Document Processing with Insurance Documents

Kevin SERRE, Dinojan KARTHIGESU, Simon BOZONNET

kevin.serre@axa.fr,
dinojan.karthigesu@axa.fr,
simon.bozonnet@axa.fr

DATA SCIENCE SUMMER SCHOOL

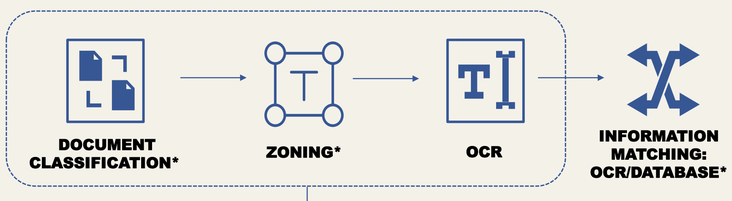
ABSTRACT

Handling informative documents with agility is a key step for the insurance industry to face the ongoing digital transformation. More precisely, the idea behind is to automate document onboarding, sorting analysis, and make business workflows more efficient and effective.
This implies, recognition of some typical pages or documents in a pdf (classification), detecting region of interests in those pages, (object detection), reading information with a customized model (retrained OCR models), signature detection, crossed-out detection, …
In this work we present a solution built at *AXA* to handle a type of insurance document called *RSE* which is used for life insurance contractualization. The goal was to automate at least 80% of the incoming flow.

OVERALL SYSTEM

The overall document processing is composed of 4 steps:

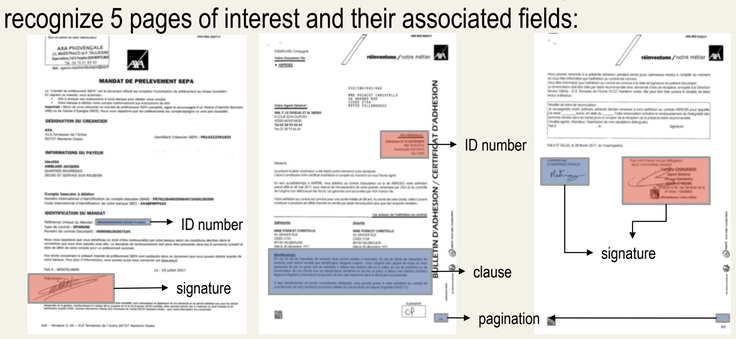
- **classify** the documents
- detect automatically the **region** of interest (ROI)
- apply **OCR**, **signature** or **crossed-out** detection
- **match** information from OCR with business database

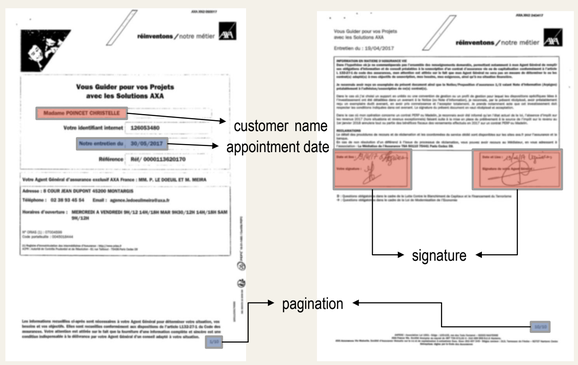


Only the first three parts are addressed in this work

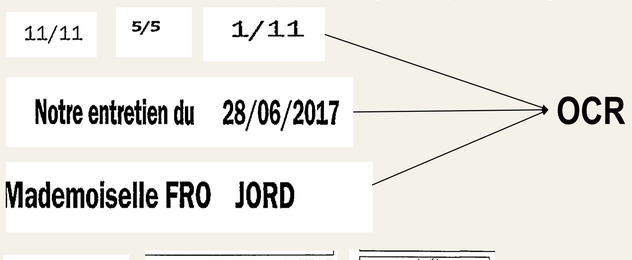
DATA DESCRIPTION

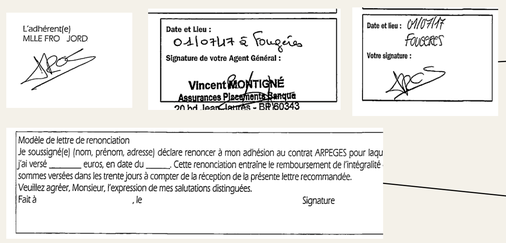
3 documents are involved in this work from which the system aims to recognize 5 pages of interest and their associated fields:



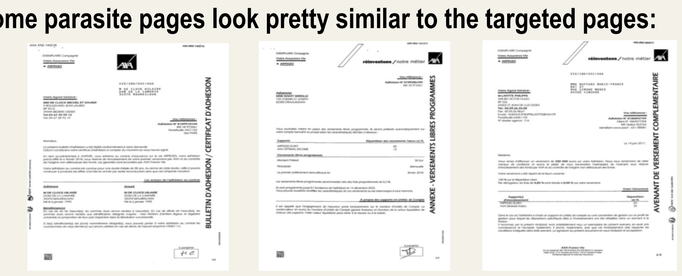


For each extracted field, we applied a specific processing:





Some parasite pages look pretty similar to the targeted pages:

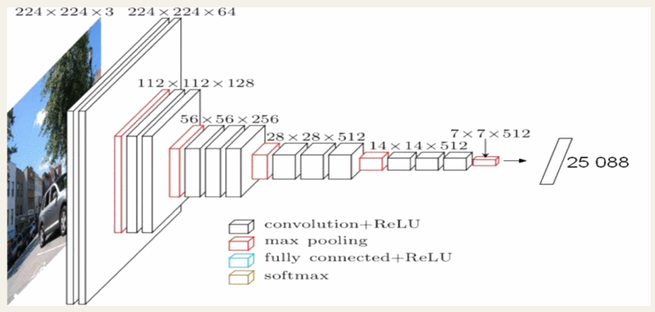


page of interest | Parasite 1 | Parasite 2

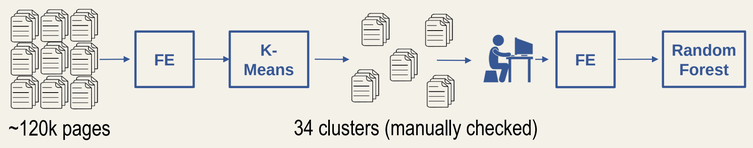
VOLUME OF DATA		
	Training set	**Testing set**
PDF	**5 750**	**2 465**
PAGES	**83 129**	**35 626**

FEATURE EXTRACTION (FE)

For each step of the pipeline, features were extracted with VGG-16 using the weights trained on *ImageNet*



DOCUMENT CLASSIFICATION



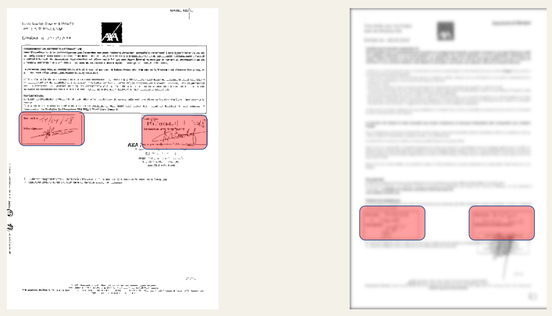
We used clustering to label our dataset (each cluster are then **checked manually**). We then trained a classifier with the following results in terms of precision and recall:

RESULTS		
	Precision	Recall
POI1	99%	99%
POI2	99%	98%
POI3	99%	99%
POI4	99%	99%
POI5	99%	98%
others	99%	99%

Aggregation of 29 classes

OBJECT DETECTION (OD)

Following the page classification step, we then needed to extract the regions of interests: e.g. signature, ID number, … (see data description). If those pages had always the same format, padding, text size, … we could have just cropped the areas easily because they would have always been the same for each type of document. Unfortunately, we had **intra-class variations**. For example, the text size :

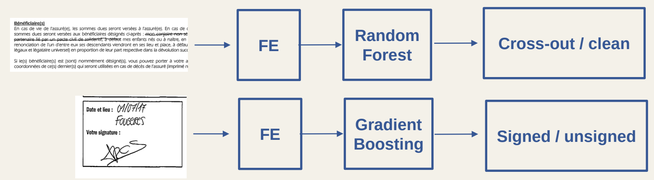


This figure illustrates two pages from the same class but where the text size variation moved the signatures to different locations.
Thus, we have to use an **object detection model** to detect those areas.
The model used was a **Faster-RCNN**

PERFORMANCE Average Precision and avg IOU		
	AP*	AIOU*
Contract ID number	0.99	0.94
Agent signature	0.93	0.88
Customer signature	0.99	0.96
Appointment date	0.91	0.85
Sepa signature	0.90	0.93
Sepa ID number	0.99	0.95
Clause	0.93	0.88

AP : Average Precision AIOU : Average Intersection Over Union

SIGNATURE AND CROSS-OUT DETECTION



Models are tuned in order to detect 100% of unsigned cases or crossed-out clause even if we wrongly predict some cases as unsigned or crossed-out.

RESULTS		
	Precision	Recall
Signed - agent	100%	99.9%
Unsigned - agent	95%	100%
Signed – customer	100%	99%
Unsigned -customer	74%	100%
Crossed-out Clause	98%	97%
Clause	99%	99%

OCR

To enhance OCR quality, **we finetuned *Tesseract* V4 'fra_best'** model for each region of interest. By retraining, we try to learn a model which will be more specific to the l**inguistic style** and to the **font**. We provide the following results in terms of *character error rate* (CER) and *word error rate* (WER).

RESULTS		
	CER	WER
NAMES	1.93%	4.05%
APPOINTMENT DATE	0.19%	1.5%
PAGINATION	0.81%	3.3%
SEPA MANDAT	0.04%	0.15%
CONTRACT ID NUMBER	0.03%	0.15%

CONCLUSION

This work is a first attempt to fully automate the contractualization process in the insurance field. The idea was to use object detection, classification models, OCR to check completeness, detection of region of interests, … on specific insurance documents which was done manually before. Today, this work is in production at *AXA France* and **automates 82% of the incoming flow**.

FUTURE WORK

There are still some potential for improvement while addressing the following points:

- **Post-processing** could outperform current results, e.g.:
  - business specific post processing dictionary to check/modify the output,
  - rules on the output (ID Number)
- **Data augmentation** could improve results by adding some types of image noises or distortions (color perturbation, rotation…)
- **Finetuning** standard bottleneck features should help for classification, signature and crossed-out detections.
- **Object detection** model was made using an old implementation. Retraining another model object detection model with a more recent implementation could be interesting.
- **Pre-processing** can be improved while addressing background removal, distortion cancellation…

REFERENCES

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

Itzeez 2015: Open Source Computer Vision Library, Itseez, https://github.com/itseez/opencv

Shaoqing 2015: Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*

Rbgirshick2015: Py-Faster-RCNN, Rbirshick, https://github.com/rbgirshick/py-faster-rcnn