A SEGMENTATION FREE APPROACH FOR INDEXING DIGITIZED SYRIAC MANUSCRIPTS

P. Bilane[1], S. Bres[1], K. Challita[2], H. Emptoz[1]

[1]LIRIS Laboratory, INSA of Lyon, 20, Av. Albert Einstein, 69621 Villeurbanne Cedex, France
[2]Faculty of Sciences and Computer Engineering, USEK, PO.Box : 446, Jounieh, Lebanon
{pbilane, sbres, hemptoz}@liris.cnrs.fr, khalilchallita@usek.edu.lb

ABSTRACT

This paper presents a method to assist the indexation of digitized Syriac manuscripts. Syriac belongs to the Aramaic branch of Semitic languages, it is written from right to left, and the characters hold the feature of being intentionally tilted by an angle of approximately 45°. For the indexation purpose we propose a word spotting approach that should locate all the occurrences of a certain query word image. The method is based on a selective sliding window technique from which directional features are extracted. Matching between features is done using Euclidean distance correspondence. The proposed method does not require any prior information, it is also fully independent of a word to character segmentation algorithm, which would be extremely difficult to realize due to the tilted nature of the handwriting.

1. INTRODUCTION

The digitization project has become a world wide concern for being the most reliable tool for the preservation of ancient manuscripts. These documents having been severely degraded by multiple elements are at risk of disappearing forever. Digitizing the manuscripts will create an electronic copy of them in the form of an image. An image is a static representation of the text; it does not allow its immediate processing. However, indexation of the documents requires research by textual content which would be very tedious if done manually, the need for an automated tool to perform this task has thus considerably grown.

Our documents of interest are digitized Syriac manuscripts. Syriac belongs to the Aramaic branch of Semitic languages. The oldest Syriac documents can date back to the 1st century A.D, however the language itself is not dead, Syriac literature is still written until our current days. Syriac is written from right to left titled by approximately 45°. There are three Syriac calligraphies, the Estrangelo which is the oldest and is rather rounded, the Serto which is the most cursive and widespread, and the Nestorian which is the most elegant.

2. RELATED WORK

The research community working on Syriac documents is very restricted. Besides the works of William Clocksin [1], and [2], and some works of our own [3], no previous work has been published on Syriac manuscripts analysis and recognition. Clocksin has chosen the Estrangelo calligraphy to perform his works; he used order structure invariance for the recognition of isolated Syriac characters. He was one of the few people who attempted a character level approach rather than a word level approach. In our previous works [4], we were interested in global information for document classification based on the handwriting style. In this paper, we focus on word spotting.

Our test database consists of digitized ancient Syriac manuscripts that were provided to us by the Central Library of the Holy Spirit University of Kaslik in Lebanon, and from the scanning department of Gorgias Press and Beth Mardutho the Syriac Institute, some of which present much degradation considered as noise by the domain of image processing, and handwriting recognition. The additional intentional tilt of the handwriting will make a text to character segmentation process become almost impossible. Thus the word becomes the smallest element we could work with.

Different methods exist for handwriting recognition in ancient manuscripts, and most authors agree on a word level approach rather than a character level approach. Each group of authors and researchers has chosen a particular type of documents depending on the availability in their region.

Manmatha et al. [5], and [6], performed their works on Georges Washington's collection, they simulated a word searching engine based on word spotting. The features they used were multidimensional profile features. The matching of the extracted features between word images was performed using Dynamic Time Warping (DTW).

Leydier et al. [7], and [8], chose medieval Latin manuscripts, the purpose of their works was the extraction and recognition of words using gradient based orientation features. The matching algorithm relies on hovering pixels gradients over a search area to find a correct superposition.

The works of Terasawa et al. [9][10], were performed on Japanese manuscripts. They performed word spotting based on an Eigenspace method. The signatures are extracted from sliding windows. The relative levels of gradients in 8 main directions are computed. This leads to features that are robust to scale changes. Morphological differences between words are overcome by a DTW matching algorithm.

3. PROPOSED METHOD

The method that we propose consists on a word spotting approach based on a correspondence between directional features extracted from selective sliding windows. Unlike Terasawa et al. [9] and [10] who took into consideration all

sliding windows, we have chosen to keep only a few sliding windows after performing several eliminations. Regions of interest were possible occurrences of a query word image may be located are detected. Saliency coefficients of directional roses are extracted from sub-windows within the selected sliding windows in the pre-detected regions of interest. They are matched with the ones extracted from the query word image by Euclidean distance correspondence.

3.1 Pre-processing

We are working with digitized ancient manuscripts and they present much degradation that we consider as noise. Our test documents are written with the Serto calligraphy, they were digitized with a very low resolution of 96 dpi in reversed binary mode. The first step was to invert the binarization of the documents since black ink on a white background is more pleasant for the reader's eye. Then we performed a slope detection using the Hough transform [11], to detect the line of greater slope indicating the maximum inclination of the text lines, followed by a slope correction using a shear transformation with a shear factor the tangent of the angle of greater slope. Afterwards we segmented the page into individual lines using horizontal projections. Figure 1 shows a sample from our test material.

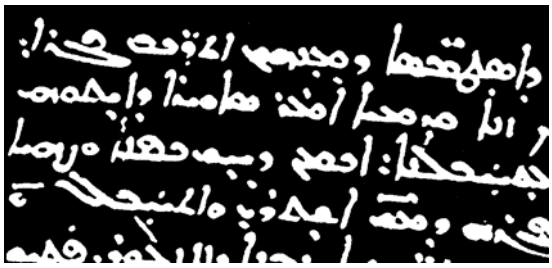

Figure 1 – Sample document written with the Serto calligraphy from Gorgias Press.

3.2 Selective sliding windows

After the pre-processing phase, we proceed to the selection of the sliding windows. In each text line, a sliding window of size 96x32 pixels is passed from left to right at a step of 1 pixel. At each step, the content of the sliding window is analysed, and it will be retained only if it responds to the following criteria:

- A minimum prefixed black pixel density.
- A centre of gravity of the black pixels that has moved significantly along the x axis compared to its predecessor.
- Not covering more than half its preceding window.

The size of the sliding window has been chosen while taking into consideration the thickness of the handwriting and the average height of the text lines. Figures 2, and 3 show an example of a word and it's retained sliding windows.

Figure 2 – A sample word.

Figure 3 – Retained sliding windows for the word in Figure 2.

3.3 Directional roses

The retained sliding windows are divided into 12 sub-windows of size 16x16 pixels. We compute the autocorrelation function in each sub-window. The obtained patterns represent the main direction in each quadrant of the sub-window. We represent this information in a directional rose of 8 directions. Each of the 12 sub-windows is represented by a signature of 8 values, resulting in a feature vector of 96 values for a retained sliding window. Figure 4 shows the 12 sub-windows of a certain selected sliding window from the word shown in Figure 3 and their autocorrelation functions.

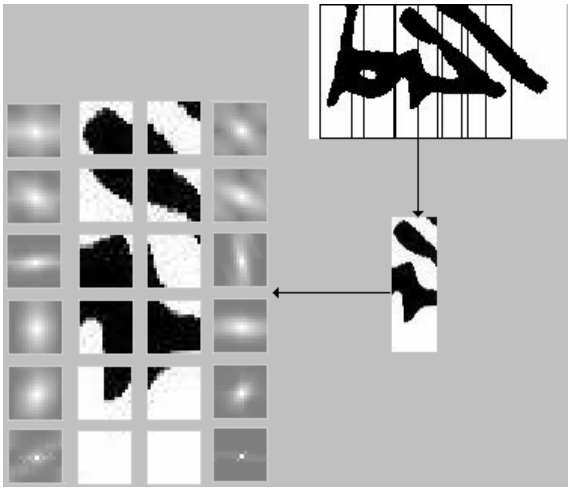

Figure 4 – Autocorrelation functions of the sub-windows.

3.4 The matching algorithm

Once the signatures of all the selected windows are extracted, they are compared to the ones extracted from the query word image. The most similar ones are detected and the regions having an agglomeration of sliding windows similar to those of the query word image are considered a possible match. However a decision based only on this criterion is not very reliable as the database or the search corpus increase in size a large number of similar agglomerations may exist resulting in possible yet incorrect matches for the query word image.

We proceed by a pre-location phase of regions of interest where possible occurrences may be located, the detection of regions of interest is done by studying the movement of the centre of gravity of the sliding windows of the query word image and then plotting it, the positions of the centres of gravity of the sliding windows from the search corpus are also plotted and then the portions of graph that are the most similar to the query plot are detected, the similarity measure is done by minimum Euclidean distance. A region of interest correctly detected is to be considered as a possible match for the query word image, also a mistaken region of interest will result in an incorrect but possible match. Figure 5 shows the directional roses representing each of the 12 sub-windows.

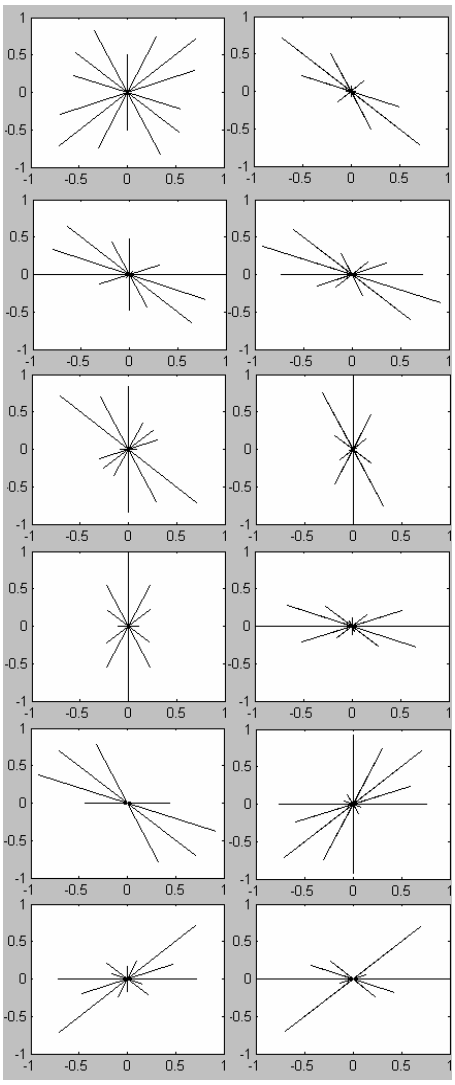

Figure 5 – Directional roses of the sub-windows shown in Figure 4.

By combining both of the preceding steps together we have performed a double elimination. The first will eliminate the agglomerations of incorrect but potentially matching windows that are not located inside the regions of interest, and the second will eliminate the regions of interest that are mistaken for occurrences if there are no sufficient windows matching those of the query word image inside them. Even though some incorrect matches may remain, their allure is not very different from the one of the query word image.

4. THE OBTAINED RESULTS

4.1. Overall results

Some frequent words were chosen as sample query word images. A frequent word is a word, not a preposition, which occurred several times in the corpus in different pages. Prepositions by their nature of usage are the most frequent words in any text; however, they offer no valuable contribution for indexing purposes. The occurrences of these frequent words being situ-ateeted in different pages; our purpose is to locate all these occurrences throughout the corpus. Even though a book is written by the same scribe, the way the same word is written differs from one page to another. Some of these differences are:

- The thickness of the ink, some pages present a thicker handwriting than others. This may be due either to the angle which the pencil or the brush is held, or to their sharpness, or to the quantity of ink available, or to the degree of absorption of the paper, or to fading agents that influenced pages more than others.
- The spacing between words, in some pages words may look closer to each other than other pages.
- The line spacing differs from one page to another.
- The length and extent of the word itself may differ not only from one page to another but also in the same page, in an attempt to justify the writing, the scribe may stretch a word more than another.
- The position of the handwriting, after de-sloping relatively to the middle of the text line.

All these differences, along with the usual differences such as the mood of the scribe, a shiver in his hand, the conditions surrounding him, and some differences introduced by the process of digitization, will influence on the similarity measures between a query word and its occurrences. Table 1 shows examples of variations in the writing of the word in Figure 2.

Table 1 – Some occurrences of the word in Figure 2 showing the variations in its writing.

A single image of a frequent word is used to locate all its occurrences in the different pages of the corpus, the diversity created by the conditions mentioned here above will prove the ability of our method to overcome these differences.

The obtained results by combining the directional signatures of the sliding windows, and the pre-selection of regions of interest are very promising, our method proved the ability to locate all the occurrences of our query word images. Some

false answers were kept by the method; these words although different in meaning from the query word image, were quite similar to it in the drawing of the letters.

Figure 6 shows the pre-detection of regions of interest where possible occurrences of the word image in Figure 2 may be located in a sample image from the test corpus from Gorgias Press. The regions of interest do point out the obvious locations of the occurrences of our query word image; however they also point out some incorrect locations. The similarity measure between regions of interest relies on the similarity in the plots of the centres of gravity of the black density regions, it is clear that a conclusion based only on this assumption is not enough.

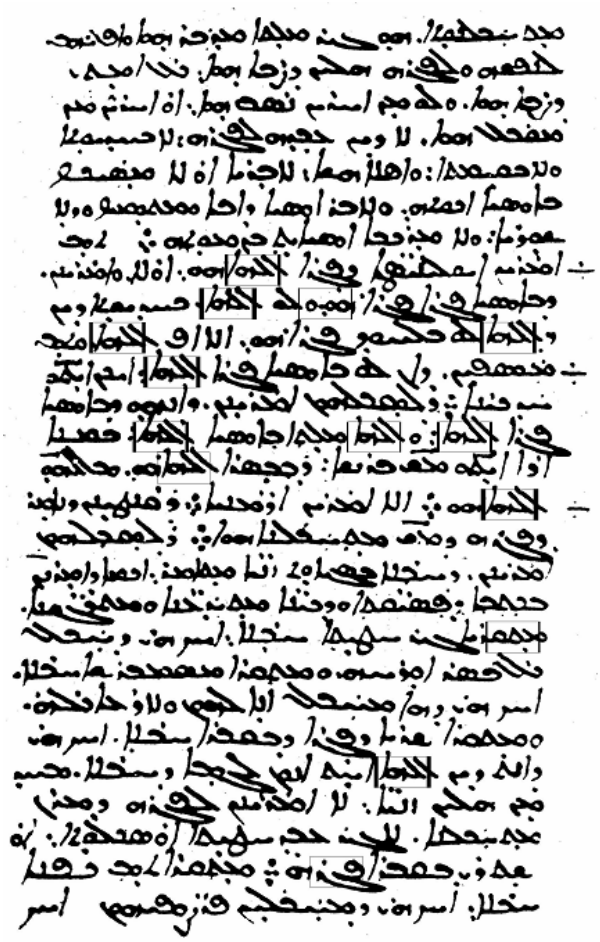

Figure 6 – Pre-detection of regions of interest for the word shown in Figure 2 in a sample test page.

Figure 7 shows the results after combing the pre-detection of regions of interest with the retained sliding windows while searching for all the occurrences of the word image in Figure 2 on a sample page from the test corpus from Gorgias Press. The positions of the regions of interest help reduce the ambiguity by only concentrating the chosen windows in these particular regions. The search for the occurrences of a word becomes easier and less error prone than a manual search relying only on eyesight

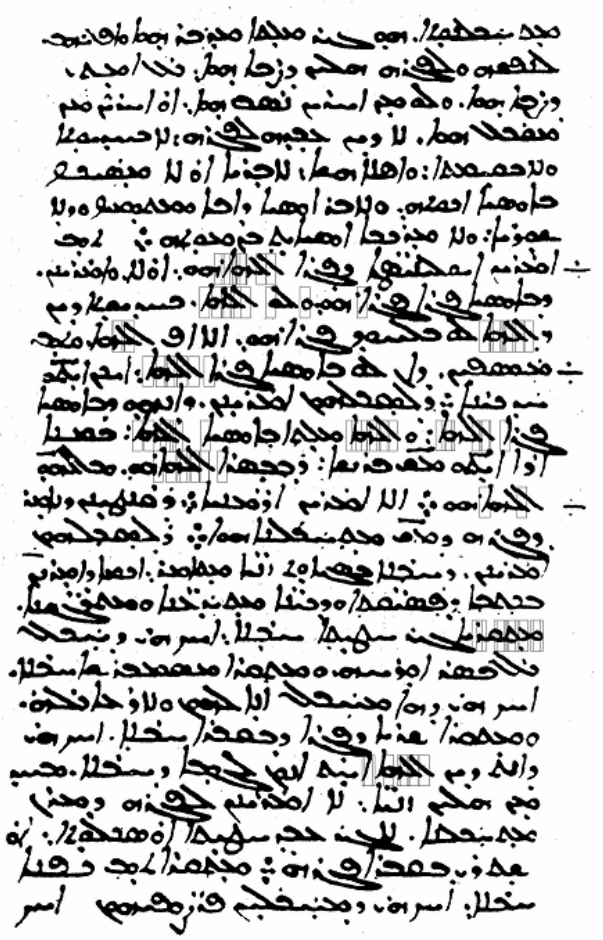

Figure 7 – Results after combining the pre-detection of regions of interest with the selected sliding windows.

4.2. Results evaluation

It is a bit difficult at the moment to present quantitative evaluation of our results since we don't have a real ground truth for our documents. Such a ground truth is not easy to build; the counting of the occurrences of query word images was done manually. If the process of counting is to be performed automatically, we need to perform some recognition and increment a counter, so we recognized to count, but to recognize the words that we are looking for, we need to know where they are located and how many exist, then perform the recognition, so we counted to recognize. However some statistics may be retrieved, by counting the number of existing occurrences, how many were found, how many were missed, and how many false occurrences were retained. We chose some frequent words to show these statistics. Figures 8, 9, and 10 are examples of frequent words.

Figure 8 – A sample frequent word.

Figure 9 – A second sample frequent word.

Figure 10 – A third sample frequent word.

Table 2 summarizes the statistical measures mentioned above for each of the frequent words examples. Our evaluation dataset is a 10 pages corpus of 1813 words.

Word	Total occur-rences.	Found occur-rences.	Missed occur-rences.	False occur-rences.
Word 1.	14	14	0	1
Word 2.	13	13	0	3
Word 3.	7	7	0	1

Table 2 – Statistics concerning the performance of the method.

The number of false occurrences is small compared to the number of existing words. It is true that these false occurrences are sometimes different in meaning from the intended query word, but they show similarity in the drawing of the letters. In some cases, a false occurrence may be a derivative of the query word image, since almost all grammatical functions in Syriac are written as prefixes and suffixes instead of separate words [1], and [2]. If we spot the exact word, or one of its derivations the initial meaning is found and indexation is achieved, the false occurrence becomes valid.

Figures 11, 12, and 13 show the first, second, and third false occurrences of query word 2. All three false occurrences are very similar and even have letters in common. We notice that the false occurrence in Figure 11 differs only by its first letter from the query word, so it can be a derivation with a different prefix, we also notice that the false occurrence in Figure 12 differs only by its last letter from the query word, so it can be a derivation with a different suffix.

Figure 11 – First false occurrence of query word 2.

Figure 12 – Second false occurrence of query word 2.

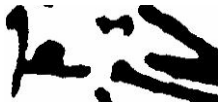

Figure 13 – Third false occurrence of query word 2.

5. CONCLUSION

The word spotting method described in this paper proves being reliable in finding all the occurrences of a query word image. The obtained results, although including some false occurrences, are enough for indexing purposes. The method is fully independent from a word to character segmentation algorithm, very difficult to realize due to the documents condition. Arabic handwriting being very similar to its Syriac ancestor can benefit from this method for indexation.

REFERENCES

[1] W. F. Clocksin, and P.P.J. Fernando, "Towards automatic transcription of Syriac handwriting", IEEE Proceedings of the International Conference on Image Analysis and Processing (ICIAP'03), Italy, pp. 664-669, Sept. 2003.

[2] W. F. Clocksin, "Handwritten Syriac character recognition us-ing order structure invariance", IEEE Proceedings of the International Conference on Pattern Recognition (ICPR'04), UK, pp. 562-565, Aug. 2004.

[3] P. Bilane, S. Bres, and H. Emptoz, "Local orientation extraction for Word Spotting in Syriac manuscripts", in Proceedings of the International Conference on Image and Signal Processing (ICISP'08), France, pp. 481-489, Jul. 2008.

[4] 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.

[5] R. Manmatha, and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'05), pp. 1212-1225, Vol. 27, Aug. 2005.

[6] T. M. Rath, and R. Manmatha, "Word spotting for historical documents", International Journal on Document Analysis and Recognition (IJDAR'07), Springer-Verlag Berlin Heidelberg, pp. 139-152, Vol. 9, Apr. 2007.

[7] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text search for medieval manuscript images", Pattern Recognition, The Journal of the Pattern Recognition Society, Elsevier, pp. 3552-3567, Vol. 40, Dec. 2007.

[8] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Omnilingual segmentation-free word spotting for ancient manuscripts indexation", IEEE Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'05), Korea, pp. 533-537, Aug. 2005.

[9] K. Terasawa, T. Nagasaki, and T. Kawashima, "Eigenspace method for text retrieval in historical document images", IEEE Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'05), Korea, pp. 437-441, Aug. 2005.

[10] K. Terasawa, and Y. Tanaka, "Locality sensitive pseudo-code for document images", IEEE Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'07), Brazil, pp. 73-77, Sep. 2007.

[11] R. O. Duda, and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures". Published in the ACM Communications, Vol. 15, pp. 11-15, Jan. 1972.