Recherche de mots dans des images de documents par appariement de caractères

K. Khurshid, C. Faure, N. Vincent

Recherche de mots dans des images de documents par appariement de caractères

Khurram KHURSHID[1] - Claudie FAURE[2] - Nicole VINCENT[1]

[1] Laboratoire CRIP5-SIP
Université Paris Descartes, 75006 Paris, France

{khurshid ; nicole.vincent}@math-info.univ-paris5.fr

[2] UMR CNRS 5141 - GET ENST
46 rue Barrault, 75634 Paris cédex 13

cfaure@enst.fr

Résumé : *Repérer des mots ("word spotting") dans les documents imprimés anciens est une tâche extrêmement difficile. Les méthodes classiques, comme la corrélation, échouent quand elles sont appliquées sur les documents anciens. Ainsi pour résoudre ce problème, nous avons défini un mécanisme multipas d'analyse de document qui repose principalement sur l'extraction des mots et la caractérisation des caractères par une représentation multidimensionnelle. Les mots sont appariés à un modèle de mot en comparant les représentations multidimensionnelles des caractères qui les composent par un algorithme de "dynamic time warping" (DTW). Nous avons expérimenté cette approche sur des documents du XIXème siècle, imprimés sur des presses mécaniques, de la BIUM (Bibliothèque Interuniversitaire de Médecine, Paris). Nos premières expériences montrent des résultats extrêmement encourageants ayant une précision de 95% avec un taux de rappel de 89%.*

Mots-clés *:* Word-spotting, dynamic time warping, représentations des caractères, RLSA.

1 Introduction

L'importance des bibliothèques numériques pour la recherche d'information ne peut pas être niée. Les livres historiques anciens contiennent une information de valeur inestimable. Mais quand les livres anciens ne sont pas transformés en version électronique, le temps nécessaire pour rechercher l'information dans ces livres papier est considérable. Néanmoins la disposition sur écran des images de document n'est pas suffisante pour rendre l'information accessible. Notre travail, dans ce domaine, vise à faciliter la recherche de l'information en repérant des occurrences de mots dans les images des pages. Avec cette capacité de recherche dans les documents historiques anciens, les bibliothèques numériques augmenteront encore plus leur importance.

Repérer des mots dans les documents écrits avec l'alphabet latin a suscité récemment une attention considérable. Bien que beaucoup de travaux aient été déjà effectués dans le domaine de la caractérisation des mots, il reste toujours un champ de recherche car les résultats obtenus jusqu'ici ne sont pas suffisants pour traiter des volumes de données importants; en particulier si la base de documents se compose d'un ensemble de documents imprimés anciens de qualité relativement dégradée, ce qui est propre aux documents composés à la main et imprimés sur des presses mécaniques.

Cet article présente une manière efficace pour la recherche documentaire reposant sur la l'appariement de mots dans les documents anciens. Le papier est divisé en sections et commence par la description des travaux dans le même domaine. Elle est suivie de la description détaillée du modèle proposé et des différentes étapes impliquées dans le traitement du document. Enfin, nous montrons les résultats obtenus avec notre méthode.

2 L'etat de l'art

De nombreux travaux ont été réalisés sur la recherche de mots par *word spotting* dans les images de documents ainsi que sur la reconnaissance des caractères dans les images de documents anciens. Les nombreux problèmes liés aux documents imprimés anciens sont en particulier discutés en détail dans [ANT 04] et [BAI 04]. Ceux-ci incluent les causes physiques telles que la qualité des documents, les marques de liquides, les encres, la poussière, etc.; et les problèmes sémantiques [ANT 04]. Ces problèmes constituent un grand défi pour les chercheurs travaillant dans ce domaine pour améliorer les résultats. Dans cet article cependant, nous n'aborderons pas ces problèmes. Nous nous intéressons à la recherche de mots dans les images des pages d'un document à partir d'un exemple de ce mot. L'exemple constitue la requête qui se présente sous la forme d'une image de mot. Les résultats de la requête sont obtenus par l'appariement d'images. Des méthodes basées sur l'analyse des manuscrits ont été développées pour ce type de problème.

Rath et Manmatha [RAT 07; RAT 03] ont présenté une approche qui implique de grouper des images de mot dans des clusters des mots semblables, en employant l'appariement d'images de mot. Ils proposent quatre

caractéristiques de profil pour les images de mot qui sont alors appariées en utilisant différentes méthodes [RAT 07]. Leur travail a été effectué sur des documents manuscrits historiques. [ROT 03] a employé les correspondances entre les points anguleux pour classer des images de mot par similitude dans des manuscrits historiques. Le détecteur de points anguleux de Harris est employé dans les images de mot. Des correspondances entre ces points sont établies en comparant des fenêtres locales et en utilisant la somme des carrés des différences. La distance euclidienne entre les points mis en correspondance donne une mesure de similarité entre mots.

Des manuscrits en langue Telugu ont été caractérisés avec des représentations par ondelettes des mots [PUJ 02]. La représentation par ondelettes fournit les informations sur le contenu de l'image à différentes échelles. Elle exploite les caractéristiques inhérentes aux caractères du Telugu. Mais cette représentation par ondelettes ne donne pas de bons résultats pour les caractères latins [PUJ 02].

Adamek et al. ont présenté l'appariement des contours de mot pour leur reconnaissance holistique dans des manuscrits historiques. Les contours fermés de mots sont extraits et mis en correspondance en utilisant une technique de contours élastiques [ADA 07].

3 Système proposé

Notre méthode est basée sur l'extraction de différentes caractéristiques multidimensionnelles pour les images de caractère avant de comparer les mots. Par opposition à [RAT 07] où des caractéristiques sont extraites à partir de l'image entière du mot, nous segmentons les caractères du mot. Les caractéristiques sont ensuite extraites à partir des images des caractères. De ce fait on extrait l'information avec plus de précision dans le mot étudié que [RAT 07], nous le montrerons plus tard en comparant les résultats obtenus.

L'image du document est d'abord binarisée en utilisant notre algorithme NICK qui est une amélioration de la formule de Niblack originale [KHU 09]. Le texte et les zones graphiques des images de document sont séparés et les mots dans le document sont extraits en appliquant la technique du *Run length smoothing algorithm* (RLSA) [WAN 82]. Les mots correspondent aux composantes connexes de l'image obtenue après traitement par RLSA. On les appellera par la suite les "CCmots". Pour chaque mot détecté, les caractères qui le composent sont trouvés en revenant aux composantes connexes de l'image binarisée. Les erreurs de segmentation en caractères sont réduites en utilisant un processus de réparation en trois étapes. On obtient alors les caractères sur lesquels un ensemble de caractéristiques seront extraites. Les mots seront recherchés dans les CCmots détectés dans le document en mettant en correspondance les caractéristiques des caractères qui les composent et celles du mot de la requête par un algorithme de type DTW [KHU 08, KEO 01].

Le traitement que nous proposons est basé sur un prétraitement qui permettra de disposer des éléments nécessaires pour retrouver un mot donné.

Les différentes étapes du traitement des documents sont illustrées dans la figure 1. Ces traitements sont effectués hors ligne pour créer un fichier d'index pour chaque image du document. Les coordonnées de chaque mot, le nombre de caractères dans le mot, la position des caractères et aussi les caractéristiques de chaque caractère, sont stockés dans les fichiers d'index.

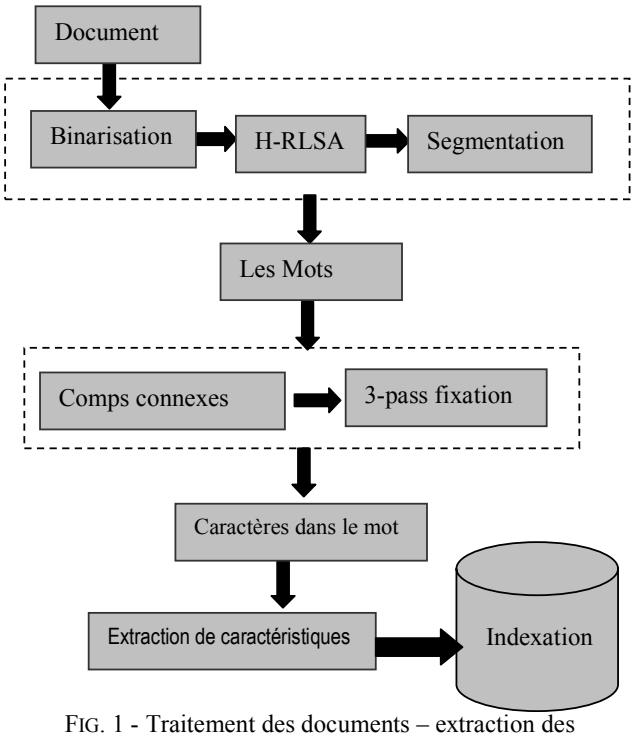

FIG. 1 - Traitement des documents – extraction des caractéristiques

La construction des fichiers d'index permet d'accélérer le traitement lors de la sélection d'un mot requête. Le choix du mot se fait en cliquant sur le mot dans l'interface graphique de notre système de traitement des documents. La requête est traitée de manière analogue au document global et les caractéristiques des caractères de la requête sont appariés avec les caractéristiques des caractères des mots déjà stockés dans le fichier d'index. Les mots pour lesquels la distance est inférieure à un seuil sont les mots acceptés (figure 2).

4 Indexation

Nous voyons maintenant en détail les différentes étapes du traitement.

4.1 Binarisation

Il n'est pas raisonnable d'utiliser un seuil global fixe de binarisation pour tous les documents. La qualité des résultats en recherche de mot dépend de la qualité de la binarisation. Aussi, nous avons modifié l'algorithme de

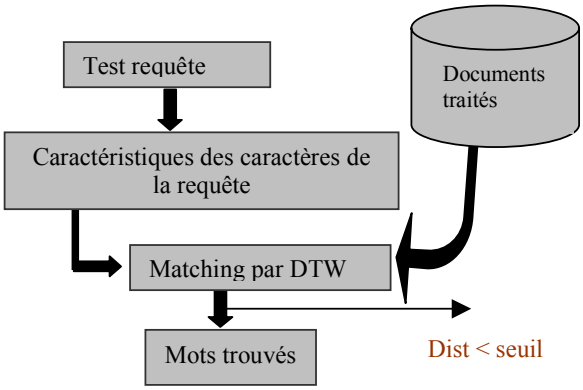

FIG. 2 - Word-spotting en utilisant le DTW

Niblack [KHU 09, LEE 03] pour le rendre plus efficace pour les documents anciens. Le seuil de binarisation, calculé pour chaque page, est calculé par la formule suivante :

$$T = m + k\sqrt{\sum(p_i^2 - m^2)/NPT}$$

Avec :
k = - 0.2
pi = niveau de gris du pixel i
m = moyenne des valeurs de gris
NPT = nombre de pixels

Nous avons expérimenté cette formule de deux manières :

- **Globale** : ce qui conduit à un seuil pour l'ensemble de l'image.
- **Locale** : un seuil local est recherché pour de petites fenêtres (15 x 15) dans l'image.

Nous disposons de différentes images issues de la BIUM [BIUM], avec deux résolutions différentes, une résolution correspondant aux images visualisées sur le site de la BIUM (550 x 913) ainsi qu'une résolution plus élevée (1536 x 2549). Les résultats obtenus sont extrêmement satisfaisants dans les deux cas. Par comparaison avec la formule originale de Niblack, nous constatons de meilleurs résultats pour des images ne présentant pas, ou très peu, d'éléments imprimés. Nous avons finalement choisi la méthode modifiée globale de Niblack. Elle présente l'avantage de fournir un seuil global tout à fait acceptable, d'avoir un temps de calcul inférieur à celui de la méthode locale et permet d'économiser les post-traitements qui sont nécessaires aux frontières des images pour la méthode locale.

4.2 Extraction des Mots par RLSA

Les images des pages sont binarisées puis traitées par RLSA. Les composantes connexes de ces images traitées sont les mots détectés : les CCmots. Nous appliquons un RLSA horizontal [WAN 82] avec le seuil égal à 5. Cette valeur est liée à la moyenne des largeurs des composantes connexes de l'image binarisée. Un exemple de CCmots détectés est montré sur la figure 3.

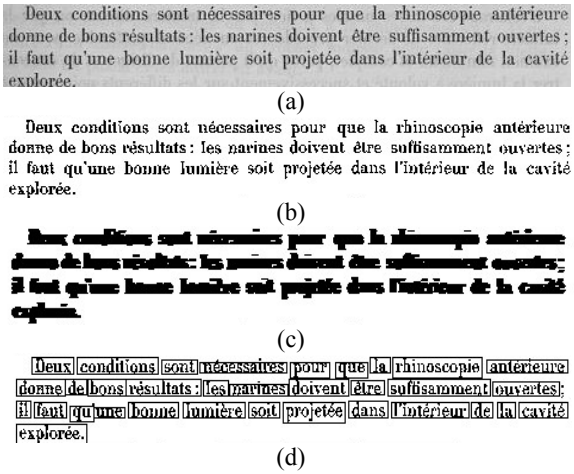

FIG. 3 - a) Image originale b) Image binaire c) RLSA d) les mots

4.3 Séparation texte/image

Les CCmots trouvés ne correspondent pas toujours à des mots, les composantes graphiques provoquent aussi la détection de CCmots. Elles vont être éliminées des CCmots sur des critères de taille. Les composantes graphiques vérifient les conditions suivantes :

- Aire composante > (aire moyenne x A)

ET

- Taille composante > (taille moyenne x B)

où l'aire moyenne est la moyenne des aires de toutes les composantes connexes dans cette image particulier. A et B sont des constantes que nous avons choisies égales à 5 et 4 respectivement en nous basant sur l'expérience des images de la BIUM. Les résultats prouvent que cette méthode fonctionne pour presque tous les types de documents pour séparer le texte des illustrations (fig 4).

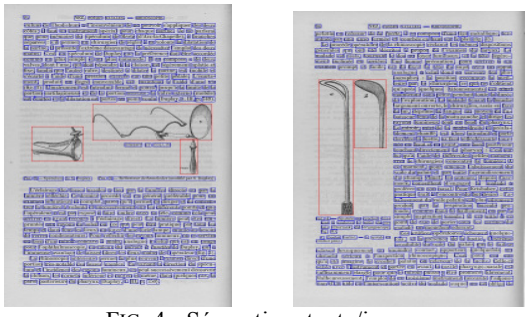

FIG. 4 - Séparation texte/image

4.4 Extraction des caractères

Un mot est un ensemble de caractères et ces caractères, dans le cas idéal, devraient être les composantes connexes. En fait une composante connexe ne correspond pas toujours à un caractère à cause de fusions, de ruptures et des points diacritiques. Nous effectuons donc une analyse des composantes connexes

extraites sur les images de mot pour améliorer la segmentation en caractères. Cette amélioration porte sur le regroupement de composantes connexes dans le cas de caractères fragmentés et de caractères comportant des marques diacritiques, ainsi que l'élimination de pseudo caractères dans le mot. Une méthode de réparation en trois étapes a été définie.

Dans la premère étape, nous considérons la projection des composantes connexes sur l'horizontale. Les composantes qui ont une partie commune dans leurs projections horizontales sont regroupées (figure 5a). A cette étape, les caractères accentués, les i et j sont reconstruits

Dans la deuxième étape, le recouvrement des composantes connexes est considéré. Les composantes dont les boîtes englobantes ont une intersection non nulle sont regroupées (figure 5b). Cette étape concerne des caractères comme r, g etc.

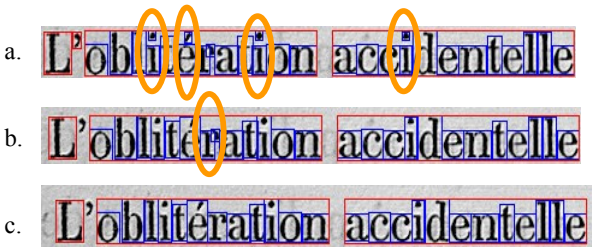

FIG. 5 – a) composantes originales b) Passe1 c) Passe2

Dans la troisième étape, nous enlevons les signes de ponctuation (comme " , " ou " . ") qui, inexactement, ont été inclus dans le CCmot (figure 6) car ils ne sont pas séparés des mots par un espace suffisamment grand. Pour cette étape, nous considérons la moyenne des aires des caractères du mot et supprimons les caractères ne remplissant pas la condition suivante :

Aire de la boîte englobante du caractère > 0,4 x aire du caractère moyen du mot

Les signes de ponctuation et les tout petits caractères constituant du bruit dans notre problématique sont marqués dans ce passage et ne sont plus considérés ainsi comme une partie du mot.

FIG. 6 - Suppression des caractères de ponctuation

4.5 Extraction de caractéristiques

Nous avons utilisé un ensemble de six vecteurs de caractéristiques pour représenter les images des caractères. Contrairement à [RAT 07], où il n'y a que quatre caractéristiques pour caractériser l'image du mot dans son ensemble, nous avons défini six caractéristiques pour les images de caractères, ce qui nous donne une meilleure représentation des mots dans un espace de caractéristiques, comme les résultats nous le révélerons plus tard. Les deux caractéristiques ajoutées permettent de mieux spécifier les formes. Ces six caractéristiques sont:

1. Le profil de projection verticale – la somme des valeurs d'intensité dans la direction verticale ; il est calculé dans l'image du caractère en niveaux de gris et normalisé pour obtenir des valeurs entre 0 et 1.

2. Le profil supérieur du caractère - dans l'image binarisée du caractère, pour chaque colonne, nous retenons la distance entre le rectangle circonscrit au caractère et le premier pixel du caractère.

3. Le profil inférieur de caractère - comme dans le profil supérieur, ici nous trouvons la distance entre le dernier pixel noir du caractère et la boîte englobante.
Les profils inférieur et supérieur sont normalisés entre 0 et 1. La Figure 7 montre les profils inférieur et supérieur calculés pour une image du caractère *p*.

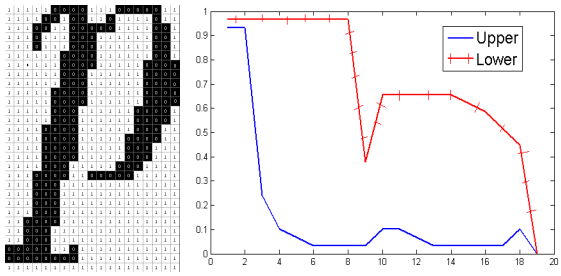

FIG. 7 - Profils supérieur et inférieur de l'image du caractère P

4. L'histogramme vertical – Le nombre de pixels noirs dans une colonne de l'image binaire du caractère.

5. Les transitions encre/non-encre - pour capturer la structure intérieure d'un caractère, nous calculons le nombre de transitions de non-encre à encre dans chaque colonne de l'image binarisée du caractère.

6. Transitions dans la ligne du milieu - pour la ligne centrale de l'image du caractère, nous trouvons le vecteur transitoire pour des transitions encre/non-encre. Nous plaçons un 1 pour chaque transition et 0 pour toutes les non-transitions dans la rangée. Au commencement nous avons considéré les transitions sur les trois rangées centrales et nous avons appliqué un OU logique sur celles-ci pour obtenir un vecteur transitoire moyen. Il suffisait qu'il y ait un pixel encre dans l'une des trois rangées centrales pour considérer que la colonne correspondante était un pixel encre. En comparant les deux approches, nous avons constaté que l'emploi d'une seule ligne centrale donne de meilleurs résultats et nous avons retenu cette caractéristique.

Pour chaque mot, nous trouvons les caractéristiques pour chacun de leurs caractères. Après le traitement des images d'un document, son fichier d'index est créé dans lequel sont stockés : la localisation des CCmots, la position de chacun de ses caractères et les caractéristiques de chaque caractère qui apparaissent comme une chaîne.

5 Appariement de mots

Pour que deux mots soient considérés éligibles à un appariement, nous avons placé des limites sur le rapport de leurs longueurs (nombre de caractères). Si le rapport est au-dessus d'une valeur spécifique, nous n'essayons pas d'apparier les deux mots. Pour mettre en correspondance les mots, nous considérons les caractéristiques des caractères des mots en employant le DTW. L'avantage d'employer le DTW est qu'il est en mesure de tenir compte de l'étirement et de la compression non-linéaires des mots car il trouve un axe de référence commun [KEO 01]. De cette manière, deux mots identiques qui diffèrent par leurs tailles seront mis en correspondance correctement, à la différence de la corrélation [BUR 82] où les mots doivent être de la même dimension pour être appariés.

Ainsi pour apparier deux caractères, nous traitons les vecteurs de caractéristiques des deux mots comme deux suites $X = (x_1 \ldots x_m)$ et $Y = (y_1 \ldots y_n)$. Pour déterminer la distance/coût du DTW entre ces deux suites, nous calculons une matrice D d'ordre m x n qui contient le coût/distance d'alignement des couples de sous-suites débutant en 1 et finissant respectivement en m et n. Les valeurs de la matrice D sont exprimées par :

$$D(i,j) = \min\begin{Bmatrix} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{Bmatrix} + d(x_i, y_j)$$

Ici pour définir d(xi , yj), nous avons employé la distance euclidienne dans l'espace des caractéristiques car celles-ci ont été normalisées :

$$d(x_i, y_j) = \sum_{k=1}^{p} (x_{i,k} - y_{j,k})^2$$

où p représente le nombre de caractéristiques qui dans notre cas est 6.

Une fois que toutes les valeurs de D sont calculées, le chemin d'appariement est déterminé comme le chemin de coût minimum entre la position (1, 1) dans la matrice et la position (m, n). Le coût d'appariement final est le coût D(m, n) divisé par le nombre d'étapes du chemin d'appariement. Deux caractères sont semblables si ce coût final est inférieur à un seuil (qui a été fixé après un ensemble d'expérimentations). Plus de détails sur DTW peuvent être trouvés dans [KEO 01] et [RAT 07].

La distance entre deux mots n'est pas la simple somme des distances entre les caractères ordonnés des mots. Le meilleur appariement global des caractères est recherché pour assurer plus de robustesse à la méthode. Chaque caractère du mot requête est mis en correspondance avec un nombre différent de caractères voisins dans le mot inspecté. Ce nombre est fonction de la taille du mot de requête. Pour chaque caractère de la requête, nous trouvons son meilleur appariement dans les caractères du mot inspecté et sommons les coûts de mise en correspondance des caractères. Après appariement des caractères des deux mots, nous normalisons le coût associé au mot en divisant par le nombre de caractères mis en correspondance. Si pour deux mots, ce coût normalisé est inférieur à un seuil, nous indiquons que les deux mots sont les mêmes.

Pour évaluer l'efficacité de notre méthode, nous l'avons comparée avec la méthode utilisant la corrélation pour réaliser l'appariement des CCmots et avec la méthode de [RAT 07] sur les mêmes images de documents de la BIUM. Ces résultats sont donnés dans la section suivante.

6 Experimentation

Il n'existe pas de base de données de documents imprimés anciens disponible pour le développement et la validation des méthodes d'appariement de mots. Afin de comparer notre méthode à d'autres, nous les avons testées sur la base réalisée pour cette étude à partir des documents de la BIUM. Les autres approches incluent la corrélation pour mesurer la similarité entre mots, l'appariement de mots représentés par l'ensemble des caractéristiques définies dans [RAT 07] et l'appariement de mots qui utilise les six caractéristiques que nous avons définies mais en les appliquant aux images de mot (et non plus aux caractères des mots). Nous avons comparé les résultats de ces trois méthodes à notre approche.

Nous avons entrepris deux ensembles d'expériences, le premier pour la comparaison des différentes méthodes et le second pour l'évaluation détaillée de notre méthode. Dans le premier ensemble d'expériences, nous avons choisi 35 images de mots requête de longueurs différentes dans différents documents. L'évaluation est faite en calculant les pourcentages de rappel et de précision. Les résultats pour l'approche par corrélation ne sont pas très satisfaisants. Ils donnent une précision juste au dessus de 70%. En employant l'appariement de caractéristiques des images de mots entiers, avec les caractéristiques proposées par [RAT 07], la précision tombe à 53%. En utilisant nos six caractéristiques pour les images de mots entiers, on peut noter dans la figure 8 que les taux de rappel et de précision sont plus élevés qu'avec l'approche de [RAT 07]. Enfin, pour notre méthode d'appariement des mots qui considère les caractères du mot, les résultats réalisés sont bien meilleurs qu'avec la corrélation ou avec [RAT 07] suivant les indications de la figure 8.

Pour l'évaluation détaillée de notre méthode, nous avons effectué un test en utilisant 500 exemples de requêtes de différentes longueurs dans cent images de différents documents. Dix valeurs de seuil différentes (à partir de 0,13 avec un incrément de 0,01) ont été appliquées et les variations des pourcentages du rappel et de la précision ont été notées (figure 9).

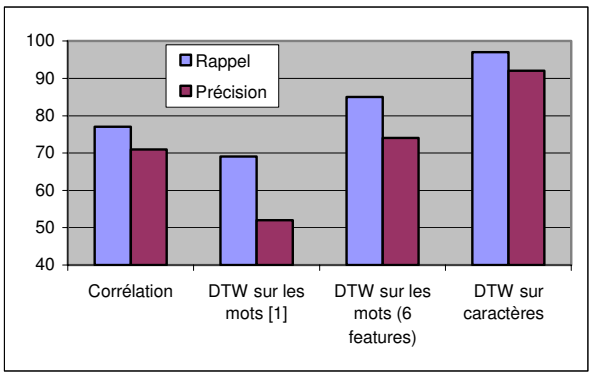

FIG. 8 – Comparaison entre la corrélation ; en utilisant 4 caractéristiques pour les mots, puis 6 caractéristiques pour les mot et enfin 6 caractéristiques pour les caractères des mots.

Nous avons également examiné l'effet de la longueur du mot sur le seuil convenable, pour de plus petits mots (avec peu de caractères), un seuil plus élevé donne de meilleurs taux de précision et de rappel, tandis que pour des mots plus longs, une valeur inférieure pour le seuil s'avère meilleure. Considérant des valeurs de seuil optimum pour différentes longueurs de mot, notre système réalise une précision de 95% tout en obtenant un rappel de 89%.

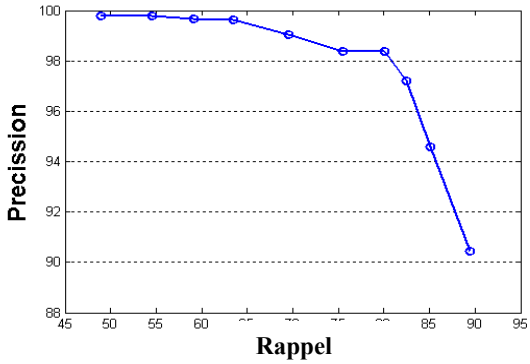

FIG. 9 - Précision en fonction du rappel pour différents seuils

7 Conclusion

Nous avons proposé une nouvelle méthode pour rechercher un mot dans une image de document. Elle est basée sur la mise en correspondance des caractéristiques des images de caractères contenus dans les mots par un algorithme de DTW et la définition d'un ensemble de caractéritiques de représentation. Les résultats obtenus en employant cette méthode sont très encourageants. Le nombre de faux positifs est très inférieur aux résultats obtenus par la corrélation et aussi par la méthode de [RAT 07]. Des améliorations sont encore possibles en affinant les différentes étapes, en particulier en scindant les caractères fusionnés, et en ajoutant davantage de caractéristiques. Actuellement, la recherche de mot est faite uniquement sur le texte horizontal mais notre travail prochain se concentre sur la recherche de mots dans des lignes verticales de texte. Ce travail peut également être adapté et testé sur des manuscrits.

Références

[RAT 07] Tony M. Rath, R. Manmatha, Word Spotting for historical documents, IJDAR (2007) 9:139-152

[KHU 08] K. Khurshid, C. Faure, N. Vincent, "Feature based word spotting in ancient printed documents", *8th International workshop on pattern recognition in information systems*, Spain, 2008.

[WAN 82] K. Y. Wang, R. G. Casey and F. M. Wahl, Document analysis system, IBM J. Res.Development, Vol. 26, 1982, pp. 647-656.

[LEE 03] Graham Leedham, Chen Yan, Kalyan Takru, Joie Hadi Nata Tan and Li Mian, Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images, *7th International Conference on Document Analysis and Recognition (ICDAR)*, 2003.

[BUR 82] P. J. Burt, C. Yen, X. Xu, Local Correlation Measures for Motion Analysis: a Comparative Study, *IEEE Conf. Pattern Recognition Image Processing* 1982, pp. 269-274.

[PUJ 02] A. K. Pujari, C.D. Naidu, B.C. Jinaga, An adaptive character recogniser for telugu scripts using multiresolution analysis and associative memory, *ICVGIP*, 2002

[KEO 01] Keogh, E. and Pazzani, M., Derivative Dynamic Time Warping, *First SIAM International Conference on Data Mining, (Chicago, IL),* 2001.

[ROT 03] Jamie L. Rothfeder, Shaolei Feng and Toni M. Rath, Using corner feature correspondences to rank word images by similarity, *Conference on Computer Vision and Pattern Recognition Workshop*, Madison, USA, 2003, pp. 30-35.

[BIUM] Bibliothèque Interuniversitaire de Médecine, Paris, http://www.bium.univ-paris5.fr/histmed/medica.htm

[ADA 07] Adamek, T., O'Connor, N. E. and Smeaton, A. F., Word matching using single closed contours for indexing handwritten historical documents, IJDAR, 2007, 9, 153 - 16

[ANT 04] A.Antonacopoulos, Karatzas D., Krawczyk H. and Wiszniewski B., The Lifecycle of a Digital Historical Document: Structure and Content, *ACM Symposium on Document Engineering,* 2004, 147 –154.

[RAT 03] Tony M. Rath, R. Manmatha, Features for Word Spotting in Historical Manuscripts, *Seventh International Conference on Document Analysis and Recognition (ICDAR),* 2003.

[BAI 04] Baird H. S., Difficult and urgent open problems in document image analysis for libraries, *1st International workshop on Document Image Analysis for Libraries,* 2004.

[KHU 09] Khurshid, I. Siddiqi, C. Faure, N. Vincent, "Comparison of Niblack inspired binarization techniques for ancient document images", *16th International conference on Document Recognition and Retrieval*, 2009 (submitted)