Text detection and post-OCR correction in Engineering Documents

Mathieu FRANCOIS, Université de Lyon, INSA-Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France, francois.mathieu@orinox.com
Véronique EGLIN , Université de Lyon, INSA-Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France, veronique.eglin@insa-lyon.fr
Maxime BIOU, Orinox, Vaulx-en-Velin, France, biou.maxime@orinox.com

Objectives

- Digital transformation of engineering documents
- Extraction of textual entities on large and unstructured documents
- Short text contextualisation and correct OCR predictions of Tags.

Related works

Post-OCR Correction

[ICADL2020] When to Use OCR Post-correction for Named Entity Recognition?, [ICDAR2019] A Cost Efficient Approach to Correct OCR Errors in Large Document Collections, [2019] Leveraging text repetitions and denoising autoencoders in OCR post-correction, [ACL2017] OpenNMT: Open-source toolkit for neural machine translation

Text Detection

[CVPR2017] EAST: An Efficient and Accurate Scene Texte Detector, [ICPR2020] DUET: Detection Utilizing Enhancement for Text in Scanned or Captured Documents, [CVPR2019] Character Region Awareness for Text Detection, [ICDAR2021] Context Free TextSpotter for Real-Time and Mobile End-to-End Text Detection and Recognition

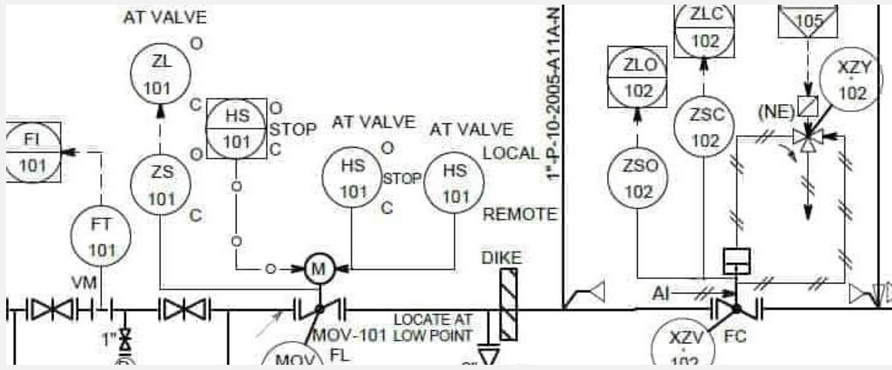
Contribution

Detection :

- Based on the FCN model EAST with pre-trained Resnet V1.
- NMS part is removed by our own filtering method.
- Adaptation of the system to the use case

Post-OCR Correction:

- Clustering of Tags using Affinity Propagation
- Identification of potential errors made by the OCR by analyzing the tag structure
- Proposal to correct misrecognized characters



Experimentations

- Industrial data are sensitive and cannot be shared, so we have tested on our own dataset
- Calculation of the Precision, Recall & F1 Score for the detection module
- The Affinity Propagation clusters are the tags of the plan “shot by shot”
- Difference between the WER of the OCR output and the post-OCR Correction : 7% (75% OCR–82% post-

Conclusion & future works

- Focus on graphic symbols to help the post-OCR correction
- Association of symbolic and textual entities
- Improve initial step of text recognition by OCR competition

P0	P1	OCR	Post-OCR
(15,200)	(36,212)	AA-2504-Xx	AA-2504-XX
(135,200)	(148,212)	AA-25Q7-XX	AA-2507-XX
(264,108)	(278,120)	AA-2513-XX	AA-2513-XX

	Precision	Recall	F1 Score
EAST (original)	0.818	0.619	0.705
Our Approach	0.824	0.863	0.843

Evaluation of the detection system with our own dataset