Spot Words in Printed Historical Arabic Documents

Fattah Zirari[1,2], Abdel Ennaji[1], Driss Mammass[2], and Stéphane Nicolas[1]

[1] LITIS Laboratory, University of Rouen, Rouen, France
[2] IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco
zirari_fattah@yahoo.fr

Abstract. Libraries contain huge amounts of arabic printed historical documents which cannot be available on-line because they do not have a searchable index. The word spotting idea has previously been suggested as a solution to create indexes for such a collecton of documents by matching word images. In this paper we present a word spotting method for arabic printed historical document. We start with word segmentation using run length smoothing algorithm. The description of the features selected to represent the words images is given afterwards. Elastic Dynamic Time Warping is used for matching the features of the two words. This method was tested on the arabic historical printed document database of Moroccan National Library.

Keywords: Segmentation, text/no text Separation, Document Image, Graph, modelization, structural analysis.

1 Introduction

Historical library collections across the world hold huge numbers of historical printed documents. By digitizing these documents, their content can be preserved and made available to a large community via the Internet or other electronic media. Due to the large volume, it is imperative to provide fast and efficient collection access methods to this document analysis. However, the current tools for indexing and searching in large databases are not appropriate to deal with this type of data. Moreover, the use of OCR-based methods was proved an expensive option of the computational point of view [3]. An interesting alternative is the group of methods that aim to make possible the word spotting in document images without using OCR. Word Spotting is an approach to find and retrieve all the occurrences of a query word in a set of documents. Research in word spotting on the Latin documents was initially suggested by Manmatha in [5] and [6] and has produced a number of publications that offers algorithms and features for the approach [9], [1], [10], [8]. word spotting in the Arabic script dates back several years when we unfortunately offered little work and solutions for Arabic script.

You et al. [14] presented a hierarchical Chamfer matching scheme as an extension to traditional approaches of detecting edge points, and managed to detect interesting points dynamically. They created a pyramid through a dynamic thresholding scheme to find the best match for points of interest. The same hierarchical approach was used

A. Elmoataz et al. (Eds.): ICISP 2014, LNCS 8509, pp. 289–296, 2014.

by Borgefors [2] to match edges by minimizing a generalized distance between them. Rothfeder et al. [11] and Srihari et al. [13] presented a system for spotting words in scanned document images for three scripts: Devanagari, Arabic, and Latin. Their system retrieved the candidate words from the documents and ranked them based on global word shape features.

Saabni and El-Sana [12] segmented the documents into Arabic word-parts, ; they used Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs) for matching in two different systems, and then additional strokes were used by means of a rulebased system to determine the final match. Moghaddam et al. [7] presented an Arabic word spotting system that is based on shape matching, They extracted the connected components from the documents and then created their library of word-parts using an Euclidean distance technique and DTW. Then word-parts were clustered into metaclasses to improve the accuracy and reduce the computational complexity. Both approaches [12] and [7] searched for word-parts rather than words, and they were tested on historical Arabic documents.

Most of the work in the field of word spotting has been done on handwritten manuscripts. The reason for that mainly being the irregular writing styles that prevent commercial OCRs from achieving higher recognition rates. Printed document images are usually considered ‘OCR friendly’, as OCR software achieves relatively better on printed documents compared to handwritten ones. But if the printed text is from old historical ancient documents, then OCR results on these document images degrade significantly. In that case, word spotting comes as a lucrative alternate of OCR as B. Gatos in [4] remarked “OCR is a very difficult problem to solve, especially for historical printed documents”.

As part of this paper, we propose a word spotting in the arabic historical printed document. It starts with word segmentation using run length smoothing algorithm. The description of the features selected to represent the words images is given afterwards. Elastic Dynamic Time Warping is used for matching the features of the two words.

This paper is organized as follows. In Section 2, we first describe the steps used by our approach. And then, the experimental results are given and discussed in Section 3. Finally, the last section concludes.

2 Proposed Method

The proposed word spotting methodology receives as input the word image query and the document image and produces as output the word spotting results which correspond to a set of rectangular areas of document image which delimits the word images that match the word image query. It consists of several distinct steps: (a) words segmentation in document image based on a RLSA smoothing [16]; (b) word-based feature extraction of document image as well of image query, (c) Elastic Dynamic Time Warping is used for matching the features of the two words. The proposed methodology is detailed in this section while a flowchart is presented in figure 1.

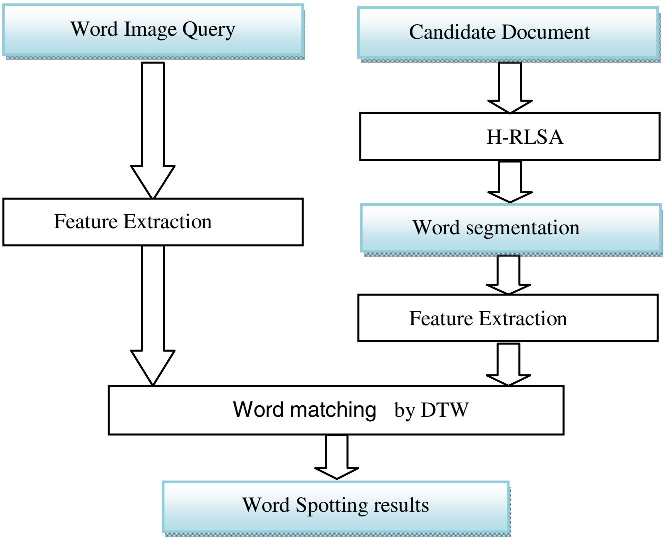


Fig. 1. The flowchart of the proposed word spotting methodology

2.1 Word Segmentation

RLSA has been used previously in text/non-text segmentation. Wong et al. [16] used a combination of RLSA in horizontal and vertical direction to segment blocks of text and non-text. Afterwards, the text blocks are analyzed for the extraction of words. In our case, we proceed to segment the document image directly into word. To do so, we apply RLSA only in horizontal direction. The basic RLSA is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's. The algorithm transforms binary 'x' into an output 'y' according to the following rules:

1. 0's in x are changed to 1's in y if the number of adjacent 0's between two 1's is less than or equal to a predefined limit C.
2. 1's in x are unchanged in y.

For example, with C = 4 the sequence x is mapped into y as follows:

x : 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 01
y : 0 0 0 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1

H-RLSA has a same effect that dilation of black areas in horizontal direction. The word-parts in a word are dilated and get blacked/connected to the other word-parts of the same word. The distance between two neighborings word-parts of two adjacent words is greater than the value of C, so that gap remains there meaning that each word becomes a connected component.

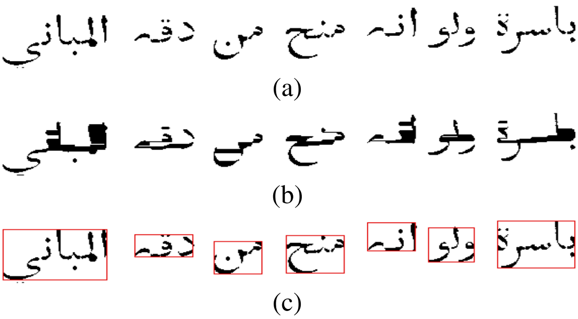


Fig. 2. Word segmentation process; a) Original image; b) H-RLSA image with C=15; c) Connected components (words)

2.2 Feature Extraction

We describe a selection of the more useful features each having a length equal to the width (height) of that particular word. It means that for different word, the length of the sequences may be different depending on their widths (heights), some of which have been previously reported in the literature (e.g. see [15]). These feature sequences include vertical/horizontal histogram, upper word profile, lower word profile, black-non-black transitions and transitional vector. These feature vectors are built for each pixel column and row of a word image. All of the features have been extracted on the binarized image. All feature plots presented in this work are extracted directly from the image (66x46) in figure 3.

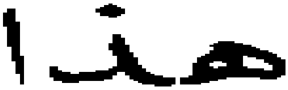


Fig. 3. Preprocessed example image

Vertical/Horizontal Histogram. Number of black pixels in each column (row) of a binarized word image. Figure 4 shows the normalized vertical (horizontal) histogram of the binarized word image.

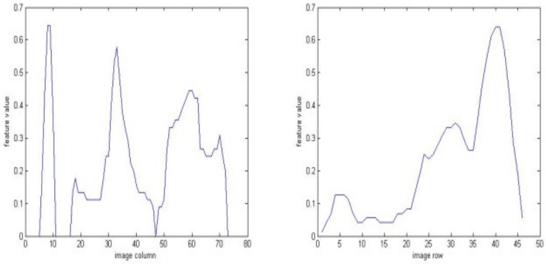


Fig. 4. Vertical histogram; Horizontal histogram

Upper/Lower Word Profile. Upper/lower word profile features are computed by recording, for each image column, the distance from the upper/lower boundary of the word image to the closest "black" pixel. If an image column does not contain black, the feature value is computed by linear interpolation between the two closest defined values. Figure 5 shows two typical profiles (feature values are inverted).

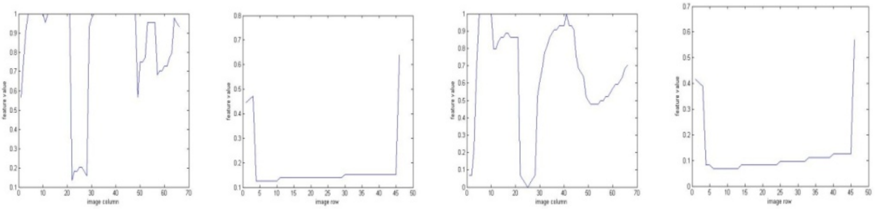


Fig. 5. Vertical lower word profile; Horizontal lower word profile; Vertical upper word profile; Horizontal upper word profile

Background to Black Transitions. This feature (see Figure 6) records, for every image column, the number of transitions from the background to a "black" pixel (determined by thresholding). The value of that threshold comes to be 10 since it is the maximum number of transitions we may have in any row (column) of a word of figure 3.

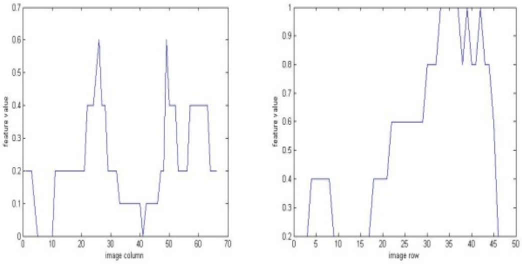
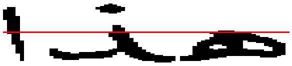
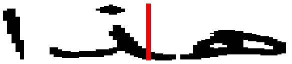


Fig. 6. Normalized number of background to-black transitions feature



0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 0

(a)



0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 00 0

(b)

Fig. 7. (a) Mid row black to non-black transitional sequence; (b) Mid column black to non-black transitional sequence

Transitional vector. For the central row of the word image, we find a transitional sequence accounting for all the black/non-black transitions. A ‘1’ is placed for every transition from black to non-black or non-black to black, and a ‘0’ for all the non-transitions in that row.

2.3 Word Matching

To match two words, we use the Levenshtein Edit distance [17]. At the word level, a non-linear elastic matching is more appropriate to compute the word similarity. Some deformations may occur in the word; therefore, the length of the vector sequences may be different for occurrences of the same word. Elastic matching cannot take into account the nonlinear stretch or compression of words.

A word is represented by a sequence of feature vectors: $X = (x_1 \ldots x_m)$ where x_i is a 5-dimensional vector. To determine the *DTW* distance between two sequences, X and $Y=(y_1 \ldots y_n)$, $D(m,n)$ is computed as:

$$\mathrm{D(i,j)} = \min \left\{ \begin{matrix} \mathrm{D(i,j-1)} \\ \mathrm{D(i-1,j)} \\ \mathrm{D(i-1,j-1)} \end{matrix} \right\} + d(x_{i,}y_j) \tag{1}$$

$d(x_i, y_j)$, is the Euclidean distance in the feature space, i varies from 1 to m, j from 1 to n. The distance between two words is equal to $D(m,n)$ divided by the number of steps of the warping path. Two words are similar if their matching distance is lower than an empirically fixed threshold.

3 Experiment and Results

We tested our methodology on an Arabic historical printed documents database provided by Moroccan National Library. The documents images suffer from several problems such as degradations and typesetting imperfections. We selected 40 documents from this database and manually marked 12 keywords. These keywords are semantically significant and frequently repeated in the database. In the selected documents we marked 1616 instances of all keywords. Then, we applied the proposed word spotting methodology. A word spotting result is illustrated in Figures 8 and 9.

Let N be the total number of word instances for every keyword, M the total number of detected keyword instances and Corr the correctly detected keyword instance. Evaluation metric of recall (RC), precision (PR) and F-measure (FM) are defined as follows:

$$RC = \frac{\mathrm{Corr}}{N} 100\% \qquad PR = \frac{\mathrm{Corr}}{M} 100\% \qquad FM = \frac{2*\mathrm{RC}*\mathrm{PR}}{RC+PR}$$

Table 1 presents the word spotting results for the 12 keywords in terms of recall and precision as well the F-measure. As it can be observed, we can achieve high recall rates (95.75% on average) while keeping the precision on acceptable levels (96.47% on average) resulting to an F-measure equal to 96.04% on average.

Table 1 shows also that vertical features, i.e. the features extracted on the columns and feeding an horizontal sequential model, outperforms the horizontal features extracted on the rows. This difference in performance can be explained by the nature of the DTW algorithm and by the lower stability of writing in the horizontal direction.

Table 1. Word spotting results in terms of recall and precision for the 12 keywords used in the experiments

		Vertical features					**Horizontal features**				
	N	*M*	*Corr*	*RC(%)*	*PR(%)*	*FM(%)*	*M*	*Corr*	*RC(%)*	*PR(%)*	*FM(%)*
في	427	396	389	91,10	98,23	94,53	345	323	75,64	93,62	83,67
هذا	78	75	71	91,02	94,66	92,80	84	65	83,33	77,38	80,24
مجلة	45	44	44	97,77	100	98,87	49	13	28,88	26,53	27,65
من	406	388	387	95,32	99,74	97,47	397	360	88,66	90,68	89,65
المغرب	46	43	42	91,30	97,67	94,37	52	29	63,04	55,76	59,17
الى	134	134	134	100	100	100	125	112	83,58	89,60	86,48
ان	169	162	161	95,26	99,38	97,27	143	92	54,43	64,33	58,96
على	177	177	177	100	100	100	159	141	79,66	88,67	83,92
الذي	48	50	47	97,91	94	95,91	62	45	93,75	72,58	81,81
اذا	43	47	42	97,67	89,36	93,33	61	38	88,37	62,29	73,07
العالم	12	13	11	91,66	84,61	87,99	28	10	83,33	35,71	49,99
محمد	31	31	31	100	100	100	47	30	96,77	63,82	76,91
Total	**1616**	**1543**	**1536**	**95,75**	**96,47**	**96,04**	**1552**	**1258**	**76,58**	**68,41**	**70,96**

4 Conclusion

We have proposed a word spotting system for Arabic historical documents that makes use of the run length smoothing algorithm for word segmentation and the word-based feature extraction of image as well of image query. Thereby, Elastic Dynamic Time Warping is used for matching the features of the two words.

Our method was tested on document database of Moroccan National Library. These results are promising but are still preliminary ones. Our ai mis now to optimize our approach to be able to analyze the entire database which contains nearly 2950 documents of Moroccan National Library. Our system must allow us to generate semi-automatically the ground truth for a part of this database. Then the annotated documents will serve as a learning database to train reliable classifiers.

Acknowledgment. We would like to acknowledge the financial support of our project by the France-Morocco-Hubert Curien Program-PHC-Volubilis n° MA/10/233.

References

1. Adamek, T., O'Connor, N.E., Smeaton, A.F.: Word matching using single closed contours for indexing handwritten historical documents. IJDAR 9, 153–165 (2007)
2. Borgefors, G.: Hierarchical Chamfer Matching: Aparametric Edge Matching Algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 10(6), 849–865 (1988)
3. Doermann, D.: The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding (CVIU) 70(3), 287–298 (1998)
4. Gatos, B., Pratikakis, I.: Segmentation-free word spotting in historical printed documents. In: 10th International Conference on Document Analysis and Recognition (2009)
5. Manmatha, R., Han, C., Riseman, E.M.: Wordspotting: A New Approach to Indexing Handwriting. In: Conference on Computer Vision and Pattern Recognition (CVPR), p. 631 (1996a)
6. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: 1st ACM Internationall Conference on Digital Libraries (1996b)
7. Moghaddam, R., Rivest-Hénault, D., Cheriet, M.: Restoration and Segmentation of Highly Degraded Characters using a Shape Independent Level Set Approach and Multi-level Classifiers. In: Proc. ICDAR 2009, Barcelona, Spain, pp. 828–832 (2009)
8. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Seventh International Conference on Document Analysis and Recognition (ICDAR), p. 218 (2003)
9. Rath, T.M., Manmatha, R.: Word spotting for historical documents. IJDAR 9, 139–152 (2007)
10. Rothfeder, J.L., Feng, S., Rath, T.M.: Using corner features correspondences to rank word images by similarity. In: Conference on Computer Vision and Pattern Recognition, pp. 30–35, USA (2003)
11. Rothfeder, J.L., Feng, S., Rath, T.M.: Using Corner Feature Correspondences to Rankword Images by Similarity. In: Proc. DIAR 2003, Madison, WI (June 2003)
12. Saabni, R., El-Sana, J.: Keyword searching for Arabic handwritten documents. In: Proc. 11th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), pp. 271–277 (2008)
13. Srihari, S., Srinivasan, H., Huang, C., Shetty, S.: Spotting Words in Latin, Devanagari and Arabic scripts. Vivek: Indian Journal of Artificial Intelligence 16(3), 2–9 (2003)
14. You, J., Pissaloux, E., Zhu, W., Cohen, H.: Efficient Image Matching: A hierarchical Chamfer Matching Scheme via Distributed System. Real-Time Imaging 1(4), 245–259 (1995)
15. Kolcz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.V.: A line-oriented approach to word spotting in handwritten documents. Pattern Analysis & Applications, pp. 153–168 (2000)
16. Wong, K.Y., Casey, R.G., Wahi, F.M.: Document analysis system. IBM Journal of Research Development 26, 647–656 (1982)
17. Wagner, R.A., Fischer, M.J.: The string-to-string correction Problem. Journal of ACM 21, 168–173 (1974)