PROPOSITION OF POSTDOCTORAL POSITION

L3i laboratory
University of La Rochelle
France

Topic

Document analysis and description for hybrid document authentication

Summary of the proposed work

The recruited person will integrate the SHADES project within the L3i lab (University of La Rochelle in France). The objective of this project is to provide a new tool for authenticating the entirety of the content of a document through an advanced compact signature in order to fight against fraud and falsification. This signature is based on the document's content (text and graphics) and structure (spatial relationships) what we call a semantic signature**.** Thanks to a hashing of the document's information during the signature computation, no information from the original document will be deduced from its signature alone. The signature can then be inserted in the document or used in company content management software in order to check the authenticity of the document without compromising its confidentiality.

Currently, there is ongoing working on this subject. The objective of this postdoctoral position is to propose, based on this work, stable algorithms for document content analysis and description methods, as well as hashing.

Key words

Image/document processing, stability of image/document analysis algorithms, hashing, document authentication

Context

The L3i is a research lab in La Rochelle. La Rochelle is a city in the south west of France on the Atlantic coast and is one of the most attractive and dynamic cities in France.

The L3i works since several years on fraud detection in documents and document security and has become a worldwide reference in this domain. The SHADES (Semantic Hash for Advanced Document Electronic Signature) project is an interdisciplinary project on document authentication, financed by the French National Research Agency (ANR). It involves the company ITESOFT, the FNTC (a professional federation representing the trusted third parties at national and international level), and two research labs in computer science (the L3i of University of La Rochelle and the LIPADE of the Paris Descartes University) and one in law (the CEJEP of the University of La Rochelle). The recruited person will work in strong collaboration with the LIPADE and ITESOFT, but will also interact with the other partners.

Description of the subject

Many documents need to be secured, ideally by the means of an electronic signature. Typically, the electronic signature is obtained by computing a hash code on the document's pixels values. If two documents have the same signature, then they are authentic copies of each other and if their signatures are different, one of the two documents is fraudulent or at least different from the other one. This concept works well for naturally born digital documents. However, nowadays a document, the so-called hybrid document, is often used in electronic or paper form according to the need. Hence, the hybrid document undergoes a life cycle of printing and scanning and thus different degraded versions of the document exist as the printing and scanning process introduces specific degradations, such as print and scan noise, in the document. Thus, the concept of electronic signature cannot be applied. For this reason, our work intends to develop an advanced

electronic signature for the field of securing hybrid documents, the so-called hybrid security. Our idea is to extract the layout, the text and the images from the document, to describe the page in a stable manner, and to compute a hash that will be the same for all the authentic copies of the document. In consequence, this requires document analysis techniques with an extreme stability especially with regard to print and scan noise.

Many document analysis algorithms have been evaluated with respect to accuracy. Anyhow, the concept of accuracy does not apply to our security context. The two concepts of accuracy and stability should not be confused. Accuracy requires a ground truth to evaluate how close a result is to this ground truth. Accuracy can be evaluated with only one result as long as there is also a ground truth. Stability does not require a ground truth. Stability requires at least two results with similar inputs to see how close these results are together compared to how close the inputs were. In our case, similar inputs are two photocopies of the same document. A consequence of this is that an algorithm can be very stable and yet not be accurate. For instance, this can be an algorithm that always makes the same mistakes or for instance in the case of a segmentation algorithm, an algorithm producing always one region covering the whole image. Such an algorithm would have an absolute stability and zero accuracy. The contrary is not true. An algorithm with an absolute accuracy will always produce results that are identical to the ground truth and hence will be identical between each other. Furthermore, stability should also not be confused with robustness. A robust algorithm is an algorithm capable of providing a relevant output even when a certain amount of noise is contained in its input. The variation of this output is not constrained as it is for stability.

Our recent work has shown that traditional document analysis algorithms such as optical character recognition (OCR) and segmentation algorithms are unstable [1,4] as they contain thresholds and parameters [3]. Our first approach of developing a layout descriptor, without using thresholds and parameters, shows that stable algorithms can be achieved. The objective of this postdoctoral position is continue the ongoing work and to develop stable document analysis algorithms and description methods.

Profile

The applicant should have a completed PhD in computer science, signal processing or applied mathematics. The ideal candidate will have a strong background in image or document analysis. Good programming skills are required. The recruited person will be involved in the management of the project, participate to consortium meetings and contribute to deliverables. Therefore, good communication skills and autonomy are mandatory.

How to apply

The application should include a brief description of research interests and past experience, a CV, degrees and grades, motivation letter, relevant publications, letter(s) of recommendation and contact information of reference persons.

Details

Starting date: preferably between September and December 2016

Duration: 18 months

Salary: approximately 2300 € net per month

Contacts

Petra Gomez-Krämer: petra.gomez@univ-lr.fr

Jean-Marc Ogier: jean-marc.ogier@univ-lr.fr

References

[1] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier. Evaluation of the stability of four document segmentation algorithms. In *International Workshop on Document Analysis Systems (DAS)*, 2016.

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

[3] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier. Let's be done with thresholds. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015.

[4] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier. When document security brings new challenges to document analysis. In *International Workshop on Computational Forensics (IWCF)*, Lecture Notes in Computer Science (LNCS 8915), pages 104-116. Springer, 2015.