Doctorant: Maroua MEHRI (L3i/LITIS, Université de La Rochelle/Université de Rouen)

Encadrants: Rémy MULLOT, Pierre HÉROUX, Petra GOMEZ-KRÄMER, Alain BOUCHER

Titre: Catégorisation de contenus d'images de documents anciens par analyse multirésolution et approche texture

Résumé: Les récents progrès dans la numérisation des collections de documents anciens a ravivé de nouveaux défis dans la recherche d'information dans les bibliothèques numériques et l'analyse du contenu des documents numérisés. Par conséquent, afin de contrôler la qualité de la numérisation de documents et pour répondre à la nécessité d'une caractérisation de leur contenu à l'aide des métadonnées de niveau intermédiaire (entre l'image et la structure du document), nous proposons une catégorisation rapide et automatique du contenu d'images de documents anciens. Cette catégorisation s'appuie tout d'abord sur le calcul des indices de texture calculés à partir de la fonction d'autocorrélation. Les descripteurs d'autocorrélation sont obtenus par une analyse multirésolution et servent par la suite à extraire les zones homogènes de l'image du document numérisé à l'aide d'une méthodologie non supervisé de clustering. La méthode proposée se veut complètement non paramétrable et indépendante de la structure du document. L'originalité de ce travail vient aussi de l'absence de connaissances a priori, que ce soit sur le modèle de document (structure physique), ou les paramètres typographiques (structure logique). Pour évaluer notre approche et montrer sa pertinence en termes de bonne segmentation et caractérisation de contenu d'un corpus hétérogène, nous l'appliquons sur 316 images de documents anciens de la bibliothèque numérique Gallica. Ce corpus comprend six siècles (1200-1900) de l'histoire française. Par ailleurs, nous définissons une nouvelle métrique supervisée d'évaluation de clustering, nommée la mesure d'homogénéité. Nous obtenons une moyenne de 85% d'homogénéité. Ces résultats permettront de représenter le contenu d'un document par structure hiérarchique et de définir une ou plusieurs signatures pour chaque page, sur la base d'une représentation hiérarchique des blocs homogènes et leur topologie.