musterstrasse 123 12345 Musterstadt Sabine Schmidt Schmidt GmbH Musterweg 456 67890 Beispielstadt Musterstadt, 15.10.2023 Termination Mobilephone Contract Dear Sir or Madam Please terminate my mobilephone contract for 077 300 00 00 as soon as possible. May I kindly ask you to confirm the receipt of this letter. Many thank and kind regards Max Mustermann
1	0.9981981981981982	1

Advisor
Prof. Dr. Marco Lehmann

Co-Examiner
Dr. Johanni Brea, EPFL, Lausanne, VD

Subject Area
Miscellaneous, Software