Système générique et omni-langage de navigation dans des bases de documents anciens basé sur de la recherche de mots par composition interactive de requêtes

Quang Anh BUI [1]

Laboratoire L3I, Université de La Rochelle

ABSTRACT. *Word retrieval for browsing old digitized document collections is an active field of research. Indeed, because of the bad quality of this type of documents and the use of an ancient language, applying a basic OCR is not enough in general. In order to circumvent these difficulties, we are currently working on a generic, omni-language and interactive word retrieval system for browsing old document collections. This approach enables the user to retrieve words in any old collection of documents, whatever the alphabet, pictographs or ideograms used, without previously detecting an occurrence of the word in the collection, and even without mastering the language.*

RÉSUMÉ. *La recherche de mots ou de groupe de mots pour la navigation dans des collections de documents anciens numérisés est un sujet de recherche actif dans la communauté internationale. En raison en particulier de la qualité médiocre de ce type de documents et de l'utilisation d'un langage ancien ou rare, appliquer un simple OCR sur ces documents ne suffit pas, d'autant que certains alphabets ne disposent d'aucun système de reconnaissance automatique. Afin de contourner ces difficultés, nous proposons un système générique, omni-langage et interactif de recherche de mots dans des collections de documents anciens. Cette approche permet de travailler sur n'importe quelle collection de documents anciens, utilisant n'importe quel alphabet, pictogrammes ou idéogrammes. Dans ce contexte, l'utilisateur peut composer sa requête et il n'a pas besoin de maîtriser le langage ni de détecter préalablement une occurrence du mot-clé.*

KEYWORDS: *word retrieval, old documents, omni-language*

MOTS-CLÉS : *recherche de mots, documents anciens, omni-langage*

1. Encadrants: Rémy MULLOT et Muriel VISANI. Laboratoire L3I, Université de La Rochelle, avenue Michel Crépeau 17042 La Rochelle CEDEX 1

1. Introduction

Chaque année, un nombre croissant de documents anciens sont numérisés pour être préservés et également pour diffuser le patrimoine culturel au plus grand public. Cette numérisation offre en outre aux historiens et autres chercheurs de nouvelles possibilités d'indexation et de recherche associées à leur discipline.
Il existe différentes méthodes pour l'indexation et la recherche d'information depuis de ce type de document. La plus simple d'entre elles est la transcription manuelle depuis sous forme de documents textuels. Le résultat de la recherche est précis et correct, mais le coût de la transcription est important.
Pour éviter cette transcription manuelle, des méthodes de transcription automatique sont disponibles. Il suffit alors d'appliquer un système de reconnaissance (OCR) pour convertir les documents originaux en documents textes. Cependant, la qualité médiocre des documents anciens et/ou l'utilisation d'un langage ancien ou rare rend la segmentation du mot en caractères et la reconnaissance de caractères difficiles, d'autant que certains alphabets ne disposent d'aucun systèmes de reconnaissance automatique.

Le "Word-spotting" (Manmatha *et al.,* 1996), (Lu *et al.,* 2002) est une méthode plus récente et générique qui peut être appliquée à n'importe quel document écrit dans n'importe quelle langue, utilisant n'importe quel alphabet, pictogrammes ou idéogrammes. En pratique, le "word-spotting" consiste à retrouver dans le document toutes les occurrences d'une image d'un mot-clé qui est fournie par l'utilisateur qui doit donc préalablement en détecter une occurrence dans le document. Le "word-spotting" est basé sur une similarité entre cette image-requête et les mots segmentés extraits de la collection de documents. Cette technique, plus proche des systèmes de recherche d'images par le contenu (CBIR), peut être utilisée lorsque les systèmes de transcription automatique ne peuvent pas s'appliquer. Le principal inconvénient de cette méthode est que l'utilisateur doit trouver une occurrence du mot-clé dans le document, ce qui peut s'avérer fastidieux.
Pour résoudre ce problème, des chercheurs ont proposé une alternative: le "word retrieval". Cette technique permet la génération d'images-requêtes à partir d'un codage prédéfini. Les images-requêtes peuvent être synthétisées à partir de polices anciennes simulées, mais la forme des caractères est idéalisée et la variation du style d'écriture ne correspond pas à la variabilité que l'on trouve dans la réalité. Une autre solution pour créer les images-requêtes est la technique de la transcription semi-automatique (Le Bourgeois *et al.,* 2007). Le système segmente quelques pages d'un document en caractères. Ensuite, il applique une méthode de clustering sur ces caractères. Puis, en comparaison avec la solution précédente, cette solution crée des requêtes qui correspondent mieux à la réalités, car elles sont composées à partir du document dans lequel on fait la recherche. Par contre, l'utilisateur ne peut composer sa requête que si le caractère qu'il souhaite est effectivement dans la liste proposée par le système. De plus, le système de segmentation est spécifique à un alphabet donné, et notamment dans le cas de documents de qualité médiocre ou de langages rares, l'étape de segmentation du document en caractères peut s'avérer problématique.

Afin de contourner ces difficultés, nous travaillons à la conception d'un système générique, omni-langage et interactif de recherche de mots (word retrieval) dans des collections de documents anciens, fonctionnant même si l'alphabet n'est pas connu. Ce système repose sur trois étapes : la première phase est une extraction d'invariants depuis la collection de documents. Puis, la seconde phase consiste en une composition interactive par l'utilisateur de sa requête, et ce, en utilisant les invariants préalablement extraits. La troisième phase de recherche se fait en utilisant une distance entre les suites d'invariants de la requête et des mots extraits de la collection de document. Cette approche permet de travailler sur n'importe quelle collection de documents anciens, utilisant n'importe quel alphabet, cursif ou non-cursif, pictogrammes ou idéogrammes. Dans ce contexte, l'utilisateur n'a pas besoin de maîtriser le langage ni de détecter préalablement une occurrence du mot à rechercher, comme dans le word spotting. De plus, l'utilisateur peut composer librement et intuitivement sa requête.

Ce papier est organisé comme suit. En section 2, nous présentons le système proposé, tandis que la section 3 dresse un résumé et les perspectives de ce travail.

2. Système générique, omni-langage et interactif de recherche de mots

2.1. *Extraction d'invariants*

Dans la première étape, après un pré-traitement permettant d'enlever les bruits, d'améliorer la qualité du document et de binariser l'image, nous segmentons les mots du document en segments élémentaires. Puis, nous appliquons un algorithme de clustering afin d'extraire les invariants. Ces invariants doivent être suffisamment significatifs pour que l'utilisateur puisse composer facilement sa requête dans la deuxième phase. Pour cette raison, nous définissons les invariants comme "strokes" qui est défini comme une partie de la trajectoire d'écriture, délimitée par une pose du stylo à une extrémité et un retrait du stylo à l'autre extrémité.

Les méthodes d'extraction de strokes peuvent appartenir à deux catégories. La première catégorie, (Liu *et al.,* 1999), (Qiao *et al.,* 2004) est basée sur l'utilisation de squelettes, peu robustes au bruit. La seconde catégorie, (Lee *et al.,* 1998) utilise le contour de formes, tandis que (Su, 2003) est basée sur les filtres de Gabor. Pour ces deux catégories, l'une des principales causes d'erreur est liée aux ambiguïtés et incertitudes dans les régions d'intersection ou de jonction. Il faut alors être capable de déterminer quel "sous-stroke" est la continuté du stroke courant. Afin de résoudre ce problème, (Su *et al.,* 2009) propose un nouvel algorithme utilisant les points caractéristiques du squelette et l'information de contour englobant pour détecter les zones ambigües. L'algorithme n'utilisant pas d'estimation de courbure, la complexité de calcul est réduite et l'efficacité améliorée. Après la détection de zones ambigües, les mots sont divisés en deux parties: les sous-strokes et les zones ambigües. Les branches appartenant au même stroke sont déterminés en utilisant un classificateur bayésien. Nous avons choisi d'appliquer la méthode de (Su *et al.,* 2009) pour extraire des strokes d'un document. La Figure 1 montre quelques résultats sur un document manuscrit et sur un document imprimé.

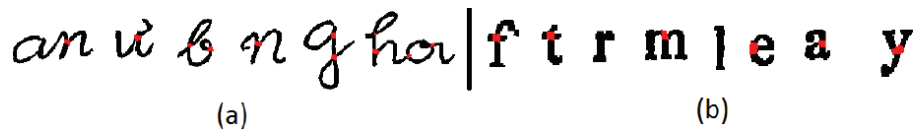

Figure 1. *Les traits de l'écriture dans un document (a) manuscrit et (b) imprimé. Les zones ambigües sont en rouge*

Les strokes sont décrits par des caractéristiques à la fois structurelles et statistiques, globales et locales: projection horizontale et verticale (Wu *et al.,* 2000), moments de Zernike, moments de Hu, ondelettes de Haar (Papageorgiou *et al.,* 1998). A partir des vecteurs de caractéristiques associés au stroke, nous appliquons un algorithme de clustering Hiérarchique, K-Moyennes, BIRCH (Zhang *et al.,* 1996)) pour extraire les invariants. A l'issue de cette phase, nous disposons d'invariants extraits directement du document qui seront à la base de la composition interactive de requête par l'utilisateur. La figure 2 donne des exemples d'invariants extraits avec notre méthode, dans 10 pages d'un document imprimé "Lord Byron" (Macaulay, 1856).

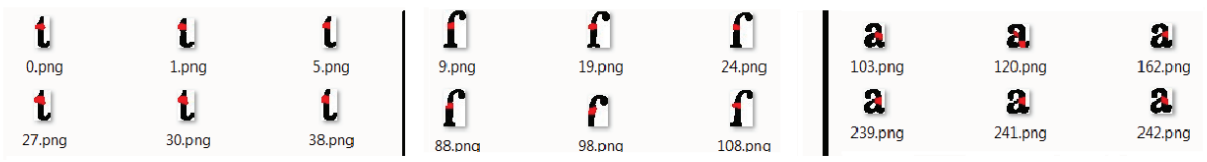

Figure 2. *Des exemples d'invariants extraits avec notre méthode*

La figure 3 donne une évaluation (non supervisée) de la qualité de la solution de clustering en utilisant la mesure Silhouette Width (Rousseeuw, 1987) pour 3 méthodes de clustering: Hérarchique, K-Moyennes et BIRCH, utilisant des différentes caractéristiques, appliquées sur les strokes extraits de 10 pages du document "Lord Byron". Pour chaque méthode de clustering, le nombre de clusters est fixé à 50. Nous pouvons conclure que l'utilisation du descripteur de projection et de la méthode de clustering BIRCH nous donne la meilleure solution (d'un point de vue non supervisé), constatation corroborée par nos observations visuelles.

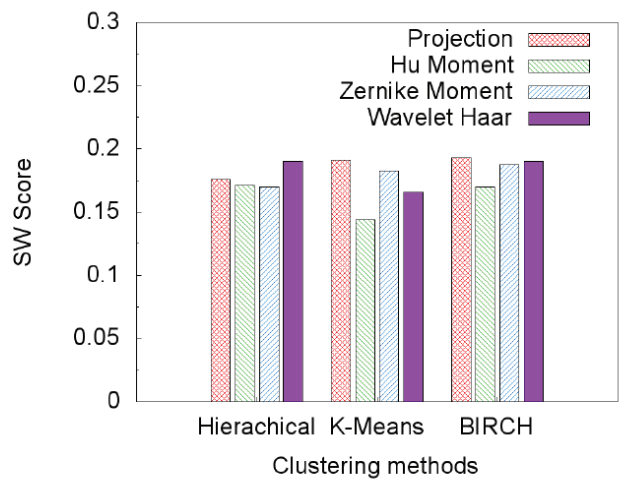

Figure 3. *Les mesures SW de 3 méthodes de clustering utilisant différents descripteurs*

2.2. *Composition interactive de requêtes*

Cette phase consiste à créer le mot à rechercher par composition interactive de l'utilisateur. L'idée est de proposer à l'utilisateur une interface intuitive lui permettant de composer sa requête à partir de rien, quand il n'a pu trouver une occurence du mot à rechercher dans le document. Dans le contexte de la recherche d'images par leur contenu (CBIR), (Fauqueur *et al.,* 2006) ont présenté un système permettant à l'utilisateur de composer lui-même sa requête, à partir de la représentation mentale qu'il se fait de l'image à rechercher et en se basant sur des catégories de régions extraites d'une base d'images. Dans notre cas, l'idée de base est similaire, mais la composition interactive se fera à partir de l'ensemble des prototypes des invariants extraits dans la première étape. Nous souhaitons à terme proposer une interface de composition interactive, basée sur le signal en-ligne transmis par l'utilisateur à l'aide d'une tablette par exemple. Cette interface permettra la suggestion automatique à l'utilisateur de la liste des invariants les plus proches des traits composant sa requête.

2.3. *Module de recherche*

A partir de la requête, il s'agit de rechercher dans l'ensemble de l'ouvrage les mots similaires. Plusieurs algorithmes de recherche de mots ont été proposées dans la litérature. La méthode de (Le Bourgeois *et al.,* 2007) détecte les zones d'intérêt et prend l'orientation de gradient dans ces zones comme caractéristique. Puis, la mise en correspondance avec les mots segmentés du document se fait en utilisant la méthode de "cohensive elastic matching". (Tan *et al.,* 2003), représente chaque mot comme une chaîne de symboles calculée à partir des composants de ce mot. La distance entre deux mots utilisée est la distance d'édition entre les deux chaînes de symbols. Dans notre système, les invariants sont extraits du document, et la requête est obtenue par la composition interactive de ces invariants. Il est donc logique que ces invariants soient réutilisés pour la recherche. Si nous n'avons pas encore eu le temps d'implémenter cette étape, nous nous concentrons a priori sur l'utilisation d'une distance d'édition entre chaînes d'invariants.

3. Conclusion

Dans ce papier, nous avons présenté notre système générique, omni-langage et interactif de recherche de mots dans des collections de documents anciens. Cette approche permet de travailler sur n'importe quelle collection de documents anciens, utilisant n'importe quel alphabet, pictogrammes ou idéogrammes. Ce système repose sur trois étapes. Pour les premiers mois de cette thèse, nous avons focalisé notre travail sur la première étape et nous avons quelques résultats préliminaires, suffisamment encourageants a priori pour envisager à court terme des tests sur des ouvrages anciens utilisant des écritures non latines (Khmer, Chinois, Arabe, Indien, etc.)

4. References

Fauqueur J., Boujemaa N., « Mental image search by boolean composition of region categories », *Multimedia Tools and Applications*, vol. 31, p. 95-117, 2006. 10.1007/s11042-006-0033-3.

Le Bourgeois F., Emptoz H., « DEBORA: Digital AccEss to BOoks of the RenAissance », *IJDAR*, vol. 9, p. 193-221, April, 2007.

Lee C., Wu B., « A Chinese-character-stroke-extraction algorithm based on contour information », *Pattern Recognition*, vol. 31, n° 6, p. 651-663, 1998.

Liu K., Huang Y., « Identification of fork points on the skeletons of handwritten Chinese characters », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, n° 10, p. 1095 -1100, oct, 1999.

Lu Y., Tan C. L., « Word spotting in Chinese document images without layout analysis », *ICPR, 2002*, vol. 3, p. 57 - 60 vol.3, 2002.

Macaulay T. B., *Lord Byron*, Longman, 1856.

Manmatha R., Han C., « Word spotting: a new approach to indexing handwriting », *Proceedings of CVPR 1996*, p. 631 -637, jun, 1996.

Papageorgiou C. P., Oren M., « A General Framework for Object Detection », *Proceedings of the International Conference on Computer Vision*, ICCV '98, IEEE Computer Society, Washington, DC, USA, p. 555-, 1998.

Qiao Y., Yasuhara M., « Recovering dynamic information from static handwritten images », *Proceedings of IWFHR 2004*, p. 118 - 123, oct., 2004.

Rousseeuw P. J., « Silhouettes: A graphical aid to the interpretation and validation of cluster analysis », *Journal of Computational and Applied Mathematics*, vol. 20, n° 0, p. 53 - 65, 1987.

Su Y., « A novel stroke extraction method for Chinese characters using Gabor filters », *Pattern Recognition*, vol. 36, n° 3, p. 635-647, 2003.

Su Z., Cao Z., « Stroke extraction based on ambiguous zone detection: a preprocessing step to recover dynamic information from handwritten Chinese characters », *IJDAR*, vol. 12, p. 109-121, June, 2009.

Tan C. L., Huang W., « Text Retrieval from Document Images Based on Word Shape Analysis », *Applied Intelligence*, vol. 18, p. 257-270, May, 2003.

Wu P., Manjunath B., « A texture descriptor for browsing and similarity retrieval », *Signal Processing: Image Communication*, vol. 16, n° 1-2, p. 33 - 43, 2000.

Zhang T., Ramakrishnan R., « BIRCH: An Efficient Data Clustering Method for Very Large Databases », *in* , H. V. Jagadish, , I. S. Mumick (eds), *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, ACM Press, p. 103-114, 1996.