PROPOSITION DE STAGE MASTER 2

Année 2016

Laboratoire L3i

Sujet de stage :

Amélioration des algorithmes de hachage de document

Résumé du travail proposé :

Ce stage s'intègre dans les travaux du projet SHADES au sein du laboratoire L3i en partenariat avec l'entreprise ITESOFT, la FNTC, et deux autres laboratoires français sur la protection de documents administratifs. L'objectif de ce projet est de fournir un nouvel outil permettant l'authentification de l'intégrité du contenu d'un document quelle que soit sa forme (numérique, numérisé, faxé, etc.) par le biais du calcul d'une signature robuste et compacte afin de lutter contre la fraude et la falsification. Cette signature sera basée sur le contenu (textuel et graphique) du document et prendra également en considération la structure interne sous-jacente aux éléments de base composant ce document (relations spatiales). Grâce à un hachage de l'information du document lors du calcul de cette signature, aucune information du document original ne pourra être déduite de sa seule signature. La signature pourra alors être insérée dans le document ou utilisée dans un logiciel de gestion de contenu d'entreprise afin de vérifier l'authenticité du document, sans toutefois compromettre sa confidentialité.

Des travaux sur l'analyse du document et le calcul de la signature sont actuellement en cours. Le but du stage sera de compléter ces travaux, en particulier l'analyse et hachage des éléments graphiques d'un document et la stabilité des outils de reconnaissance du texte.

Mots clés :

Sécurité, hachage, stabilité des algorithmes, analyse de texte, analyse d'image

Informations complémentaires :

Encadrant(s) : Petra Gomez-Krämer, Sébastien Eskenazi, Jean-Marc Ogier

Thématiques :

X Analyse et gestions de contenus

Domaine d'application :

X Pertinence – contenu – interactions

Cadre de coopération :

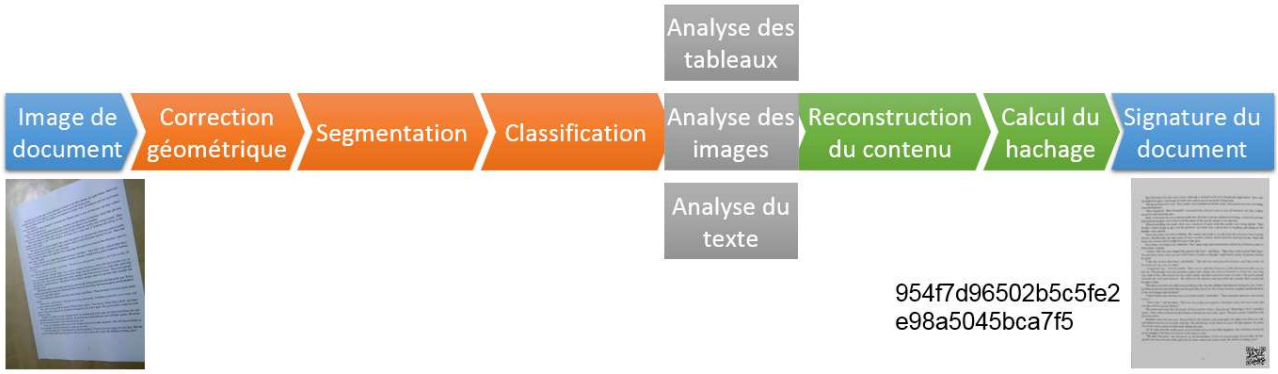
Date de début du stage : mars

Durée du stage : 5 mois minimum

Financement : Projet ANR SHADES

Contexte de l'étude:

L'objectif du projet est de proposer des méthodes valables indépendamment de la forme du document (numérique, numérisé, faxé, etc.). Par contre, le processus d'impression et de numérisation introduit du bruit et des déformations dans le document. Donc, un hachage cryptographique calculé au niveau des valeurs des pixels ne peut pas être utilisé pour sécuriser le document car la signature résultante ne sera pas la même pour deux versions du même document (par exemple le document original numérique et le document imprimé et numérisé). Par contre, en cas de modification frauduleuse du document (par exemple la modification de la date, d'un montant ou du logo) la signature calculée doit différer de celle obtenue à partir du document original. La difficulté réside dans la stabilité des algorithmes proposés face aux déformations d'impression et de numérisation, mais qui doivent être suffisamment précis pour produire une signature différente en cas de modification frauduleuse du document.
Le calcul de la signature est décrit dans la figure ci-dessous :



Le sujet du stage s'intégrera principalement dans les tâches « Analyse du texte » et « Analyse des images ». Il s'agit d'évaluer des travaux déjà effectués, de les améliorer ou de proposer des méthodes plus pertinentes.

Description du sujet :

Les travaux actuels portent sur l'utilisation de la corrélation dans le processus de hachage de logos et de signatures manuscrites. Ces travaux ont été partiellement évalués, mais seront à compléter par une approche de cross-validation. Des techniques d'analyse d'image fréquentielles et de techniques de hachage flou seront à étudier si la méthode existante s'avère insuffisamment stable.

L'analyse de texte par des outils de reconnaissance de caractères (OCR), quant à elle, n'est pas assez performante dans son état actuel. Une étude sur une amélioration par des réseaux de neurones est en cours. Le cas échéant, l 'étudiant pourra être amené à reprendre ce travail.

Ce stage pourra donner lieu à des publications ainsi qu'à des présentations devant des partenaires académiques ou industriels. Il est à noter que la problématique abordée constitue un enjeu majeur dans le secteur industriel de la gestion électronique de document (GED) ainsi que dans celui de la dématérialisation des documents. Le candidat retenu devra donc faire preuve de discernement au regard de la confidentialité de certaines parties de son travail.

Prérequis et contraintes particulières :

- Langages : C++, Matlab, Python
- Outils de programmation pour l'analyse d'image : OpenCV, Matlab image processing toolbox
- Connaissances scientifiques : analyse d'images, des compétences d'analyse de documents et/ou des algorithmes de hachage sera un plus
- Langues : français, anglais
- Niveau Master 2

Références bibliographiques :

[1] A. Malvido Garcià, “Secure Imprint Generated for Paper Documents (SIGNED),” 2013.

[2] S. Eskenazi, P. Gomez-Krämer, and J-M. Ogier, “When document security brings new challenges to document analysis,” in *Proc. of 6th International Workshop on Computational Forensics (IWCF)*, 2014, pp. 1–13.

[3] S. Eskenazi, P. Gomez-Krämer, and J-M. Ogier, “Let’s be done with thresholds !” in *Proc. of the 13th IEEE International Conference on Document Analysis and Recognition (ICDAR)* , 2015, pp. 1–5.

[4] S. Eskenazi, P. Gomez-Krämer, and J-M. Ogier, “The Delaunay document layout descriptor” in *Proc. of the 15th ACM SIGWEB International Symposium on Document Engineering*, 2015, pp. 1–10.

Contacts – liens :

Renseignements et discussions : sebastien.eskenazi@univ-lr.fr, L3i, bât. Pascal, bureau 129
Candidature : petra.gomez@univ-lr.fr, jean-marc.ogier@univ-lr.fr, sebastien.eskenazi@univ-lr.fr

Merci de fournir un CV, une lettre de motivation, les relevés de notes des deux années de Master et un descriptif/rapport d’un projet/travail significatif que vous avez réalisé dans les deux dernières années.