L'INSA de Rennes recrute

un Post-doctorant en informatique

Combinaison de connaissances logiques et statistiques pour la reconnaissance d'images de registres anciens

Affectation

Etablissement : **INSA de Rennes, laboratoire de l'IRISA**

Service : **équipe de recherche Intuidoc de l'IRISA**

Poste à pourvoir dès que possible, pour un CDD de **12 mois**

Contexe

Le projet Intuidoc (https://www.irisa.fr/intuidoc) de l'Irisa travaille notamment sur la reconnaissance de la structure de documents anciens, manuscrits ou dégradés (partitions musicales, registres d'archives, journaux, courriers manuscrits, schémas électriques ...).

Doptim (https://www.doptim.eu) est une start-up créée en 2017 et spécialisée en data science avec deux activités, conseil auprès des entreprises qui souhaitent valoriser leurs données, et R&D sur des produits logiciels qui mettent en œuvre l'état de l'art en machine learning et en technologie big data.

La collaboration entre Doptim et Intuidoc concerne les images de manuscrits anciens principalement utilisés par les généalogistes amateurs du monde entier dans leur travail de reconstitution d'histoires familiales. Doptim développe une plate-forme web, Geneafinder, (http://www.geneafinder.com) pour rendre moins fastidieuse la recherche d'information dans les millions d'images disponibles en ligne.

Une technique en particulier est considérée : la possibilité de découper automatiquement les images de registres qui sont des successions d'actes de naissance, de mariage, de décès ou de minutes notariales, etc. pour afficher plus rapidement chaque acte dans un format lisible, sans manipulation, et enchaîner sur une transcription dans un contexte sémantique simplifié.

Missions

Les recherches de l'équipe Intuidoc ont mené à la mise en place de la méthode DMOS, (Description et MOdification de la Segmentation), qui est une méthode grammaticale pour la reconnaissance d'images de documents. Une extension de cette méthode, DMOS-P permet d'ajouter une dimension supplémentaire en prenant en compte plusieurs niveaux de perception d'une même image, par exemple à des résolutions différentes. Cette méthode est

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

générique et peut être appliquée pour la reconnaissance de n'importe quel type de documents.

Ce projet se focalise sur l'analyse de registres paroissiaux datés de 1675 à 1790 dans le but d'en faciliter la lecture pour des généalogistes. Des premiers travaux ont été menés pour proposer un découpage des pages en actes en se basant sur la localisation des lignes de texte. Les résultats préliminaires ont montré la nécessité d'introduire d'autres indices visuels, issus d'analyse statistiques plus poussées ou de systèmes de reconnaissance partielle de l'écriture.

Les objectifs visés par ce post-doc seront donc de :

- prendre en main les différentes techniques existantes dans l'équipe sur l'analyse de documents ;
- proposer des indices basés sur une analyse de propriétés statistiques de la page pour améliorer le découpage en actes de la page ;
- regrouper les différentes briques existantes en fusionnant les données intelligemment pour augmenter le taux de reconnaissance des pages de registres ;
- appliquer le système produit sur une base de données fournie par Doptim ;
- créer un référentiel pour pouvoir mesurer la qualité de découpage des images ;
- rendre le logiciel performant, capable de découper une image à la volée (un utilisateur charge une image, elle est découpée en quelques millisecondes) ;
- rendre le logiciel flexible, disposant de jeux de grammaire pour analyser des images de structure différente, à la volée ;
- documenter le logiciel pour permettre son intégration dans le cloud : multi-serveurs, robustesse, statistiques opérationnelles, debugage et mise à jour à distance, contrôle de version.

Compétences principales

Nous cherchons un docteur ayant une thèse dans le domaine de la reconnaissance de documents ou de l'analyse statistique. Des connaissances en grammaires et langages et/ou en programmation logique seraient un plus.

Environnement

Le CDD s'effectuera dans les locaux du laboratoire de l'Irisa, au sein de l'équipe Intuidoc. Il sera encadré par Bertrand Coüasnon et Aurélie Lemaitre, enseignants chercheurs.

Doptim intégrera les résultats régulièrement dans le projet d'expérimentation qui se met en place en Ille-et-Vilaine. Le projet est donc construit en mode agile avec des livraisons rapprochées et des retours terrain. Doptim mettra à disposition une plate-forme de test du service final, un support technique, notamment sur les algorithmes statistiques et organisera des réunions hebdomadaires de synchronisation.

Pour plus de renseignements, vous pouvez nous contacter par email :

bertrand.couasnon@irisa.fr, aurelie.lemaitre@irisa.fr sophie.tardivel@doptim.eu

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr