Séminaire MIDI : Wassim Swaileh

15 Mai 2020, 10:00

Titre du séminaire et orateur

Reconnaissance d'écriture et extraction d'information dans des documents anciens scannés.

Wassim Swaileh, MCF contractuel, ETIS, CY Cergy Paris Université

Date et lieu

Vendredi 15 mai 2020, 10h.

Par visioconférence

Résumé

La reconnaissance de l'écriture et l'extraction de l'information dans des images de documents anciens joue un rôle important dans la numération et la structuration de grandes masses d'archives de documents. La reconnaissance de l'écriture introduit un ensemble des difficultés liées à la qualité des images, la mise en page, la variabilité de l'écriture et la langue du document. Les systèmes de référence s'appuient sur un modèle optique, appris à reconnaître des caractères dans des images de lignes de texte, et des modèles de langues venant apporter des corrections linguistiques à la reconnaissance optique. Nous présentons un système de reconnaissance d'écriture générique au sein duquel nous proposons une méthode d'identification de lignes de texte dans les images et un modèle de langues basé sur des sous-unités lexicales appelées Multigrammes.

Suite à cette reconnaissance, extraire de l'information peut être difficile en raison d'erreurs de reconnaissance et du type d'information à extraire. En effet, certaines informations peuvent être reconnues par elles-mêmes (ville, marque,...), par leurs typographiques (majuscule, grammaire,...) ou selon le contexte. Nous présentons une approche d'extraction d'entités nommées dans des documents financiers s'appuyant sur l'apprentissage actif des champs aléatoires conditionnels sans et avec un modèle de réseaux de neurones récurrents type BLSTM.

http://pagesperso.litislab.fr/wswaileh/