Local Orientation Extraction for Wordspotting in Syriac Manuscripts

P. Bilane, S. Bres, and H. Emptoz

LIRIS, INSA-Lyon, F-69621, France
{petra.bilane, stephane.bres, hubert.emptoz}@insa-lyon.fr

Abstract. This paper presents a contribution to Word Spotting applied for digitized Syriac manuscripts. The Syriac language was wrongfully accused of being a dead language and has been set aside by the domain of handwriting recognition. Yet it is a very fascinating handwriting that combines the word structure and calligraphy of the Arabic handwriting with the particularity of being intentionally written tilted by an angle of approximately 45°. For the spotting process, we developed a method that should find all occurrences of a certain query word image, based on a selective sliding window technique, from which we extract directional features and afterwards perform a matching using Euclidean distance correspondence between features. The proposed method does not require any prior information, and does not depend of a word to character segmentation algorithm which would be extremely complex to realize due to the tilted nature of the handwriting.

Keywords: Word Spotting, orientation features, directional roses.

1 Introduction

The Syriac language belongs to the Aramaic branch of the Semitic languages. The oldest Syriac manuscripts can date back to the 1st century AD, however the literature itself flourished from the 3rd century AD onward, then began to decline in the 7th century in the face of Arabic culture [2].

The documents we are interested in are old manuscripts. Most of the time these documents present much degradation that can be interpreted as noise in the context of text extraction and recognition. If we take into account the variability of the handwriting and the fact that the segmentation of the text into letters is most often not possible, we understand why classical OCR is useless. The word is then the smallest element we can identify.

2 Related Work

Very few are the people who launched themselves in the study of Syriac manuscripts. Besides the works of William Clocksin [1], no previous work has been published on Syriac handwriting recognition. Different approaches exist for handwriting recognition

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 481 – 489, 2008.
© Springer-Verlag Berlin Heidelberg 2008

in historical manuscripts. In our previous work, we were interested in global information for document classification based on handwriting style [4]. In this paper, we focus on word spotting.

Old documents can be treated as they are or a pre-processing can be performed, like binarization to highlight the text [5][6]. Most of the approaches focus on words and not letters [7][8][10][14], because of the problems of segmentation on manuscripts data. For most of the authors, a word level approach is better than a letter level approach. In this domain of old documents study, there are a lot of existing documents. It is not always that easy to have access to them and thematic studies on documents coming from a precise origin are especially interesting because they can take advantages of some constant characteristics of the database. In this paper we focus on Syriac documents. Other researches were made on the famous French writer Flaubert corpus [9] for layout extraction using Markov fields, Manmatha [8][10] targeted manuscripts from the Georges Washington's collection and Leydier targeted Medieval Latin manuscripts [7]. The purpose of their worls is to extract and recognize words using description based on computed features. Terasawa et al. [11] also performed word spotting inspired by an Eigen space method [11] or gradient [12]. In this case, word signatures are extracted from sliding windows. The relative levels of the gradient in the 8 main directions are computed in these sliding windows. This leads to features that are robust to scale changes. To overcome the morphological differences between the words, the matching is performed using a Dynamic Time Warping (DTW) algorithm. DTW is also used in [1] to match whole words. Another segmentation-free approach which uses HMMs and statistical language models for handwritten text recognition is described in [13].

3 Proposed Method

The method that we propose consists of an eliminatory process. First we start with a preprocessing phase to select sliding windows of interest, unlike Terasawa et al. [10] who took into consideration all extractable windows. The elimination begins at this step: windows that do not respond to certain criteria are eliminated (the elimination process is further detailed in the following sub sections). Regions of interest are then detected. Afterwards, saliency coefficients of directional roses are extracted from sub windows within the regions of interest, and are matched to those extracted from the query word image using a point to point correspondence of the Euclidean distance.

3.1 Preprocessing

The material that we will be working on consists of Karshuni manuscripts that are written with the Serto calligraphy. They were supplied to us by the Central Library of the Holy Spirit University of Kaslik in Lebanon. These manuscripts date from the beginning of the 19th century and were digitized with a resolution of 300 dpi. Figure 1 shows a sample from these manuscripts.

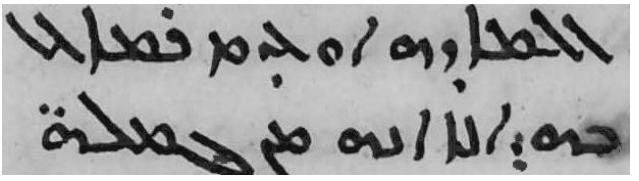

Fig. 1. Sample image from a Karshuni manuscript written with the Serto calligraphy

The manuscripts that we have contain mainly three color categories, the background which is a light yellow tone, the main text which is dark brown or rather black, and the emphasized words which are written in red. The preprocessing phase consisted converting the document in grey level preserving the contrast between the text and the background. A skew detection and correction of the line is then performed using the Hough Transform. An upper and a lower baseline detection using horizontal projections are finally performed (see Figure 2).

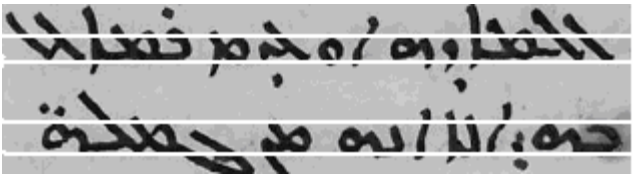

Fig. 2. Example of an upper and a lower baseline detection

3.2 Selective Sliding Windows

Once the upper and lower baselines are detected, the text line is divided into three regions (upper, middle, and lower). In each of these regions a sliding window of size 32x32 pixels is passed along from left to right at a step of one pixel. Afterwards, an analysis of the content of the window is conducted: first we reject the windows having less than a certain minimum density of black pixels; then we study the movement of the center of gravity of the black pixels. We only keep the window having a center of gravity that moved significantly along the x axis compared to its predecessor; in a third elimination, windows that cover more than half their predecessor are eliminated.

The size of the sliding windows and the allowable covering rate were chosen while taking into consideration the image resolution (300 dpi) and the thickness of the handwriting (10 pixels).

3.3 Directional Roses

The remaining windows are divided into four sub windows of size 16x16 pixels. We compute the autocorrelation function on each of these four windows. The patterns we obtain from this autocorrelation function represent the main directions in the four quadrants of the current window.

We summarize this information in a directional rose of eight directions. Each 16x16 pixels sub window is represented by a signature of eight values resulting in a total of 32 values for the current window. Figure 3 illustrates the sub window extraction process from a selected window taken from a word image from a sample line, and figure 4 shows their respective directional roses. The length of a direction is obtained by summation of the grey levels of the autocorrelation function in this direction.

To keep the most discriminative information, we only keep the relative variations of the different directions above the least represented direction, which is then set to 0. The salient direction is then normalized to 1, to reduce the influence of the dynamic of the original image. Moreover, the use of the autocorrelation function for the local signature reduces the influence of noise or degradations because it is the main structure of the sub image that influences the result. We will give more details on this specific point later in this paper.

As shown in figure 4, the directional rose highlights the dominant orientation in each sub window; the salient direction is quite obvious. This description is only based

Fig. 3. Illustration of the sub windows extraction process and their corresponding autocorrelation functions

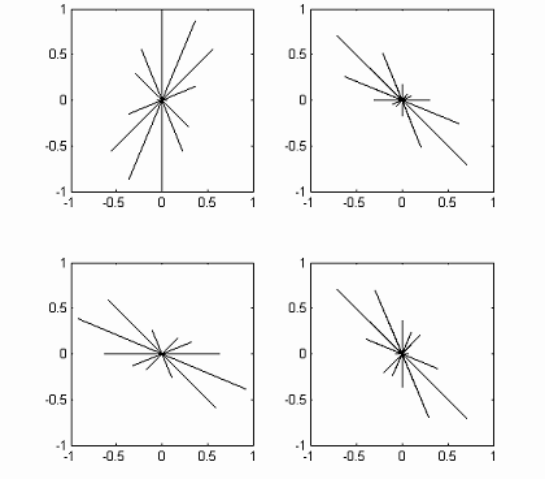

Fig. 4. Directional roses extracted from the four sub windows

on directional information. Even if this directional information is scale invariant, we do not keep this property in our signature because of the decomposition in four windows that have a defined size. However, a normalization step can be performed as a pre-processing to fit the size of the letters in the guide lines of figure 2. Reduction of size introduces no loss of accuracy in the tests we perform because the size we defined for the letters and thus for the words contains enough information. Increasing the size will lead to blurred images. This specific case is discussed latter in this paper.

3.4 The Matching Algorithm

Once the signatures of all the selected windows of each part of the query word image are extracted, the search begins in order to spot all their occurrences. They are compared to all those of the test page, the most similar ones are detected, and the region with most agglomeration of sub windows similar to those of the query word image is considered as a possible match. However, a decision based only on this assumption requires some computation time due to the great number of comparisons to perform. Moreover, as the number of lines and thus windows increases, this first simple matching is not very reliable due to superposition of matching candidates in all three parts of the line resulting in possible yet incorrect matches.

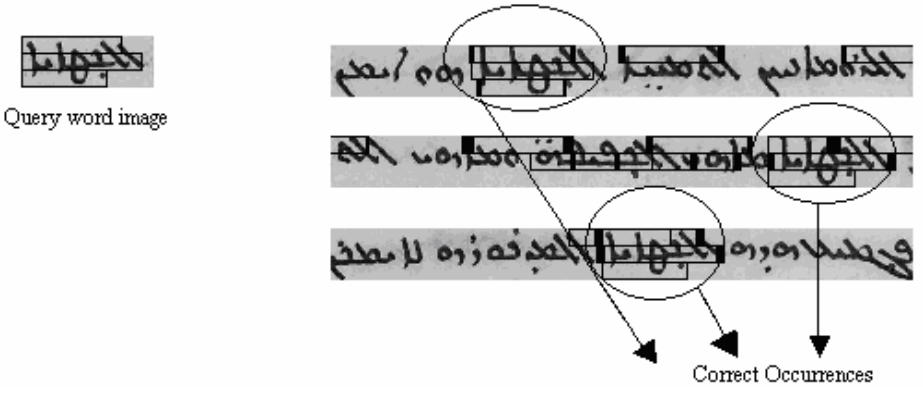

Fig. 5. Detection of regions of interest

In order to surpass the confusion problem, we proceeded by a pre detection of regions of interest where possible occurrences may be located. This was done by studying the movement of the gravity centers of the selected windows in each part of the query word image. The positions of the gravity centers are plotted, and a search is conducted to find portions of the plot that are similar to the query plot based on a minimum Euclidean distance. As a result, confusion is removed as shown in figure 6.

Normalized saliencies coefficients are extracted from the query word image and are compared to the ones extracted from the regions of interest. The matching is based on a minimum Euclidean distance, and the correct occurrences are those which have the largest number of matched coefficients in the three superposed parts.

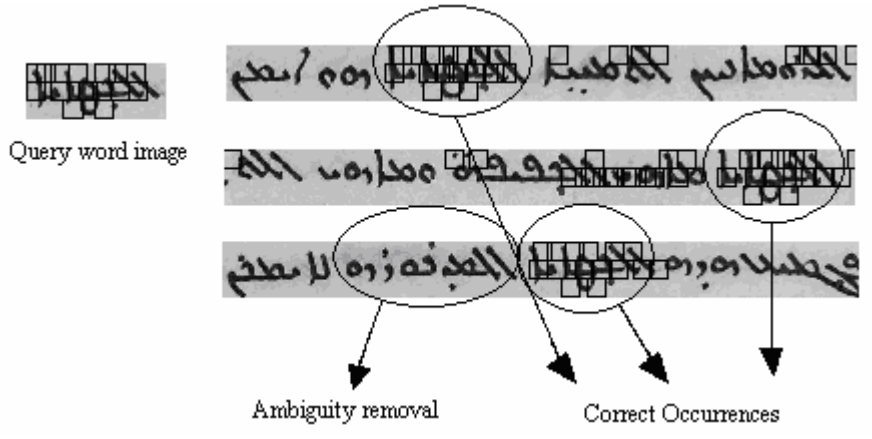

Fig. 6. Ambiguity removal

4 Results

The results we obtain using the combination of these two indicators (directional signature of the windows and pre selected area of interest using gravity center motions) are really promising because, on the tested pages so far, we always find the two or three occurrences of the query word in the top first best matches.

It is difficult to present at the moment, a quantitative evaluation of our method because we do not have yet a real ground truth on our documents. It is not that easy to build it because we have to search manually for words and their occurrences in the documents. We only did this work for some words yet and as mentioned earlier, the results are correct in every tested case so far. A more complete set of test is in progress to give more quantitative results.

Moreover, we tested our indicators in different situations and especially situations involving degraded documents. These degraded versions correspond to the two most common degradations to digitized manuscripts. The first is excessive and lossy compression, and the second is poor resolution.

4.1 Excessive Compression and Poor Resolution

Librarians and book keepers have a tendency to over compress the manuscripts images. In most cases, lossy JPEG compression is chosen. Since the degradations resulting form this compression are irreversible, many people proceed by a restoration phase that consists of a smoothing of the artifacts, sometimes even an attempt to recover the dissolved portions of the texts usually by morphological approaches as used in [16] or by active contours as attempted in [15] and [17], as a result they fall most of the time in a paradigm which is “restore to recognize and recognize to restore”. This is why many approaches fail in front of the degradations introduced by excessive JPEG compression.

The poor resolution type of degradations is introduced either by having manuscripts with large pages, and in an attempt to fit them into a standard size page, librarians and

book keepers tend to reduce the resolution of the digitizer, or just like in the preceding degradation, it is done only by fear of not having enough storage capacity for all the documents. We imitated this degradation by downscaling a test image and afterwards rescaling it back to its original resolution using cubic interpolation.

4.2 Results on Degraded Images

Figure 7 illustrates the ambiguity removal and the ability of finding all occurrences of the query word image. Just as the case for the previously degraded test image, the

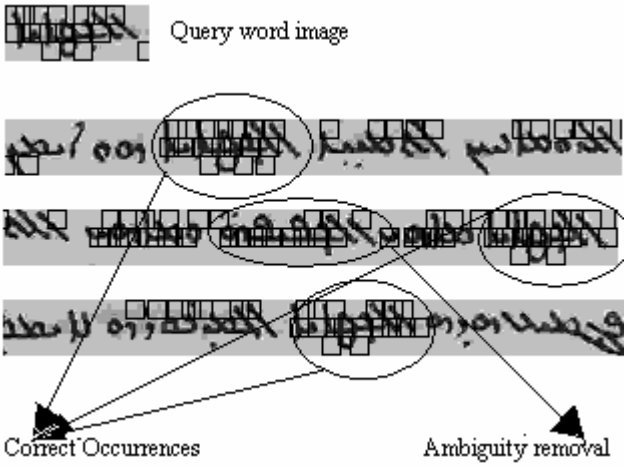

Fig. 7. Ambiguity removal on over-compressed image after interest regions extraction

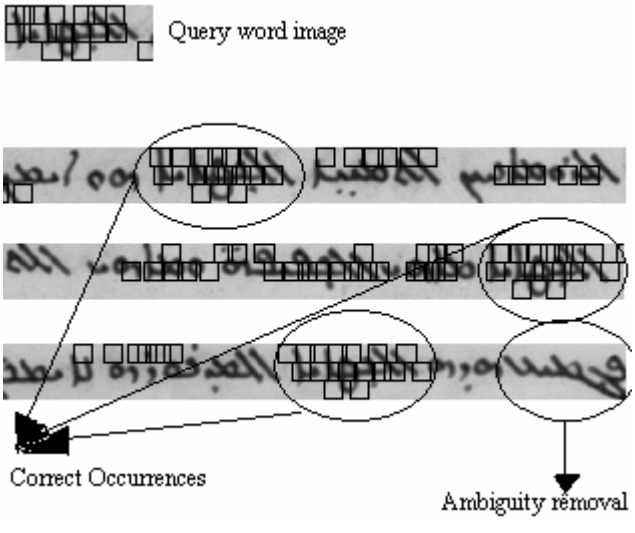

Fig. 8. Ambiguity removal

results for the rescaled test image support the robustness of the algorithm in the face of another type of degradation. Figure 8 reveals the ambiguity removal and the ability to find all occurrences of the query word image after the regions of interest detection.

The results prove the effectiveness of these features and their consistency in finding all occurrences of the query word image within reasonable iterations and processing time. The selective sliding windows, the region of interest detection, the feature extraction, and the matching algorithm were all performed exactly the same as in the word spotting in the original test image which we take now as a reference for the performance comparison

6 Conclusion

In this paper we presented a word spotting algorithm to assist the indexing of Syriac manuscripts. Our method does not require any prior information for the spotting process. It is also fully independent from any word to character segmentation algorithm. Moreover, the way the signature is computed leads to a less sensitivity to noise and degradations that are really common on that type of documents such as an excessive JPEG compression and a rescaling for low resolution documents images. An extension of this method could be used as a basis for a classification algorithm for the automatic separation of the three calligraphies.

References

[1] Balasubramanian, A., Meshesha, M., Jawahar, C.V.: Retrieval from document image collections. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 1–12. Springer, Heidelberg (2006)

[2] Clocksin, W.F., Fernando, P.P.J.: Towards automatic transcription of Syriac handwriting. In: IEEE Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP 2003), Mantova, Italy, September 2003, pp. 664–669 (2003)

[3] Clocksin, W.F.: Handwritten Syriac character recognition using order structure invariance. In: IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, pp. 562–565 (August 2004)

[4] Eglin, V., Bres, S., Rivero, C.: Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts. International Journal on Document Analysis and Recognition (IJDAR 2007) 9, 101–122 (2007)

[5] Gatos, B., Pratikakis, I., Perantonis, S.J.: An adaptative binarization technique for low quality historical documents. In: Marinai, S., Dengel, A. (eds.) DAS 2004. LNCS, vol. 3163, pp. 102–113. Springer, Heidelberg (2004)

[6] Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization., Pattern Recognition. The Journal of the Pattern Recognition Society 39, 317–327 (2006)

[7] Leydier, Y., Lebourgeois, F., Emptoz, H.: Text search for medieval manuscript images, Pattern Recognition. The Journal of the Pattern Recognition Society 40, 3552–3567 (2007)

[8] Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2005) 27, 1212–1225 (2005)

[9] Nicolas, S., Paquet, T., Heutte, L.: Extraction de la structure de documents manuscrits complexes à l'aide de champs Markoviens. In: Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 2006), pp. 124-129 (September 2006)
[10] Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal on Document Analysis and Recognition (IJDAR 2007) 9, 139–152 (2007)
[11] Terasawa, K., Nagasaki, T., Kawashima, T.: Eigenspace method for text retrieval in historical document images. In: IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, Korea, August 2005, pp. 437–441 (2005)
[12] Terasawa, K., Tanaka, Y.: Locality sensitive pseudo-code for document images. In: IEEE Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, September 2007, pp. 73–77 (2007)
[13] Vinciarelli, A., Bengio, S., Bunke, H.: Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI 2004) 26, 709–720 (2004)
[14] Weihua, H., Tan, C.L., Sung, S.Y., Xu, Y.: Word shape recognition for image-based document retrieval. In: IEEE Proceedings of the International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece, October 2001, pp. 1114–1117 (2001)
[15] Allier, B., Emptoz, H.: Degraded character image restoration using active contours: A first approach. In: Proceedings of the ACM Symposium on Document Engineering, Virginia, USA, pp. 142–148 (2002)
[16] Zheng, Q.J., Kanungo, T.: Morphological degradation models and their use in document image restoration. In: IEEE Proceedings of the International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece, October 2001, pp. 193–196 (2001)
[17] Allier, B., Bali, N., Emptoz, H.: Automatic accurate broken character restoration for patrimonial documents. International Journal on Document Analysis and Recognition (IJDAR 2006) 8, 246–261 (2006)