Apprentissage et reconnaissance de structures tabulaires dans des documents manuscrits anciens

Encadrement : Bertrand COÜASNON, Irisa/Intuidoc, bertrand.couasnon@irisa.fr
Aurélie LEMAITRE, Irisa/Intuidoc, aurelie.lemaitre@irisa.fr
Lieu du stage : IRISA – Rennes

Mots-clés : Analyse d'images de documents, documents anciens, structures tabulaires, manuscrits, modélisation de connaissances, grammaire bi-dimensionnelle, apprentissage de structures, interprétation de structures

L'équipe de recherche Intuidoc de l'Irisa (http://www.irisa.fr/intuidoc) travaille notamment sur la reconnaissance du contenu et de la structure de documents anciens, manuscrits ou dégradés (partitions musicales, registres d'archives, journaux, courriers manuscrits, schémas électriques...). Dans ce contexte, les travaux de l'équipe ont abouti à des chaînes de traitement complètes (méthode DMOS-P [2]), conduisant à une reconnaissance de la structure des documents, décrite par des grammaires bidimensionnelles permettant d'exprimer la connaissance visuelle du document et des mécanismes perceptifs.

Dans les documents anciens, les structures tabulaires peuvent prendre des formes très variées, du tableau pré-imprimé plus ou moins dégradé, à une structure purement manuscrite, avec ou sans filets explicites. Cette grande variabilité rend leur extraction difficile, particulièrement lorsque les lignes de texte manuscrit sortent des colonnes prévues, comme on peut le rencontrer régulièrement dans des documents anciens, tels que des registres paroissiaux ou des documents administratifs (Figure 1).

Figure 1 : Exemples de structures tabulaires dans des documents anciens provenant de la compétition internationale cBAD 2017

L'objectif du stage de Master est de travailler sur la modélisation de ces structures tabulaires, en s'appuyant sur des indices visuels tels que les segments de droites extraits de manière perceptive, et les lignes de base du texte manuscrit. Le stage se concentrera sur la construction d'un système d'apprentissage interactif de structures tabulaires hétérogènes, en pouvant s'appuyer par exemple sur les travaux menés dans l'équipe sur l'apprentissage de structures

de documents [4], afin de permettre notamment une segmentation adaptée des lignes de texte. Les expérimentations seront menées sur des bases de registres paroissiaux que possède l'équipe ainsi que sur les documents complexes de la base (Track B) de la compétition internationale cBAD 2017 (https://scriptnet.iit.demokritos.gr/competitions/5/1/) sur la localisation des lignes de texte dans des registres anciens. Ce travail se fera également en collaboration avec la startup Doptim (https://www.doptim.eu) dans un contexte d'interprétation des registres paroissiaux et d'état-civil pour la généalogie.

A la suite de ce stage, un financement de thèse pourra être proposé dans le cadre du projet ANR HBDEX qui vient d'être accepté. Une possibilité de thèse CIFRE est également envisagé.

Référence

[1] Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon. **Multiresolution Cooperation Improves Document Structure Recognition**. International Journal on Document Analysis and Recognition (IJDAR), 11(2):97-109, Novembre 2008.

[2] B. Coüasnon. **DMOS: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems**. In *ICDAR, International Conference on Document Analysis and Recognition*, Seattle, USA, Septembre 2001.

[3] Coüasnon, B. & Lemaitre, A. **Recognition of Tables and Forms.** Handbook of Document Image Processing and Recognition, Doermann, D. & Tombre, K. (ed.), Springer London, 2014, 775-804.

[4] Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon. **Eyes Wide Open: an interactive learning method for the design of rule-based systems.** *International Journal on Document Analysis and Recognition*, Springer Verlag, 2017, 20 (2), pp.91-103.