Sujet Thèse : « Indexation d'images de documents.»

Equipe : Image et Son
Thème : Structuration et Analyse d'image (SAI)

Nom des encadrants : Jean Philippe Domenger et Nicholas Journet.

Email : domenger@labri.fr, nicholas.journet@labri.fr

Contexte:

Les nombreuses campagnes de numérisation de documents patrimoniaux mises en place ces dernières années ont permis de constituer, un peu partout dans le monde, des entrepôts de données présentés sous la forme de collection d'images. Cette numérisation massive de documents soulève une problématique liée à l'indexation de grosses quantités d'images.

En effet, des premiers travaux de recherche ont mis en avant la difficulté relative à la recherche d'éléments dans une collection d'ouvrages numériques dont le contenu se trouve être fortement variable.

Enjeux scientifiques et objectifs de la thèse:

L'objectif de ces travaux de recherche est de proposer une plateforme opérationnelle de numérisation et de traitements logiciels intégrés, appuyée sur un moteur de recherche, permettant de produire, enrichir et de diffuser numériquement des contenus patrimoniaux.

Les recherches s'orienteront vers la mise en place d'une architecture informatique permettant de piloter une chaîne complète de traitement d'images (de la numérisation à la recherche par le contenu). Pour cela, il sera nécessaire de définir des descripteurs d'images robustes, permettant ainsi de piloter des modules d'OCR de segmentation et de (re)numérisation des ouvrages.

L'objectif des travaux de recherche consistera également à proposer une solution permettant le passage à l'échelle de ces nouvelles méthodes d'indexation. Actuellement aucun travail n'a abordé le traitement de masse de documents patrimoniaux. Il paraît donc indispensable de proposer une solution logicielle permettant de rationaliser le processus global en proposant à la fois une indexation et une recherche par le contenu la plus rapide possible.

Domaine de recherche:

Les travaux menés dans ce sujet de thèse concernent premièrement le domaine de la segmentation d'image et plus particulièrement celui de la segmentation du document. La segmentation du document se fera à plusieurs niveaux, découpage du document à un haut niveau (table des matières, chapitre, paragraphe, etc) puis à des niveaux plus bas (ligne, mot, caractère, image, lettrine, etc). Pour chacun des niveaux, il faudra caractériser et définir les descripteurs qui permettront de piloter au mieux la segmentation en fonction des informations recueillis au niveau supérieur et/ou inférieur.

Un deuxième domaine concerné par ce sujet de recherche est celui de l'indexation du document. En fonction des données extraites par l'OCR, il s'agira de classer les informations

présentes dans le document en fonction d'une classification précise. Cette classification se fera en accord avec les requêtes qui peuvent être traitées par le moteur de recherche.

Le candidat devra être intéressé par la développement logiciel, en effet les recherches menées dans cette thèse devront s'accompagner de la réalisation de prototypes afin de les intégrer dans une chaîne complète de numérisation de document. Ce projet se fera en collaboration avec des partenaires possédant des collections d'ouvrage à numériser (bibliothèque) et avec des industriels impliqués dans le développement/fabrication de chaîne de numérisation.