ABSTRACT

Handling informative documents with agility is a key step for the insurance industry to face the ongoing digital transformation. More precisely Optical Character Recognition (OCR) technology is critical in this respect, as it enables automated document scanning, sorting, and analysis, and makes business workflows more efficient and effective.

In this work we present AXA OCR sequence-to-sequence system to address recognition of typical insurance documents such as accident cards or accident reports. This system can handle typed and handwritten characters. We compare performance with a standard baseline system 'Tesseract' [Smith 2007]. Note that this work is still on progress.

OVERALL SYSTEM

The overall document processing is composed of 4 steps:
- classify the documents
- detect automatically the zones of interest
- apply OCR inside the zone
- match information from OCR with business database

DATA DESCRIPTION

2 corpus are involved in this work:
- French identity cards: the system aims to recognize each of the following fields:
  - Accident reports: license plate are extracted to identify the 2 vehicles.
  - Fields can have important length variations, e.g. first name:

DATA AUGMENTATION

Due to the lack of data, to avoid over training and increase robustness, we used data augmentation to increase the size of the training set. We distinguish 2 types of operation:
- Random shifting:

FEATURE EXTRACTION

Features are extracted with a 3 layers CNN

ENCODER - DECODER

We used a sequence-to-sequence system inspired from [Wojna and Al, 2017]. The decoder part involves an attention mechanism, it allows to focus on a specific part of the previously recognized characters to help for next prediction.

RESULTS

We compared our system performance to Tesseract system [Smith 2007]. However Tesseract does not handle handwriting so the benchmark is available only for ID card processing.

FUTURE WORK

There are still some potential for improvement while addressing the following points:
- Results were given without any specific post-processing which may have lead to better performance. Post processing specific to the business use-case could be easily applied (dictionary to check/modify the output), rules on the output (license plate)...
- Current data augmentation could be improved while adding other type of image distortions
- Using transfer learning could ensure better convergence
- Compared to the degrees of freedom of our system, our dataset is rather small, increasing its size may help a lot while allowing to train a deeper model.
- More complete pre-processing can be performed while addressing background removal, distortion cancellation...

REFERENCES


Wojna and Al, 2017: Zdzisław Wojna, Alex Gorban, Dan-Shyeong Lee, Kevin Murphy, Qian Yu, Yejing Li, Julian Ibarz, 2017, Attention-based Extraction of Structured Information from Street View Imagery

Smith 2007: Ray Smith, An Overview of the Tesseract OCR Engine