PROPOSITION DE STAGE

Année 2018

Laboratoire L3i

Sujet de stage :

Analyse et description de la structure du document pour la sécurisation

Contexte du stage :

Le projet SHADES, soutenu par l'Agence Nationale de la Recherche, est un projet interdisciplinaire qui vise à sécuriser les documents rassemblant des laboratoires de recherche, une entreprise et une association de professionnels issus du domaine informatique et du droit. L'objectif du projet est de proposer de nouveaux outils pour garantir l'intégrité du contenu d'un document au travers d'une signature compacte avancée, afin de lutter contre la fraude et la falsification.
Cette signature repose sur l'analyse du contenu des documents (texte, logos, graphiques), ainsi que la structure (organisation spatiale de ces éléments), afin d'obtenir une signature sémantique. Grâce aux techniques de hachage utilisées, aucune information confidentielle du document ne pourra être déduite de cette signature, et celle-ci pourra être insérée dans un document sous la forme d'un code barre 2D (QR-CODE, 2D-DOC, …). Cette technologie est développée conjointement avec des juristes afin de garantir son utilisabilité dans un cadre juridique.

Résumé du travail proposé :

Dans nos travaux précédents, nous avons développé une méthode d'analyse de la structure du document (layout). Le résultat est un descripteur du layout très performant et stable [2]. Par contre, l'analyse du layout nécessite une segmentation de la page stable afin d'extraire les régions composant ce layout. Notre analyse des algorithmes de segmentation de page a montré que les algorithmes de segmentation actuels ne sont pas stables du tout [1]. Le résultat est que la description de layout est trop contraignante face aux résultats de segmentation instables. L'objectif de ce stage est de développer un descripteur de layout plus tolérant que [2].

Mots clés :

Sécurisation des documents, vérification de l'intégrité des documents, traitement d'images, analyse d'image de document, analyse/description du layout

Informations complémentaires :

Encadrant(s) : Petra Gomez-Krämer, Mickaël Coustaty

Equipe :

- **X Images et Contenus**
- **☐ Dynamique des systèmes et adaptativité**
- **☐ Modèle et Connaissance**

Domaine d'application stratégique :

- **☐ E-éducation**
- **☐ Environnement et développement durable**
- **☐ E-culture**
- **X Valorisation de contenus numériques**

Cadre de coopération : Projet de recherche national
Date de début du stage : Janvier ou plus tard en fonction de la disponibilité du candidat
Durée du stage : 5 ou 6 mois
Financement : ANR SHADES

LOCALISATION DU STAGIAIRE : L3i

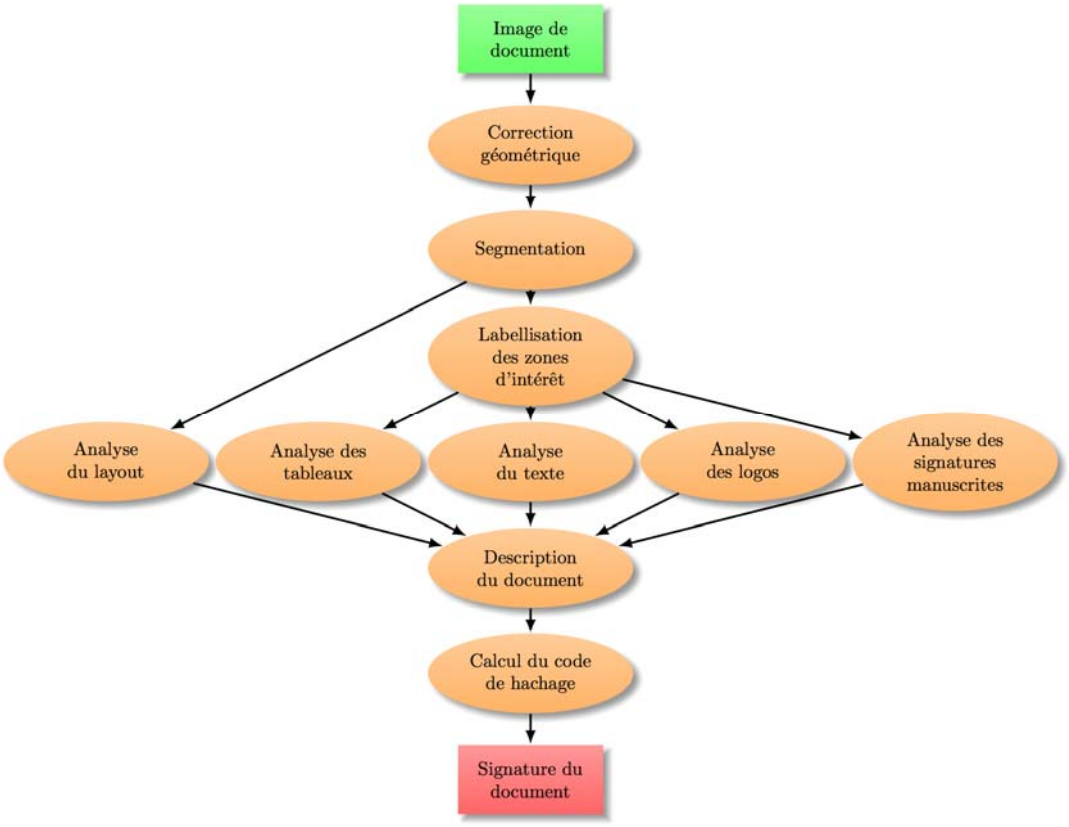
BESOIN MATERIEL : Ordinateur

Ce matériel est-il fourni ? oui

Contexte de l'étude:

Ce stage s'intègre dans les travaux du projet SHADES au sein du laboratoire L3i en partenariat avec l'entreprise ITESOFT, la FNTC, et trois 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.

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 dans les tâches « Segmentation » et « Analyse du layout ».

Description du sujet :

Le calcul de la signature décrite ci-dessus nécessite des algorithmes d'analyse de document stables face au bruit d'impression et de numérisation. La stabilité des algorithmes d'analyse de documents est un nouveau domaine de recherche. Contrairement à la précision des algorithmes, qui a beaucoup été étudiée, la stabilité de ces algorithmes n'a été jamais prise en compte. La précision peut être évaluée avec un seul résultat s'il y a aussi une vérité terrain. La stabilité ne nécessite pas une vérité terrain. La stabilité exige au moins deux résultats avec des entrées similaires pour voir à quel point ces résultats sont proches par rapport à la proximité des entrées. Dans notre cas, les entrées similaires sont deux photocopies du même document. Une conséquence de ceci est que l'algorithme peut être très stable et ne pas être précis. Ceci n'est pas à confondre avec la robustesse d'un algorithme. Un algorithme robuste est capable de produire des résultats pertinents face au bruit dans les images d'entrée. Par contre, il ne tient pas compte de la similarité des résultats entre elles.

Les derniers résultats montrent l'instabilité des algorithmes de segmentation de page [1]. Dans ce contexte on envisage deux pistes de recherche : 1) amélioration des algorithmes de segmentation et 2) un relaxation des contraintes pour la description du layout et ainsi d'obtenir un descripteur de layout plus tolérant. Donc, le travail de ce stage sera de développer une nouvelle méthode de description de layout plus tolérante que [2] en vue de la sécurisation du layout dans les images de document.

Prérequis et contraintes particulières :

- Niveau Master 2
- Langages : C++, Matlab

- Outils de programmation pour l'analyse d'image : OpenCV, Matlab image processing toolbox
- Connaissances scientifiques : traitement/analyse d'images, reconnaissance des formes, des compétences d'analyse de documents seront un plus
- Langues : français ou anglais

Références bibliographiques :

[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)*, pages 215-220, 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.

Contacts – liens :

Email : petra.gomez@univ-lr.fr, mickael.coustaty@univ-lr.fr

Lien vers le fichier de description : (PDF) (si nécessaire)

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.