PHD OFFER

Campaign 2019

L3i laboratory

Only students, which are citizens of the European Union or Switzerland and which did not yet start their professional career, are eligible for this grant.

PhD topic:

Hybrid document authentication using their graphical and textual contents

Summary of proposed work:

The work of this PhD thesis will focus on the development of a hybrid approach image/text for authenticating documents. We propose to combine a new method for authenticating the template of the document through content-based hashing (image analysis approach) with a new method for verifying the consistency of the contents (text analysis approach).

Keywords:

Document security, fraud detection, document image analysis, text analysis (searching of textual data, natural language processing, information retrieval).

Additional information:

Supervisors: Petra Gomez-Krämer and Antoine Doucet
Place: Images and contents group, L3i laboratory, La Rochelle University, France
Starting date: From October 1st 2019
Duration of the contract: 3 years
Salary (gross): over 2 000€/month

To apply please send a mail to petra.gomez@univ-lr.fr and antoine.doucet@univ-lr.fr with a resume, a motivation letter, the grades of the two years of Master studies, and a report or significative work realized during the last two years. Applications will be accepted until the position is filled.

Context of the PhD:

La Rochelle is one of the most attractive and beautiful cities in France. It is situated in the south west of France on the Atlantic coast, near the Ré and Oléron islands. Within this wonderful environment, La Rochelle University is a dynamic and innovative university with the L3i as the computer science research laboratory. The L3i works since several years on document fraud detection and document security and has become a worldwide reference in this domain.

The more and more documents are dematerialized and processed in huge image document flows in companies, banks and administrations. Thus, the detection of frauds in these documents becomes a more and more important factor in those document flows. The fraud can be the intentional modification of a document (falsification) or the production of a fake document (counterfeit). Even if the more and more fraudulent documents are detected, a significant number stays undetected.

Currently, there exists no reliable solution to protect the companies from document fraud while several companies face the consequences and demonstrate a high interest for this type of solution: according to a study of the firm PricewaterhouseCoopers (PwC), 49 % of the responding companies declared to have been victim of fraud in the last two years, with respect to 36 % in 2016. The observed percentage of fraud in France reaches a record level: 71 % of French companies have declare to have been victim of the in the last two years.

As there is not reliable solution for document fraud detection, we wish to push our research further in this PhD thesis. More precisely, this PhD thesis aims at developing a new tool for the fraud detection (falsified and counterfeited documents) in document image flows. We propose to combine a new method for authenticating the template of the document through content-based hashing (image analysis approach) with a new method for verifying the consistency of the contents (text analysis approach). The work will be conducted in collaboration with an industrial partner and the DGA (an agency of the French ministry of defense).

Description of the PhD topic:

In spite of the important need, the automatic fraud detection in documents is few investigated. The passive approaches for the detection of fraudulent documents can be applied to whatever document type. The filtering out of fraudulent documents in document flow can be seen as classification task for which the best indices and image features have been found as well as the best classification algorithm. The difficulty of document flows is the heterogeneity of the documents. The documents can be acquired by a scanner or captured by a smartphone with different resolutions (between 150 and 600 dpi) and different color representations (black and white, gray level, color) with huge diversity of contents and layouts.

There exist very few methods in the state of the art pour the authentication or the recognition of document templates. The drawbacks of these methods are that they suppose the template of the document to be known or they conceived for document retrieval. For this reason, we propose to develop a new template authentication method using on content-based hashing. The main difficulty relies in the stability of the document image analysis algorithms. Moreover, a digest representing the template of the document has to be designed.

At the same time, there exist methods working on the textual content of the document. Very few works have been published on automatic information checking. Considering that the information to check are localized, the content checking reveals to main difficulties. On the one hand, the user does not possess systematically a reliable and verified version of the information in its data bases. On the other hand, Internet is a considerable source of information. One approach for confirming information without context can be to realize requests in search engines, combined with web scraping techniques. The linking between verified data and extracted data is not trivial as the latter one can be written in another form as is has been written in a different manner or it can be impacted by optical character recognition errors.

Until now, no hybrid approach has been proposed for document fraud detection. Hence, a method for fusing the information coming from the image and the text have to be verified.

Requirements and constraints:

The candidate should

- Hold a Master degree in computer science or equivalent with good skills in mathematics, image processing and/or natural language processing;
- Have good programming skills;
- Have excellent command of English (French language skills are irrelevant).

Bibliographiques :

[Artaud18] Artaud C., Doucet A., Ogier J.-M., Poulain d'Andecy V., Automatic Matching of Abbreviated Phrases and their Expansions without Context, CICLing 2018.

[Eskenazi15a] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier. The Delaunay document layout descriptor. In ACM International Symposium on Document Engineering (DocEng), 2015.

[Eskenazi17] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier. A perceptual image hashing algorithm for hybrid document security. In International Conference on Document Analysis and Recognition (ICDAR), 2017.

[Duthil14] B. Duthil, M. Coustaty, V. Courboulay, J.-M. Ogier, Annotation sémantique de documents administratifs. Revue des Nouvelles Technologies de l'Information, 2014; Extraction et Gestion des Connaissances, RNTI-E-26:47-52.