Proposition de recherche doctorale

Sécurisation des documents hybrides par une analyse physique des contenus

Mots clés :

- **Directeur de thèse** : nicole VINCENT
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'Informatique PAris DEscartes
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Projet Avec la multiplication des moyens de production et de reproduction des documents, leur sécurisation devient un problème de plus en plus difficile. Un même document peut prendre différents aspects selon les besoins, les logiciels et les périphériques d'acquisition ou de restitution disponibles. Par exemple, le document peut être en couleur ou fait de dégradés de gris, avoir un support papier ou avoir une nature électronique, être produit directement par un traitement de texte ou acquis par un appareil photo. Si un document papier peut être authentifié par un filigrane, ou un document électronique par un tatouage numérique, il n'existe pas aujourd'hui de solution pour authentifier un document hybride. C'est l'objet du projet ANR SHADES (2014) de proposer une approche pour résoudre le problème de l'authentification d'un document tout au long de son cycle de vie. La thèse se déroulera dans ce contexte, en collaboration avec le laboratoire L3I de La Rochelle en informatique, le laboratoire de droit, le Cejep, l'entreprise Itesoft et un utilisateur, la Fédération Nationale des Tiers de Concfiance (FNTC). Le changement de nature de l'objet conduit à l'introduction dans l'image ou le fichier électronique de bruits qui modifient l'objet sans entraver l'authenticité du document produit. De manière à créer une signature du document en non de son apparence physique, l'objectif est de construire une signature identifiant fidèlement le contenu du document. Nous entendons par contenu, l'ensemble des signes porteur de sens, sans entrer dans les aspects sémantiques. La signature sera construite à l'aide d'un processus de hachage sémantique. Le processus sera d'autant plus sûr que les objets traités seront homogènes, c'est à dire qu'un élément de signature sera d'autant plus efficace que l'on aura pu l'adapter à la nature de l'objet. Les zones que l'on peut distinguer dans un document sont le texte, les images, les logos, les signatures, les tableaux, des tampons par exemple.

Résumé du projet de recherche (Langue 2)

L'objectif de la thèse est donc de détecter les différentes zones d'un document, de manière aussi précise que possible, c'est à dire en distinguant le plus possible de types de support d'information différents. Si de nombreuses études existent, elles concernent les différents supports indépendamment ou bien font des hypothèses fortes sur les contenus des documents. Le problème sera alors de tirer parti des différentes approches pour conforter les résultats qui ne peuvent souffrir d'erreurs dans le cas de l'authentification. Le problème de la résolution de conflit devra être abordé entre différents média proches. De plus on s'attachera, par des approches multi-résolution et basées sur la géométrie discrète à discerner le graphique des images. La détection est une première étape qui doit être suivie d'une partie de description de chaque partie identifiée mais aussi de l'ensemble du document, sachant qu'une différence de répartition dans la page n'est pas un argument pour un refus d'authentification des contenus. La description est au cœur du processus d'authentification car celle-ci doit être indépendante du type de support et des déformations dues aux aléas se produisant lors des transmissions et des transformations du document. Le type de description doit être choisi en fonction de la nature de la zone étudiée. On voit ici que différentes échelles interviennent et devront être gérées. La dernière partie du travail concerne la constitution d'une signature la plus compacte et sûre possible. L'idée de départ étant de procéder à un hachage de l'information collectée sous forme symbolique / sémantique.
[BOU13] M. Bouguelia, Y. Belaid, A. Belaid, A Stream-Based Semi-Supervised Active Learning Approach for Document Classification, ICDAR 2013 [DOE98] D. Doermann, The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding, 70(3), 287-298, 1998. [OKU99] O. Okun, D. Doermann and M. Pietikainen, Page segmentation and zone classification: The state of the art. Technical report, University of Maryland. 1999 [REG12] P.P. Rege, C.A. Chandrakar, Text-image separation in document images using boundary/perimeter detection, Int. J. on Signal & Image Processing, Vol. 03, No. 01, Jan 2012 [SHI13] Z. Shi, S. Setlur, V. Govindaraju, A model based framework for table processing in degraded document images, In Conference on Document Analysis and Recognition (ICDAR), 2013.

Informations complémentaires (Langue 2)

Compétences souhaitables : Le candidat doit avoir une très bonne connaissance dans les domaines du traitement et de l'analyse d'images. Il doit aussi avoir un excellent niveau en programmation (par exemple en C, C++) et aussi une bonne connaissance de l'algorithmique et des structures de données.