GOETHE UNIVERSITÄT
FRANKFURT AM MAIN

Institut für Informatik, Goethe-Universität Frankfurt am Main

AISEL: AI Systems Engineering Lab
Wintersemester 2020/2021

Supervisors:
Dr. Michael Rammensee & Prof. Visvanathan Ramesh

OCR for Customs declaration and Transport Documents

Pascal Fischer, Alen Smajic

Frankfurt am Main
December 4, 2020

Contents

1 User View

1.1 Context

In 2016, a historic referendum took place in the UK, which is better known as the BREXIT. Four years later, this event has also reached the industry and affects the transport sector. Numerous European logistics companies have to adapt to the new import and export regularities and are confronted whit a lot of paperwork. Because of that the number of customs declarations and transport documents that a company has to deal with is growing rapidly. Our goal is to develop a modern intelligent system which automatically processes a huge amount of customs and transport documents to increase the performance and efficiency inside the logistic sector while reducing the possibility of human errors at the same time.
In practice, all customs related documents are scanned by the customer and submitted to the transporting companies. These documents can have different structure and be filled with all kinds of different information. The same document type can differ from customer to customer and the overall structure of the document is often not consistent as well as the language itself. There is also a variety of different objects which can occur on such documents such as images, tables, text etc. Another important aspect which has to be considered is the physical document itself. It occurs that these can also be crumpled or contain stains which affects the quality of the digital document. The scanning device used to digitalize the document has also a huge influence on the quality of the image. Some scanners can create strange artefacts or poorly capture the document which results in a lower quality image. Lastly we have also to consider the human operating the scanner. They also have a pretty big impact on the image quality since they can directly manipulate the rotation, shift variance and the quality of the scanned image itself depending on how they use the scanning device.

1.2 Task

The goal of our project is to develop an AI-based system which is able to capture the overall structure of the scanned document and to extract useful information from the image. This information should be processed and exported in the form of a XML or JSON file for further processing capabilities. We want to emphasize the importance of capturing the structure of the document since this is a crucial task in order to develop a deep system understanding of the underlying information and its relationship to other information artefacts inside the document. The system has to be able to

extract data stored in tables and connect the information inside the cells in a logical way.

1.3 Performance

Since the system is considered to fully automate the processing of information captured by these documents and replace the human supervision, it is required that the system has a high accuracy. Our goal is to get a quantitative accuracy of at least 80% while also developing a safe solution for artefacts which could not be recognised or processed by the system. The basic idea is, that the system reports on documents which could not be recognised or processed whit a high confidence, to be supervised by a human. Whit this in mind, our system should be able to provide as much help and automatization as possible, while maintaining the safety of operations.