Keyword Matching in Historical Machine-Printed Documents Using Synthetic Data, Word Portions and Dynamic Time Warping

T. Konidaris, B. Gatos, S.J. Perantonis and A. Kesidis

Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
{tkonid,bgat,sper,akesidis}@iit.demokritos.gr

Abstract

In this paper we propose a novel and efficient technique for finding keywords typed by the user in digitised machine-printed historical documents using the Dynamic Time Warping (DTW) algorithm. The method uses word portions located at the beginning and end of each segmented word of the processed documents and try to estimate the position of the first and last characters in order to reduce the list of candidate words. Since DTW can become computational intensive in large datasets the proposed method manages to significantly prune the list of candidate words thus, speeding up the entire process. Word length is also used as a means of further reducing the data to be processed. Results are improved in terms of time and efficiency compared to those produced if no pruning is done to the list of candidate words.

1. Introduction

Machine-printed documents processing usually involves an Optical Character Recognition (OCR) step. This is not the case in historical documents due to a number of constraints such as low paper quality, typesetting imperfections and low print contrast. These are some of the constraints that affect the performance of OCR and therefore, the segmentation of these documents to individual characters leads to poor results. However, OCR in not the only solution when processing machine-printed documents. The global or segmentation-free approach where the entire word is treated as a single entity assists the processing of these documents. A segmentation-free approach is followed in [5] where a keyword spotting on historical printed documents is performed and results are improved by a feedback process. In [7] historical handwritten documents are indexed using a scale space approach. In [6] a holistic word recognition approach is followed in low quality historical documents.
Dynamic Time Warping (DTW) and its variations has been used in a variety of applications with signature identification [11] [1] being a popular one. DTW is also used in word matching applications in historical documents. In [10] [9] word matching is performed using DTW in a set of different features.
The proposed methodology follows the segmentation-free approach where the processing of the documents takes place at word level rather than character level. The proposed methodology is generic since it allows the processing of several types of machine-printed historical documents, independently of their language, print style or font size.
The proposed methodology involves a word mathing process between synthetic word images created by their ASCII equivalents with word images segmented from the processed document collections. The aim is to find these keywords in large document collections. The word matching process is based on the DTW algorithm. The proposed method uses a pruning process in order to decrease the size of the list of candidate words and therefore speed up the entire procedure. Pruning is based on the word length of the words and on their first and last portions. Words larger than a certain threshold are excluded from the list of candidate words to be processed. The pruning based on the word portions tries to estimate the position of the first and last characters of the words in order to check their similarity and eventually determine their validity. Experiments showed that the proposed method managed not only to speed up the entire process but also to improve the results compared with the non-pruning version of the method. Figure 1 illustrates the distinct steps of the proposed method.

The paper is organized as follows: Section 2 concerns the preprocessing scheme of the historical documents and Section 3 describes that segmentation process. Section 4 fo-

978-0-7695-3337-7/08 $25.00 © 2008 IEEE
DOI 10.1109/DAS.2008.64

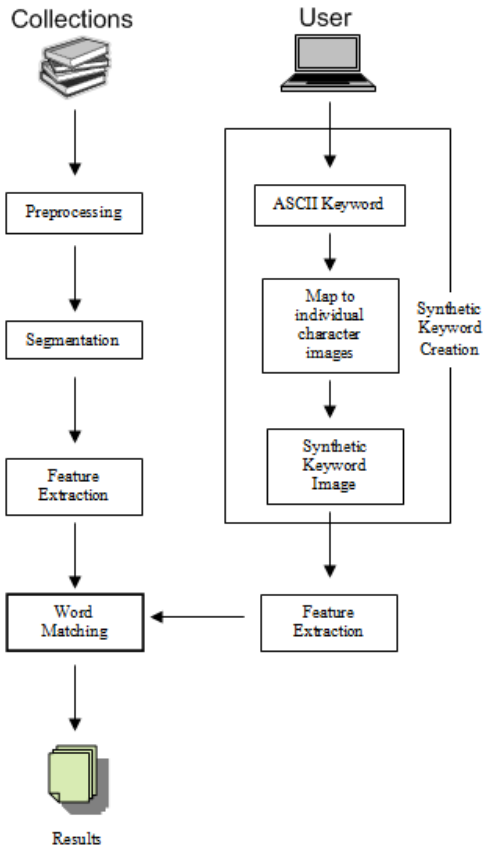


Figure 1. The distinct steps of the proposed method.

cuses on the synthetic image creation of the keywords to be searched in the documents. Feature extraction is described in Section 5 and the pruning process is described in Section 6. Section 7 concerns the word matching process. In Section 8 the results of the proposed methodology are given while Section 9 draws the conclusions.

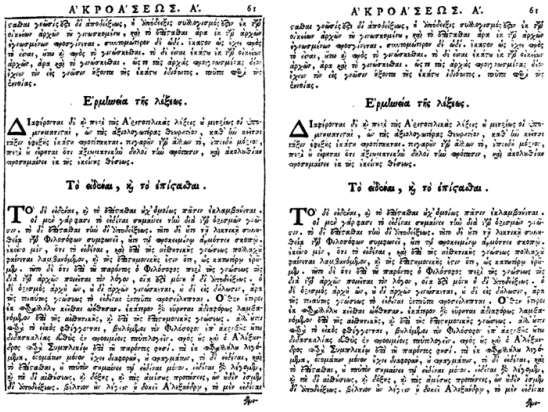
2. Preprocessing

The processed historical documents are characterized by low image quality and therefore a preprocessing procedure is essential in order to improve the quality of these documents. The preprocessing procedure consists of three distinct steps.

Firstly, image binarization and enhancement is applied to the documents. This involves the conversion of the gray scale images into binary ones and the improvement of their quality. The methodology is described in [3] [4] and in accordance with Niblack's approach [8] we get the desired results.

The second step of the preprocessing procedure involves the calculation of the average character height in the processed documents. Average character height is an important parameter since processes such as the segmentation process described in Section 3 uses it extensively. Average character height estimation uses a contour following process to determine the height of several random selected components and calculate the histogram of these heights. The maximum value of the histogram corresponds to the average character height.

The final step of the preprocessing procedure is the frame removal process. There are cases where frames surround the text areas of the processed collections resulting in wrongly segmented documents. The process is based on [2] and removes any surrounding frames enhancing the segmentation process described in Section 3. Figure 2 illustrates the results of the frame removal step.



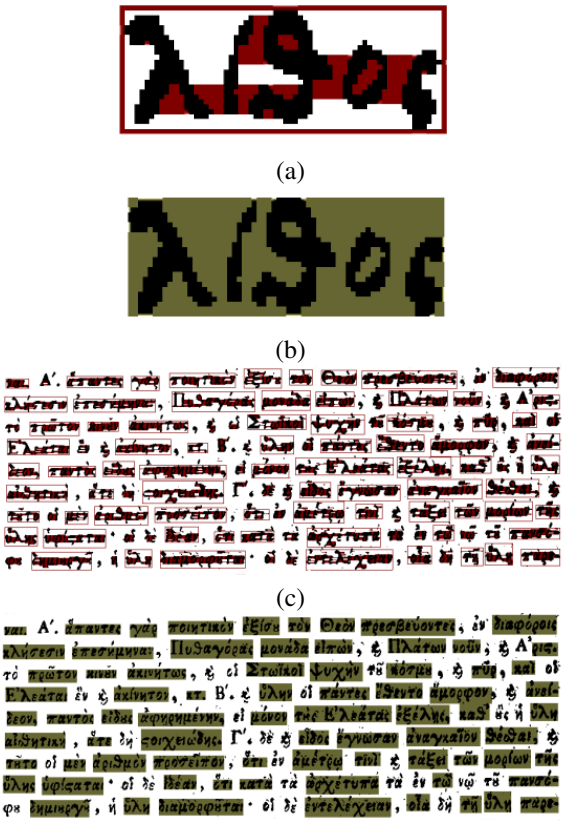
(a) (b)

Figure 2. (a) Original image; (b) Image after frame removal.

3. Segmentation

In this phase, the documents are segmented into words using the Run Length Smoothing Algorithm (RLSA) [12] [13]. The RLSA uses dynamic parameters which depend on the average character height as described in Section 2. The mechanism of the RLSA involves the examination of the white runs in both the horizontal and vertical dimensions. For each direction, the white runs that do not exceed a specified threshold are eliminated. For the documents used for our experiments, the horizontal length threshold has been experimentally defined as 50% of the average character heigth while the vertical length threshold has been experimentally defined as 10% of the average character height. The result of the RLSA is a binary image where characters

of the same word become connected to a single connected component as it can be seen in Figure 3a. The final segmented word is shoen in Figure 3b. Constraints such as the minimum expected word length are applied so as to enable stop-word rejection and eliminate undesired word segmentation. The minimum word length is defined as twice the average character height. Figure 3c illustrates the result of the RLSA in document level. Figure 3d illustrates the final segmented image having rejected words that did not satisy the minimum expected word length.



(a)

(b)

(c)

(d)

Figure 3. (a) Word Image after the RLSA (b) Final word segmentation (c) Sample of a document page after RLSA (d) The resulting word segmentation of the sample document page.

4. Synthetic Data

In order to create the synthetic keyword image, the user has to initially, manually select one example image template for each character. This task is performed once for each document collection and can be used for entire books or collections. The adjustment of the baseline of each character is also a required step during the selection of the example character image templates in order to minimize alignment problems when the synthetic keyword is created. The spacing between the characters has been experimentally defined as 10% of the average character height of the processed document collection. In Figure 4 the process of selecting a example character image is illustrated.

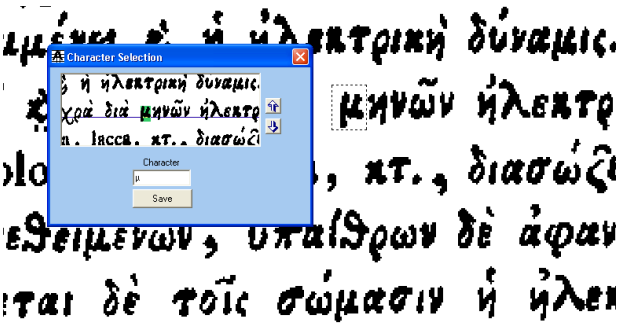


Figure 4. The process of manually selecting the individual character image templates. The adjustment of the baseline is also illustrated.

5. Feature Extraction

The feature extraction scheme used for the word matching process uses four different sets of features which are based on [9]. However, in the proposed methodology the images are binary.

The first feature set concern the number of black pixels in each image column (Figure 5b). A black pixel of an image I is denoted by $I(x, y) = 1$. For that image the number of black pixels in a column c is calculated as follows:

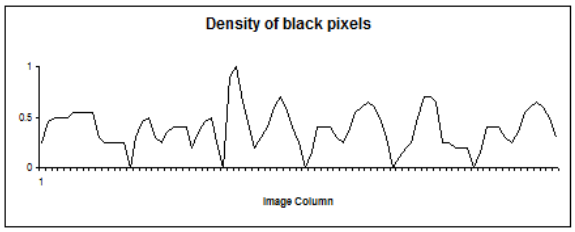
$$B(c) = \sum I(x, c) \tag{1}$$

The second feature set concerns the background-to-ink as well as the ink-to-background transitions on each image column (Figure 5c). Let wr_c be the white runs in an image column c of an image I and br_c be the black runs in the image column c of the image I. The background to ink transitions are calculated as follows:

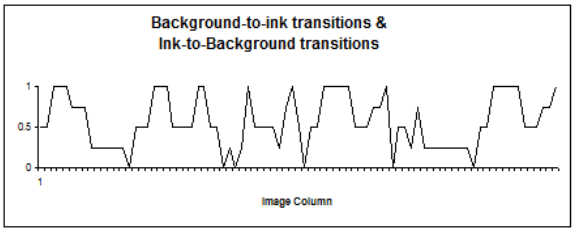
$$BGI(c) = \sum wr_c + \sum br_c \tag{2}$$

The third and fourth feature sets concern the distance to the first ink pixel from the top (Figure 5d) and bottom (Figure 5e) edge of the image respectively. Figure 5a is the word image that the extracted features correspond to. All features are normalised in the range $[0, 1]$.

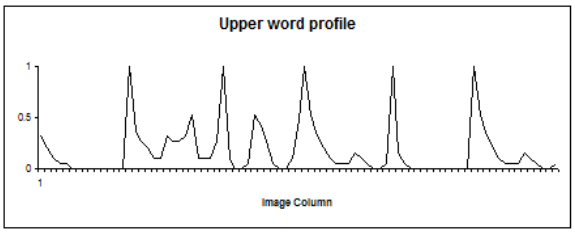
(a)



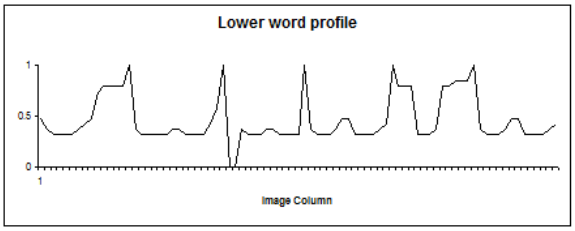
(b)



(c)



(d)



(e)

Figure 5. (a) The word image the features correspond to (b) The number of black pixels in each column of the word image (c) The background-to-ink and ink-to-background transitions of the word image (d) The upper profile of the word image (e) The lower profile of the word image.

6. Pruning

In this section we describe the proposed pruning method which leads to the rejection of words in the list of condidate words that do not satisfy certain conditions. These conditions are defined using the length of the words and the similarity of their first and last portions.

6.1 Word Length Pruning

The first condition that a segmented word must satisfy in order to be matched against the synthetic keyword is based on the word length. Let L_{synth} be the length of the synthetic word image, L_{word} be the length of a segmented word image from the dataset and c be the average letter height of the processed documents. Valid words are defined by the following equation:

$$|L_{synth} - L_{word}| \leq \frac{c \cdot L_{synth}}{100} \quad (3)$$

6.2 Pruning based on Word Portions

The aim of this phase is trying to match the first and last portions of each segmented word of the processed documents when compared with the synthetic keyword. We try to estimate the word portions of the segmented words that correspond to their first and last characters. Thus, a DTW matching is performed between the first and last character of the synthetic keyword and the first and last word portions of the segmented words.
Prior to this the example character image templates are compared in order to define the mean distance of each character image template in comparison to the others. The features used for that comparison is the background-to-ink and ink-to-background transitions and the number of black pixels per image column as described in Section 5.

For a character C_t we extract its features

$$F(C_t) = \{f_k(C_t, i)\}, \quad k = 1, 2 \quad (4)$$

where $f_k(C_t, i)$ is the k^{th} feature of the i^{th} column of character image template C_t.
The DTW between two charactes $C_t^1 = c_i \dots c_N$ and $C_t^2 = c_j \dots c_M$ is given by:

$$DTW(i,j) = min \left\{ \begin{array}{l} DTW(i-1,j) \\ DTW(i,j) \\ DTW(i,j-1) \end{array} \right\} + d(c_i, c_j) \quad (5)$$

where $d(c_i, c_j)$ is

$$d(c_i, c_j) = \sum_{k=1}^{2} (f_k(C_i, i) - f_k(C_j, j))^2 \quad (6)$$

this creates a warping path $W = (x_1, y_1) \ldots (x_K, y_K)$ between characters C_t^1 and C_t^2. The length K of the warping path W is used to determine the distance

$$DTW(C_t^1, C_t^2) = \sum_{k=1}^{K} d(c_{ik}, c_{jk}) \quad (7)$$

the overall distance $dist$ between the two character is given by

$$dist(C_t^1, C_t^2) = \frac{DTW(C_t^1, C_t^2)}{K} \quad (8)$$

Let $C = C_t^i, \ldots C_t^N$ be the set of the character image templates selected by the user. The mean distance $\overline{dist}$ of a character image template C_t^i in C towards the rest character image templates in C is given by

$$\overline{dist}(C_t^i) = \frac{1}{N} \sum_{r \neq i} dist(C_t^i, C_t^r) \quad (9)$$

where N denotes the number of character image templates in C.

The mean distances defined in equation 9 are the thresholds of the first and last character that that will determine whether a segmented word is valid for the matching process. We proceed with the matching between the first and last character of the synthetic keyword with the first and last portions of the segmented word.

Assume that the length of the first character of the synthetic keyword is x_{first} and the length of the last character is x_{last}. Let x_w be the length of the segmented word image. Since we are processing machine-printed documents a rather high similarity between the same characters is expected. However, same characters do not have equal lengths therefore, the matching process allows a variability $\pm s$ which is equal to 15% of the average character height as shown in Figure 6.

We perform a DTW matching between the first and last character features and the first and last portions of the features of the segmented word image. The range that we perform our matching on the segmented word is $[0, x_{first+s}]$ for the first character and $[x_{last-s}, x_w]$ for the last character. There are in total five DTW matches for each word portion. The minimum total distance is compared with the mean distance of the character as evaluated in Equation 9.

Assuming that the first character of the synthetic keyword is C_f and the last C_l and the first word portion of the segmented word image is FW_i and the last LW_i then for the first character we have

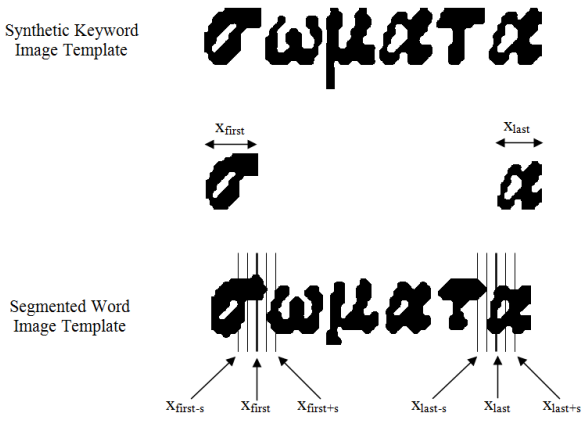


Figure 6. Matching the word portions. The process tries to estimate the position of the first and last character of the segmented word in order to prune the dataset correctly.

$$dist(C_f, FW) = \min(\frac{DTW(C_f, FW_i)}{K}) \quad (10)$$

where $i \in [x_{first-s}, x_{first+s}]$

and for the last character we have

$$dist(C_l, LW) = \min(\frac{DTW(C_l, LW_i)}{K}) \quad (11)$$

where $i \in [x_{last-s}, x_{last+s}]$.

The valid words must satisfy the following:

$$dist(C_f, FW) \leq \overline{dist}(C_f, FW) \quad (12)$$

$$dist(C_l, LW) \leq \overline{dist}(C_l, LW) \quad (13)$$

7. Word Matching

The word matching process involves the matching of the synthetic keyword with the segmented word images remained after the pruning process described in Section 6. The features used in the word matching process are all the features described in Section 5. Let S be the synthetic word image and W be the segmented word image. The features extracted for these two images are:

$$F(S) = \{f_k(S, i)\}, \quad k = 1, 2, 3, 4 \quad (14)$$

where $f_k(S,i)$ is the k^{th} feature of the i^{th} column of a synthetic keyword image S

$$F(W) = \{f_k(W,i)\}, \quad k = 1,2,3,4 \tag{15}$$

where $f_k(W,i)$ is the k^{th} feature of the i^{th} column of a segmented word image W.
We perform a DTW on these two images and we calculate their total distance

$$dist(S,W) = \frac{DTW(S,W)}{K} \tag{16}$$

where K is the length of the warping path formed after the application of the TDW algorithm.

8. Experimental Results

Experiments involved searching 25 keywords over a sample of 100 document pages. The keywords were selected among the most frequently appearing keywords in the sample document pages. The keywords were manually marked in the sample document pages in order to create a ground truth. The total number of the segmented words throughout the 100 sample document pages was 27,702.
Evaluation is performed using presicion versus recall curves. Precision is the ratio of the number of relevant words to the number of retrieved words. Recall is the ratio of the number of retrieved relevant words to the number of total relevant words marked on the sample document pages. Precision and recall are defined as follows:

$$Precision(A) = \frac{R_a}{A} \tag{17}$$

$$Recall(A) = \frac{R_a}{S} \tag{18}$$

where A denotes the number of word images retrieved, S denotes the total number of relevant marked words, and R_a denotes the retrieved relevant words from A. We have used a variety of answer sets by a step of 5% of the total word instances in the dataset of the corresponding class.
We have conducted two types of experiments. The first one applies the DTW algorithm to the list of candidate words without any pruning, that is, each keyword is compared to every other word in the list of condidate words. The second experiment uses the proposed pruning method. The proposed method is also compared against the method described in [5] where a keyword spotting in historical printed documents using a hybrid model is presented. As it can be seen in Figure 7 the proposed pruning method improved results against the aforementioned methods. Moreover, the pruning method reduces the list of candidate words by a mean of 80% concerning all keywords. This has an immediate effect on the time required to conduct the experiments since the pruning version is about 50% faster than the non-pruning one. Table 1 shows the precision, recall and f-measure values for each of the methods used for the experiments.

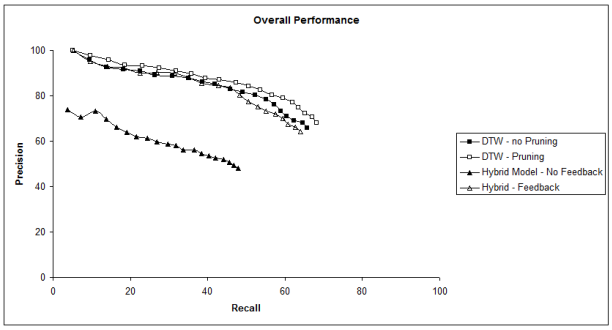


Figure 7. Average procesion/recall rates concerning all keywords.

Method	Precision	Recall	F-Measure
Pruning	84.07	50.53	41.89
nonPruning	81.50	48.98	37.69
Hybrid	57.62	34.65	26.64

Table 1. The precision/recall values concerning 60% of the total word instances

9. Conclusions

We proposed a novel technique for keyword guided image matching using DTW. The method can be used for indexing large document collections due to its ability to significantly prune the dataset. Word length and word portions based on their first and last characters are the thresholds used to reduce the data. Furthermore, the pruning process speeds up the overall matching procedure and improves significantly the precision versus recall results.
Future work involves the conduction of experiments using different feature sets in order to further improve the matching accuracy on the first and last characters thus, producing better precision versus recall results concerning 60% of the total number of word instances.

References

[1] M. Faúndez-Zanuy. On-line signature recognition based on vq-dtw. *Pattern Recognition*, 40(3):981–992, 2007.

[2] B. Gatos, D. Danatsas, I. Pratikakis, and S. J. Perantonis. Automatic table detection in document images. In *ICAPR (1)*, pages 609–618, 2005.

[3] B. Gatos, I. Pratikakis, and S. J. Perantonis. An adaptive binarization technique for low quality historical documents. In *Document Analysis Systems*, pages 102–113, 2004.

[4] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[5] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis. Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. *IJDAR*, 9(2-4):167–177, 2007.

[6] V. Lavrenko, T. M. Rath, and R. Manmatha. Holistic word recognition for handwritten historical documents. In *DIAL*, pages 278–287, 2004.

[7] R. Manmatha and J. L. Rothfeder. A scale space approach for automatically segmenting words from historical handwritten documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1212–1225, 2005.

[8] W. Niblack. *An Introduction to Image Processing*. Prentice Hall, Englewood Cliffs, 1996.

[9] T. M. Rath and R. Manmatha. Features for word spotting in historical manuscripts. In *ICDAR*, pages 218–222, 2003.

[10] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *CVPR (2)*, pages 521–527, 2003.

[11] A. P. Shanker and A. N. Rajagopalan. Off-line signature verification using dtw. *Pattern Recognition Letters*, 28(12):1407–1414, 2007.

[12] S. Theodoridis and K. Koutroubas. *Pattern Recognition*. Academic Press, New York, 1997.

[13] F. M. Wahl, K. Y. Wong, and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. *Comput. Graph. Image Process*, 20:375–390, 1982.