Détection de tableaux dans des documents
Jusqu'à leur compréhension

Thème : Analyse d'image

Structures impliquées :
- Laboratoire LIPADE (Université Paris Descartes)
- Entreprise IMDS (Montréal – Canada)

Lieu du stage : Paris (rue des Saints Pères) et un déplacement à Montréal pour une petite période.

Durée du stage : 6 mois

Prise de contact : Pr. Nicole VINCENT (LIPADE)

Contexte
L'information est souvent exprimée de façon synthétique dans des documents sous forme de tableaux. L'échange de document est souvent réalisé par des documents électroniques acquis par l'aide d'un scanner ou d'un appareil photo. L'image ne contient plus l'aspect sémantique qui doit donc être reconstitué par des méthodes d'intelligence artificielle.

L'extraction de données dans un tableau dépend de la structure de ce tableau. Les tableaux peuvent être détectés par les méthodes basées [1, 2] sur la détection des lignes horizontales et verticales si les séparateurs (les traits horizontaux/verticaux) sont présents. Néanmoins, ces traits ne sont pas souvent explicites dans le document [3]

Objectifs
Implémenter une méthode qui permette de détecter et d'analyser des tableaux dans des documents hétérogènes, par exemple, des factures, des documents administratifs, etc.
Le sujet est lié à la détection de la structure complète d'un tableau, c'est à dire la structure physique (les traits séparateurs, colonnes, lignes, cellules, l'entête du tableau à l'aide de résultat d'OCR).

Tâches à réaliser
- Réaliser un État de l'art des méthodes de détection de tableau.
- Implémenter/adapter un prototype pour détecter la structure complète dans des documents hétérogènes (en C/C++)
- Évaluer ce prototype.

Aspects techniques
Le développement logiciel sera réalisé en C++. L'intégration finale des résultats sera réalisée sur les systèmes de la société IMDS dans ses locaux à l'occasion d'une étape de transfert de la technologie qui aura lieu à Montréal.