Sujet de master recherche « Architectures logicielles distribuées » 2006–2007

Segmentation d'images contenant des graphiques et du texte manuscrit et/ou typographique

Encadrant principal : José Martinez
courriel : José.Martinez@univ-nantes.fr
tél. : 02 40 68 32 36

Co-encadrant(s) : Frédéric Bouchara, Emmanuel Bruno et Elisabeth Murisasco, LSIS Toulon Université du Sud Toulon-Var, tél. : 04 94 14 (21/26), courriel : bouchara, bruno, murisasco@univ-tln.fr

Objectif du stage

Ce stage se situe dans le cadre de la construction d'une maquette permettant d'illustrer la représentation et l'interrogation de documents numérisés composés de zones de textes manuscrits, de zones de textes typographiques, d'images et de schémas (dans un environnement XML).

La réalisation de cette maquette nécessite :

1. l'identification des éléments composant l'image ;
2. la modélisation de ces éléments (types de zones, positionnement sur le document, structure, contenu, éventuellement annotations, etc., représentés dans un schéma de données XML en vue de leur exploitation avec les outils et langages existants dans le monde XML – comme le langage d'interrogation XQuery) ;
3. l'interrogation :
   - mixte (données structurées, transcriptions de textes et contenu visuel),
   - efficiente (c'est-à-dire rapide) en exploitant classification et parallélisme. (Les données multidimensionnelles doivent faire face au problème maintenant bien connu de la « malédiction de la dimensionnalité ». En d'autres termes, lorsque le nombre de caractéristiques à prendre en compte simultanément, c'est-à-dire de manière conjonctive, croît fortement, les techniques d'indexation y compris les plus performantes (ex. : arbres-X [1]) s'effondrent.)

Travail à réaliser

Le cœur du sujet concernera sans doute les deux premières étapes, c'est-à-dire la segmentation des documents et leur modélisation correspondante. Il existe dans la littérature un certain nombres d'outils dans le domaine du traitement d'image (analyse de texture par filtres de Gabor, champs de Markov...) qui ont été appliqués avec succès à des documents similaires (séparation texte / image ou segmentation de textes manuscrits) [2, 3, 4].

Le travail de stage consistera essentiellement à :

1. étudier différentes techniques de segmentations existantes ;
2. les évaluer ;
3. les adapter à la problématique, éventuellement.
4. développer un prototype les mettant en œuvre dans le cadre des documents anciens.

Une seconde partie consistera à :

1. développer un prototype les mettant en œuvre dans le cadre des documents anciens ;
2. évaluer, et si possible améliorer, l'efficacité et l'efficience du prototype.

Références

[1] S. Berchtold, D. A. Keim, and H.-P. Kriegel. The X-tree : An index structure for high-dimensional data. In *22nd International Conference on Very Large Data Bases (VLDB)*, pages 28–39, Mumbai (Bombay), India, September 1996.

[2] S. Nicolas, T. Paquet, and L. Heutte. Extraction de la structure de documents manuscrits complexes à l'aide de champs markoviens. In *9e Colloque Francophone sur l'Écrit et le Document (CIFED 2006)*, pages 13–18, Fribourg, Suisse, 2006.

[3] Yalin Wang, Ihsin T. Phillips, and Robert M. Haralick. Document zone content classification and its performance evaluation. *Pattern Recognition*, 39 :57–732, 2006.

[4] Keechul Jung, Kwang In Kim, and Anil K. Jain. Text information extraction in images and video : a survey. *Pattern Recognition*, 37 :977–997, 2004.