Call for Bachelor Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

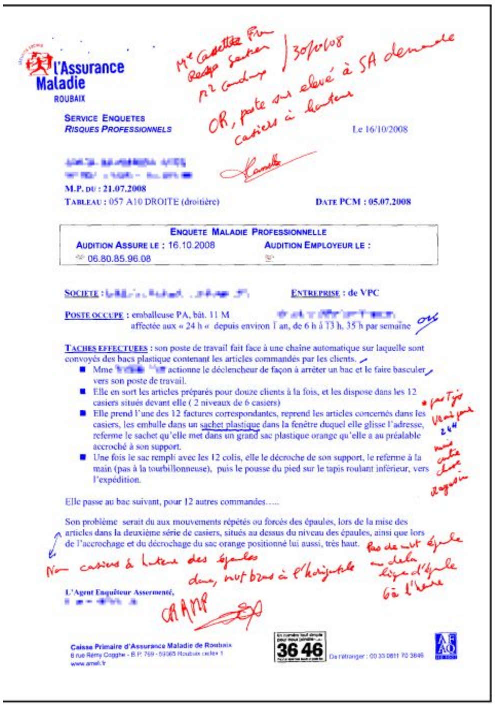
Handwritten and Printed Text Separation in Historical Documents

Objective of this work:

With the increase of digitized documents, automatic document analysis has become extremely important. The presentation of historical documents to the public introduces a variety of document types, content, quality and structure. Fundamentally speaking, documents can be skewed, noisy, and overlapped with graphics, i.e., lines, unconstrained annotations, stamps.
Most optical character recognition (OCR) systems recognize either printed or handwritten text. Hence, the task of the thesis is to separate machine printed text from handwritten text in scanned documents before feeding it to an OCR system.

In this thesis:

1. Documents containing a mix of handwritten and printed text will be collected.
2. An additional mixed dataset may be generated from historical documents.
3. The existing approaches of text separation will be reviewed and investigated.
4. A pixel-based approach for text separation based on [1] will be applied.
5. The results will be evaluated based on the ground truth data.



l'Assurance Maladie
ROUBAIX

SERVICE ENQUETES
RISQUES PROFESSIONNELS

Mme Casette Fr
Redap Santer
n2 Gardage | 30/10/08
OR, poste sur élevé à SA demande
casiers à hauteur

Le 16/10/2008

[redacted]
M.P. DU : 21.07.2008
TABLEAU : 057 A10 DROITE (droitière) DATE PCM : 05.07.2008

ENQUETE MALADIE PROFESSIONNELLE
AUDITION ASSURE LE : 16.10.2008 AUDITION EMPLOYEUR LE :
☎ 06.80.85.96.08 ☎

SOCIETE : [redacted] ENTREPRISE : de VPC

POSTE OCCUPE : emballeuse PA, bât. 11 M [redacted]
affectée aux « 24 h « depuis environ 1 an, de 6 h à 13 h, 35 h par semaine oui

TACHES EFFECTUEES : son poste de travail fait face à une chaine automatique sur laquelle sont convoyés des bacs plastique contenant les articles commandés par les clients.

- Mme [redacted] actionne le déclencheur de façon à arrêter un bac et le faire basculer vers son poste de travail.
- Elle en sort les articles préparés pour douze clients à la fois, et les dispose dans les 12 casiers situés devant elle (2 niveaux de 6 casiers)
- Elle prend l'une des 12 factures correspondantes, reprend les articles concernés dans les casiers, les emballe dans un sachet plastique dans la fenêtre duquel elle glisse l'adresse, referme le sachet qu'elle met dans un grand sac plastique orange qu'elle a au préalable accroché à son support.
- Une fois le sac rempli avec les 12 colis, elle le décroche de son support, le referme à la main (pas à la tourbillonneuse), puis le pousse du pied sur le tapis roulant inférieur, vers l'expédition.

• par Tijo Vend pour 24H mais cette chose dégagé

Elle passe au bac suivant, pour 12 autres commandes…..

Son problème serait du aux mouvements répétés ou forcés des épaules, lors de la mise des articles dans la deuxième série de casiers, situés au dessus du niveau des épaules, ainsi que lors de l'accrochage et du décrochage du sac orange positionné lui aussi, très haut.

Non casiers à hauteur des épaules
dans, mvt bras à l'horizontale

Pas de mvt épaule au delà ligne d'épaule 6 à l'heure

L'Agent Enquêteur Assermenté,
[redacted]

CRAMP

Caisse Primaire d'Assurance Maladie de Roubaix
8 rue Rémy Cogghe - B.P. 769 - 59065 Roubaix cedex 1
www.ameli.fr

36 46

De l'étranger : 00 33 0811 70 3646

AFAQ

[1] Dutly, N., Slimane, F., & Ingold, R. (2019, September). Phti-ws: A printed and handwritten text identification web service based on fcn and crf post-processing. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* (Vol. 2, pp. 20-25). IEEE.

The project work will be supervised by **Prof. Dr. Harald Sack, Mahsa Vafaie and Oleksandra Bruns**, **Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe.**

Keywords:
Machine Learning, CNN, pattern recognition
Pre-requisites:
Knowledge of Programming with Python.

Contact persons:
Mahsa Vafaie
mahsa.vafaie@kit.edu
Oleksandra Bruns
oleksandra.bruns@kit.edu