PROPOSITION DE SUJET DE THESE

Campagne 2019

Laboratoire L3i

Seuls les étudiants ressortissants de l'Union européenne ou de la Suisse, n'ayant pas entamé leur carrière professionnelle sont éligibles pour ce financement.

Sujet de la thèse :

Authentification hybride de documents par leurs contenus graphiques et textuels

Résumé du travail proposé :

Le travail de cette thèse portera sur le développement d'une approche hybride image/texte pour vérifier l'authenticité des documents. Nous proposons de combiner une nouvelle méthode d'authentification du modèle de document par hachage basé sur le contenu (approche analyse d'image) avec une nouvelle méthode de vérification de la cohérence des contenus (approche analyse du texte).

Mots clés :

Sécurité de documents, détection de fraude, analyse d'images de document, analyse de texte (fouille de données textuelles, traitement automatique des langues, recherche d'information).

Informations complémentaires :

Encadrants : Petra Gomez-Krämer et Antoine Doucet
Lieu : Équipe Images et Contenus, Laboratoire L3i, La Rochelle Université
Cadre de coopération : DGA, industriel
Date de début du contrat : à partir du 1er octobre 2019
Durée du contrat : 3 ans
Salaire : environ 2 000 € brut

Pour candidater envoyez 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 par mail à petra.gomez@univ-lr.fr et antoine.doucet@univ-lr.fr.
Les candidatures restent ouvertes jusqu'à ce qu'un candidat satisfaisant ait été sélectionné.

Contexte de l'étude :

De plus en plus de documents sont dématérialisés et traités dans des grands flux d'images de documents dans les entreprises, les banques ou les administrations. Ainsi, la détection des fraudes dans ces documents devient un facteur de plus en plus important dans ces flux. La fraude sur les documents peut être la modification intentionnelle des documents (falsification) ou la production des faux documents (contrefaçon). Même si on détecte de plus en plus de document

frauduleux, un nombre significatif reste toujours non détecté.
Actuellement, il n'existe aucune solution fiable pour protéger les entreprises des fraudes documentaires, alors que des nombreuses entreprises perçoivent l'enjeu pour leur activité et manifestent un intérêt pour ce type de solution : selon une étude du cabinet PricewaterhouseCoopers (PwC), 49 % des répondants déclarent que leur entreprise a été victime d'une fraude ces deux dernières années, contre 36 % en 2016. Le taux de fraude constaté en France atteint un niveau record : 71% des entreprises françaises ont déclaré avoir été victimes d'une fraude au cours des deux dernières années.
Comme, il n'existe aucune méthode fiable de détection de fraude, nous souhaitons dans cette thèse poursuivre nos travaux sur la détection de la fraude dans les documents. Plus précisément, cette thèse vise à développer un nouvel outil pour la détection de la fraude (documents falsifiés et documents contrefaits) dans les flux documentaires. Nous proposons de combiner une nouvelle méthode d'authentification du modèle de document par hachage basé sur le contenu (approche analyse d'image) avec une nouvelle méthode de vérification de la cohérence de contenus (approche analyse du texte).

Description du sujet :

Malgré des besoins importants, la détection des fraudes automatique dans les documents est très peu étudiée. Les approches passives de détection de faux document peuvent s'appliquer à n'importe quel type de document. Le filtrage des faux parmi un flux de documents est une tâche de classification pour laquelle il s'agit de trouver les meilleurs indices ou caractéristiques de l'image et de sélectionner et entrainer le meilleur algorithme de classification automatique. La difficulté des flux de documents est la grande hétérogénéité des documents. Les documents peuvent être numérisés par un scanner ou capturés par un *smartphone* avec des résolutions différentes (entre 150 et 600 dpi) et des niveaux de couleur différents (noir et blanc, niveaux de gris, couleur) avec une grande diversité dans les contenus et les mises en page.
Il existe peu de méthodes dans l'état de l'art pour l'authentification ou la reconnaissance des modèles de documents. Les inconvénients de ces méthodes sont qu'elles supposent que le modèle du document est connu ou qu'elles sont conçues pour la recherche de documents. Pour cette raison, nous proposons de nous appuyer sur le hachage basé sur le contenu pour développer une nouvelle méthode d'authentification du modèle de document. La difficulté principale du hachage basé sur le contenu reste la stabilité des algorithmes d'analyse d'image de document. En outre, un code de hachage pour représenter le modèle de document doit être développé.
Parallèlement à l'analyse d'image, il existe des méthodes qui s'intéressent aux informations textuelles contenues dans le document. Il existe très peu de travaux scientifiques sur la vérification automatique d'information. Si on considère que les informations à vérifier sont localisées, la vérification de contenu présente deux difficultés principales. L'utilisateur ne dispose pas systématiquement dans ses bases de données d'une version fiable et vérifiée de l'information. En revanche, Internet est une source considérable d'information. Une approche pour confirmer des informations sans contexte peut être de mettre en œuvre des requêtes sur des moteurs de recherche, combinées à des techniques de *web scraping*. Le rapprochement entre une donnée vérifiée et une donnée extraite n'est pas trivial car cette dernière peut être dans une forme rédigée différente soit car sa rédaction est elle-même différente, soit car la donnée extraite a été impacté

par des erreurs d'OCR.
Aucune méthode de détection de fraude hybride n'a à ce jour été proposée. Donc, il s'agit de proposer également une méthode de fusion d'informations provenant de l'image et du texte.

Prérequis et contraintes particulières :

- Étudiant(e) titulaire d'un Master Recherche en informatique ou équivalent, avec des bonnes bases en mathématiques, en analyse et traitement du signal et des images, ou en traitement automatique de la langue
- Avoir des bonnes compétences en programmation

Références 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.