Word spotting dans les manuscrits syriaques dégradés basé sur des caractéristiques directionnelles

Petra Bilane, Stéphane Bres, Hubert Emptoz

Word Spotting dans des manuscrits Syriaques dégradés basé sur des caractéristiques directionnelles

Petra Bilane[1] – Stéphane Bres[1]– Hubert Emptoz[1]

[1] Laboratoire LIRIS – INSA De Lyon
20 Avenue Albert Einstein – 69621 Villeurbanne – France

{petra.bilane, stephane.bres, hubert.emptoz}@insa-lyon.fr

Résumé *: Ce papier propose une contribution au « Word Spotting » pour assister l'indexation de manuscrits Syriaques numérisés.*

Mots-clés *:* « Word Spotting », caractéristiques d'orientation, roses de directions.

1 Introduction

Le Syriaque appartient à la branche Armaique des langues sémitiques. Les manuscrits Syriaques les plus anciens peuvent dater du 1er siècle ap. J. C [CLO03]. Contrairement à ce qu'on pourra penser, le Syriaque n'est pas une langue morte, de nos jours, de nombreux poètes présentent des poèmes Syriaques modernes.

Très rares sont les gens qui se sont lancés dans l'étude des documents Syriaques. En dehors des travaux de William Clocksin [CLO03], aucun travail n'a précédemment été publié sur la reconnaissance de l'écriture manuscrite Syriaque. Dans nos travaux précédents [EGL07], notre objectif était l'analyse des informations globales pour la classification de documents selon le style d'écriture. Dans ce papier, nous nous concentrons sur le « Word Spotting ».

Les documents auxquels nous nous sommes intéressés sont des documents Syriaques datant du 19ème siècle qui nous ont été fournis par la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik au Liban.

2 Méthode proposée

Nous proposons une méthode qui commence par une phase de prétraitement. Puis nous procédons au choix des fenêtres glissantes intéressantes contrairement à Terasawa et al. qui ont pris en considération toutes les fenêtres glissantes [TER05]. Une fenêtre glissante de taille 32x32 pixels est passée le long de la ligne de gauche à droite à un pas de 1 pixel. Puis une analyse du contenu est effectuée :

- Les fenêtres ne répondant pas à certains critères de choix seront rejetées (densité minimale de pixels noirs, un taux de recouvrement ne dépassant pas 50%, un centre de gravite ayant suffisamment bougé suivant l'axe des abscisses par rapport à son prédécesseur)
- Les fenêtres restantes sont divisées en quatre sous fenêtres de taille 16x16 pixels. Nous calculons la fonction d'auto corrélation dans chacune de ces quatre sous fenêtres.
- Les motifs obtenus représentent les directions principales dans les quatre quadrants de la fenêtre courante. Cette information est résumée sous forme d'une rose à huit directions.
- Chaque sous-fenêtre 16x16 pixels est représentée par une signature de 8 valeurs ce qui donne un total de 32 valeurs représentant chaque fenêtre.
- Une fois les signatures de toutes les sous-fenêtres de toutes les fenêtres sélectionnées sont extraites des trois zones de l'image du mot requête, le but est de retrouver toutes leurs occurrences.
- Elles seront comparées à celles extraites de l'image de la page de test. Celles qui leur sont les plus similaires sont détectées et la région ayant l'agglomération la plus importante de sous fenêtres similaires à celles du mot requête est considérée comme étant une occurrence possible de ce mot.

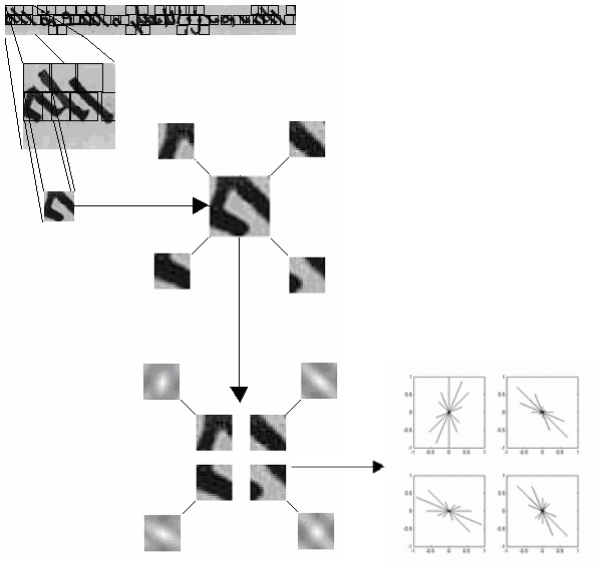

FIG. 1 – PROCESSUS D'EXTRACTION DES ROSES DE DIRECTIONS A PARTIR DES SOUS-FENETRES.

3 Premiers résultats

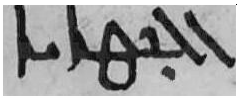

FIG. 2 – IMAGE DU MOT REQUETE.

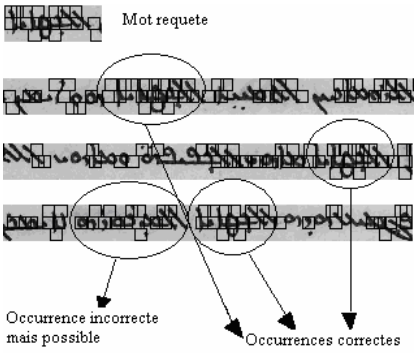

FIG. 3 – AGGLOMERATIONS PRETANT A CONFUSION.

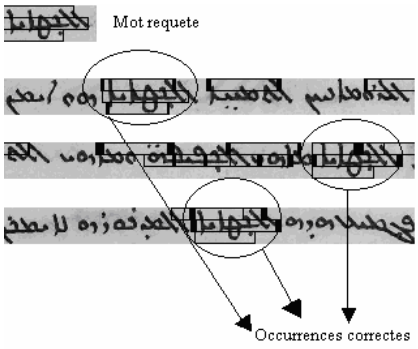

FIG. 4 – DETECTION DES REGIONS D'INTERET.

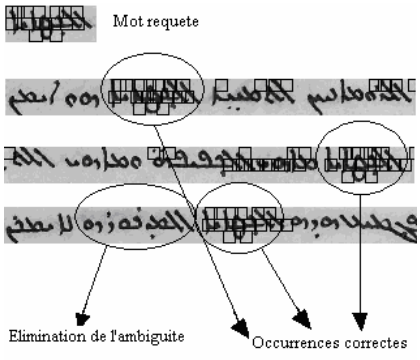

FIG. 5 – ELIMINATION DE L'AMBIGUITE.

4 Les cas dégradés : La compression excessive et la faible résolution

La compression excessive et la faible résolution sont les dégradations les plus communes aux versions numérisées de documents, le choix de la compression JPEG introduit des artefacts aux effets irréversibles, la tentative de rehausser une faible résolution aboutit à un effet de flou.

FIG. 6 – ARTEFACTS ET EFFET DE FLOU.

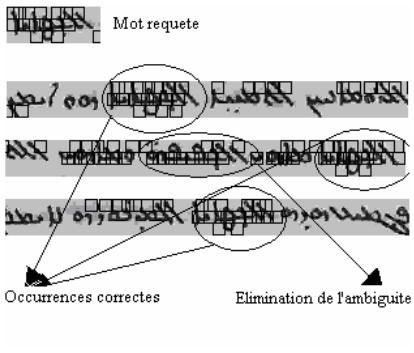

FIG. 7 – RESULTATS POUR LA DEGRADATION JPEG.

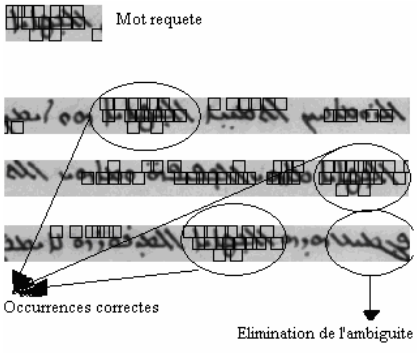

FIG. 8 – RESULTATS POUR LA FAIBLE RESOLUTION.

5 Conclusion

Nous avons présenté un algorithme de repérage de mots « Word Spotting » pour assister l'indexation de manuscrits Syriaques. Notre méthode ne nécessite aucune connaissance a priori pour le repérage. De plus elle est indépendante de tout algorithme de segmentation. La façon d'extraire les descripteurs les rend plus robustes face aux dégradations notamment la forte compression JPEG et la faible résolution.

Références

[CLO03] W. F. Clocksin, and P.P.J. Fernando, Towards automatic transcription of Syriac handwriting, *IEEE Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)*, Mantova, Italy, Sept. 2003.

[CLO04] W. F. Clocksin, Handwritten Syriac character recognition using order structure invariance, *IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, Cambridge, UK, Aug. 2004.

[EGL07] V. Eglin, S. Bres, and C. Rivero, "Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts", *International Journal on Document Analysis and Recognition (IJDAR'07)*, Springer- Verlag, Berlin Heidelberg, pp. 101-122, Vol.9, Apr. 2007.

[TER05] K. Terasawa, T. Nagasaki, and T. Kawashima, Eigenspace method for text retrieval in historical document images, *IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, Seoul, Korea, Aug. 2005.