PROPOSITION DE STAGE MASTER 2

Année 2017

Laboratoire L3i

Sujet de stage :

Développement d'un nouvel algorithme de fusion de régions de segmentation pour des composantes connexes en couleur

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 :

Nous avons récemment développé un nouvel algorithme de segmentation pour les images de document qui repose sur des composantes connexes en couleur. Cet algorithme s'est montré plus performant que les algorithmes de type super-pixel dans le contexte de la sécurité de document. Par contre, il produit une sur-segmentation de l'image nécessitant un post-traitement pour fusionner des régions de segmentation. Le travail de ce stage sera d'étudier les différents algorithmes de fusion de régions de segmentation et de développer un nouvel algorithme de fusion de régions pour des composantes connexes en couleur.

Mots clés :

Segmentation, composantes connexes, images de document, stabilité

Informations complémentaires :

Encadrant(s) : Petra Gomez-Krämer, Jean-Christophe Burie

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 mois minimum

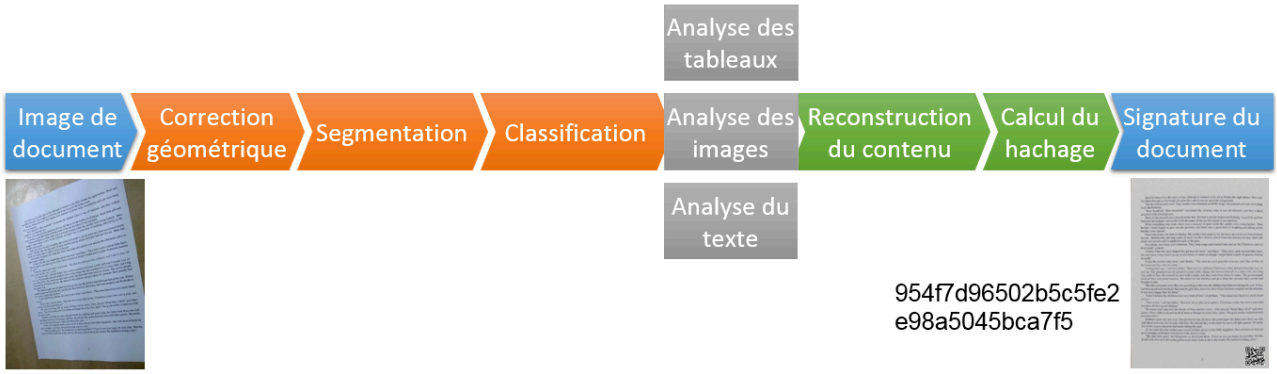
Gratification : environ 550 € net / mois (montant légal)

Financement : Projet ANR SHADES

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 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.

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 la tâche « Segmentation ».

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 de segmentation, qui a beaucoup été étudiée, la stabilité ne nécessite pas de vérité terrain. Pour évaluer la stabilité d'un algorithme de segmentation, il faut au moins deux images d'entrées et ses résultats de segmentation. Un algorithme de segmentation est stable : 1) les images d'entrée sont similaires (par exemple deux photocopies du même document), alors les résultats de

segmentation sont similaires ; 2) les images d'entrée ne sont pas similaires (deux documents différents), alors les résultats de segmentation ne sont pas similaires. Ceci n'est pas à confondre avec la robustesse d'un algorithme. Un algorithme de segmentation robuste est capable de produire des résultats de segmentation pertinents face à du bruit dans les images d'entrée. Par contre, elle ne tient pas compte de la similarité des résultats de segmentation entre elles.

Nous avons montré que les algorithmes de segmentation de document ne sont pas stables. Dans ce contexte, nous avons développé un nouvel algorithme de segmentation qui repose sur des composantes connexes en couleur. Cet algorithme est plus stable que les algorithmes de type super-pixel. Par contre, il produit une sur-segmentation de l'image nécessitant un post-traitement pour fusionner des régions de segmentation. Le travail de ce stage sera d'étudier les différents algorithmes de fusion de régions et de développer un nouvel algorithme de fusion de régions stable pour des composantes connexes en couleur.

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 d'images, des compétences d'analyse de documents sera 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.

[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.

Contacts – liens :

Email : petra.gomez@univ-lr.fr, jcburie@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.