Titre : Analyse d'images de documents patrimoniaux : une approche structurelle à base de texture

Doctorant: Maroua MEHRI (L3i-Université de La Rochelle, LITIS-Université de Rouen)
Encadrants: Rémy MULLOT, Pierre HÉROUX, et Petra GOMEZ-KRÄMER

Résumé : Les récents progrès dans la numérisation des collections de documents patrimoniaux ont ravivé de nouveaux défis afin de garantir une conservation durable et de fournir un accès plus large aux documents anciens. En parallèle de la recherche d'information dans les bibliothèques numériques ou l'analyse du contenu des pages numérisées dans les ouvrages anciens, la caractérisation et la catégorisation des pages d'ouvrages anciens a connu récemment un regain d'intérêt. Les efforts se concentrent autant sur le développement d'outils rapides et automatiques de caractérisation et catégorisation des pages d'ouvrages anciens, capables de classer les pages d'un ouvrage numérisé en fonction de plusieurs critères, notamment la structure des mises en page et/ou les caractéristiques typographiques/graphiques du contenu de ces pages.

Les systèmes actuels de caractérisation et catégorisation des pages d'ouvrages numérisés s'appuient sur plusieurs critères relatifs au contenu textuel. Cependant, des performances insatisfaisantes ont été relevées en raison de divers problèmes, et qui sont liés aux particularités des documents anciens (e.g. une grande variabilité de la mise en page, des niveaux différents de dégradation et bruit, le défaut d'orientation, la complexité de la mise en page, des alignements non-conventionnels, les polices de caractères spécifiques, la présence d'ornements, les variations de l'espacement entre les caractères, mots, lignes, paragraphes et marges, la superposition de plusieurs couches d'information). En effet, leurs performances sont étroitement liées à celles des outils de reconnaissance optique de caractères et rétro-conversion. En outre, le traitement de ce type de documents peut s'avérer complexe et pénible en raison des particularités des documents anciens mentionnées ci-dessus, et ce, sans connaissances *a priori* sur la structure des mises en page ou les caractéristiques typographiques/graphiques du contenu de ces pages.

Ainsi, dans le cadre de cette thèse, nous proposons une approche permettant la caractérisation et la catégorisation automatiques des pages d'un ouvrage ancien. L'approche proposée se veut indépendante de la structure et du contenu de l'ouvrage analysé. Le principal avantage de ce travail réside dans le fait que l'approche s'affranchit des connaissances préalables, que ce soit concernant le contenu du document ou sa structure. Elle est basée sur une analyse des descripteurs de texture et une représentation structurelle en graphe afin de fournir une description riche permettant une catégorisation à partir du contenu graphique (capturé par la texture) et des mises en page (représentées par des graphes). En effet, cette catégorisation s'appuie sur la caractérisation du contenu de la page numérisée à l'aide d'une analyse des descripteurs de texture, de forme, géométriques et topologiques. Cette caractérisation est définie à l'aide d'une représentation structurelle. Dans le détail, l'approche de catégorisation se décompose en deux étapes principales successives. La première consiste à extraire des régions homogènes. La seconde vise à proposer une signature structurelle à base de texture, sous la forme d'un graphe, construite à partir des régions homogènes extraites et reflétant la structure de la page analysée. Cette signature assure la mise en œuvre de nombreuses applications pour gérer efficacement un corpus ou des collections de livres patrimoniaux (par exemple, la recherche d'information dans les bibliothèques numériques en fonction de plusieurs critères, ou la catégorisation des pages d'un même ouvrage). En comparant les différentes signatures structurelles par le biais de la distance d'édition entre graphes, les similitudes entre les pages d'un même ouvrage en termes de leurs mises en page et/ou contenus peuvent être déduites. Ainsi de suite, les pages ayant des mises en page et/ou contenus similaires peuvent être catégorisées, et un résumé/une table des matières de l'ouvrage analysé peut être alors généré automatiquement.

En effet, une approche ascendante de segmentation exploitant des descripteurs de texture mesurés à différentes échelles est tout d'abord proposée pour l'extraction des régions homogènes. Cette approche est notamment guidée par *(i)* la nécessité de robustesse au bruit fréquemment présent sur les images de documents anciens, *(ii)* le fait de pouvoir traiter des documents dont les mises en page et caractéristiques typographiques sont variées et, *a priori*, inconnues.

Dès lors que les zones homogènes ont été extraites, la seconde étape de l'approche construit une signature structurelle de la page (*i.e.* graphe). Les nœuds du graphe ainsi produits sont associés aux

zones homogènes et sont étiquetés par les attributs caractérisants les régions. Les arcs, quant à eux, caractérisent les liens topologiques entre les différentes régions.

Cette signature structurelle associant représentation des éléments de contenu et description de la mise en page, caractérise les pages de documents anciens numérisés à différents niveaux. Elle offre ainsi plusieurs modalités de catégorisation et d'indexation permettant une navigation multi-critère dans les corpus, et ce, sans reconnaissance et en ayant introduit aussi peu de connaissances *a priori* que possible. Dans le cadre de cette thèse, nous avons notamment étudié comment la signature produite par l'approche proposée pouvait être exploitée afin de comparer et catégoriser les pages d'un même ouvrage. Pour illustrer l'efficacité de la signature proposée, une étude expérimentale détaillée a été menée dans ce travail pour évaluer deux applications possibles de catégorisation de pages d'un même ouvrage, la classification non supervisée de pages et la segmentation de flux de pages d'un même ouvrage. En outre, les différentes étapes de l'approche proposée ont donné lieu à des évaluations par le biais d'expérimentations menées sur un large corpus de documents patrimoniaux.