IJDAR (2007) 9:167–177
DOI 10.1007/s10032-007-0042-4

ORIGINAL PAPER

Keyword-guided word spotting in historical printed documents using synthetic data and user feedback

T. Konidaris · B. Gatos · K. Ntzios · I. Pratikakis · S. Theodoridis · S. J. Perantonis

Received: 21 February 2005 / Revised: 20 December 2006 / Accepted: 30 January 2007 / Published online: 10 March 2007

Abstract In this paper, we propose a novel technique for word spotting in historical printed documents combining synthetic data and user feedback. Our aim is to search for keywords typed by the user in a large collection of digitized printed historical documents. The proposed method consists of the following stages: (1) creation of synthetic image words; (2) word segmentation using dynamic parameters; (3) efficient feature extraction for each word image and (4) a retrieval procedure that is optimized by user feedback. Experimental results prove the efficiency of the proposed approach.

T. Konidaris (✉) · B. Gatos · K. Ntzios · I. Pratikakis · S. J. Perantonis
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos",
Athens, Greece
e-mail: tkonid@iit.demokritos.gr

B. Gatos
e-mail: bgat@iit.demokritos.gr

K. Ntzios
e-mail: ntzios@iit.demokritos.gr

I. Pratikakis
e-mail: ipratika@iit.demokritos.gr

S. J. Perantonis
e-mail: sper@iit.demokritos.gr

K. Ntzios · S. Theodoridis
Department of Informatics and Telecommunications,
National & Kapodistrian University of Athens,
Athens, Greece
e-mail: ntzios@di.uoa.gr

S. Theodoridis
e-mail: stheodor@di.uoa.gr

Keywords Historical document indexing · Word spotting · User feedback

1 Introduction

Historical printed documents contain a vast amount of valuable information. A robust indexing of these documents is essential for quick and efficient content exploitation of the valuable historical collections. In this paper, we deal with historical printed Greek documents that date since the period of Renaissance and Enlightenment (1471–1821) and are considered among the first Greek printed historical documents. Nevertheless, the proposed methodology is generic having the potential to be applied to other than Greek historical printed documents. The general framework of our work is the development of a system that will integrate, manage and provide access to historical printed documents.

Traditional approaches in document indexing usually involve an Optical Character Recognition (OCR) step [5]. OCR is widely used in a variety of applications [14] and it performs well in modern printed documents and documents of high quality printing. In the case of printed historical documents OCR, several factors affect the final performance like low paper quality, paper positioning variations (skew, translations, etc), low print contrast, typesetting imperfections. Usually, printed OCR systems involve a character segmentation step followed by a recognition step using pattern classification algorithms. Due to document degradations, OCR systems often fail to support a correct segmentation of the printed historical documents into individual characters [2]. This is the case of the documents we used for our experiments. The low quality of the documents as well as typesetting imperfections did not allow any OCR

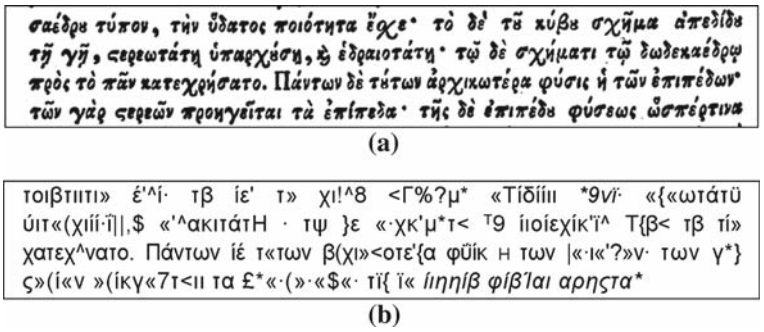

Fig. 1 **a** An extract from a historical document image used for our experiments, **b** OCR results using FineReader® [1] OCR software fails to correctly recognize the text

tasks. Figure 1 shows the results of conventional OCR of an extract taken from the processed printed historical documents used in our experiments.

In the literature, two general approaches can be identified: the segmentation approach and the global or segmentation-free approach. The segmentation approach requires that each word has to be segmented into characters while the global approach entails the recognition of the whole word. In the segmentation approach, the crucial step is to split a scanned bitmap image of a document into individual characters [9].

A segmentation-free approach is followed in [4,6,15, 16,18,23,35] where line and word segmentation is used for creating an index based on word matching. In [23], a discussion on different approaches to word matching is given. In [4], Ulam's distance is used for image matching by identifying the smallest number of mutations between two strings. In [6], a two-dimensional image is converted into a one-dimensional string. The method describes how to extract information from the strings and compute the distance between them resulting in similar matches. In the segmentation-free approach of [35], word matching is based on the vertical bar patterns. Each word is represented as a series of vertical bars that is used for the matching process. Word image matching is also applied in [18] using the weighted Hausdorff distance. Before applying the matching process using the Hausdorff distance a normalization scheme is used for each word. Word matching is also performed in [16] where global and local features based on profile signatures and morphological cavities are used for each word characterization. Matching of whole words in printed documents is performed in [3]. In this approach, a Dynamic Time Warping (DTW) based partial matching scheme is used to overcome the morphological differences between the words. Another segmentation-free approach which uses HMMs and statistical language models for handwritten text is described in [31]. In [8] a holistic approach is used in order to digitize natural history cards containing both printed and handwritten information.

In the case of historical documents, Rath and Manmatha [26] presented a word matching scheme where noisy handwritten document images are preprocessed into one-dimensional feature sets and compared using the DTW algorithm. Rath et al. [25] present a method for retrieving large collections of handwritten historical documents using statistical models. Lavrenko et al. [17] present a holistic word recognition approach for handwritten historical documents. They performed a series of experiments showing the performance of this segmentation-free approach applied to degraded historical documents where the segmentation of words into characters is not feasible.

In [20,22] a method for word spotting is presented wherein matching was based on the comparison of entire words rather than individual characters. In this method, an off-line grouping of words in a historical document and the manual characterization of each group by the ASCII equivalence of the corresponding words are required. The volume of the processed material was limited to a few pages. This process can become very tedious for large collections of documents.

Typing all unique words as well as constructing an index is an almost impossible task for large document collections. To eliminate this tedious process, we propose a novel method for keyword-guided word spotting which is based on: (1) creation of synthetic image words; (2) word segmentation using dynamic parameters; (3) efficient feature extraction for each image word and (4) a retrieval procedure that is improved by user feedback. The synthetic keyword image is used as the query image for the retrieval of all relevant words, initializing in this way, the word spotting procedure. The retrieval accuracy is further improved by the user feedback. Combination of synthetic data creation and user feedback leads to satisfactory results in terms of precision and recall. Figure 2 illustrates the distinct steps of the proposed system architecture.

The paper is organized as follows: In Sect. 2, we describe the preprocessing stage including image binarization and enhancement, calculation of the average letter height as well as frame removal. Section 3 is concerned with the word segmentation process that is based upon an efficient smoothing procedure.

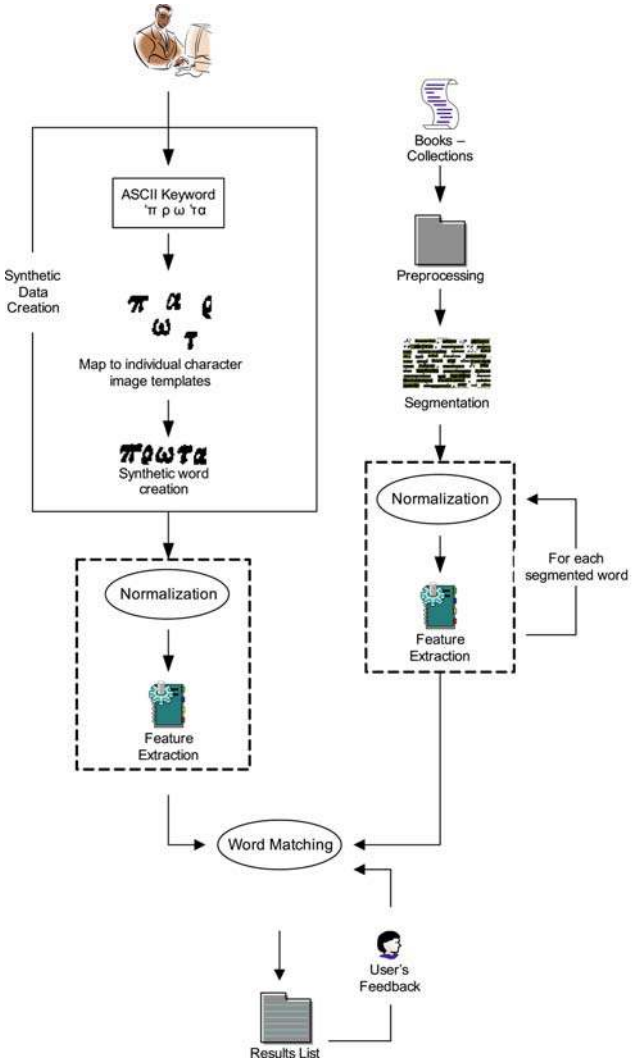

Fig. 2 The overall components of the proposed system architecture

Section 4 describes the process of creating synthetic keyword images using character image templates. Section 5 details the feature extraction process while Sect. 6 is dedicated to the analysis of the word image retrieval process that is enhanced by the incorporation of user feedback. Experimental results are given in Sect. 7, demonstrating the performance of the proposed method in terms of precision and recall. Finally, in Sect. 8 conclusions are drawn.

2 Preprocessing

2.1 Image binarization and enhancement

Binarization is the starting step of most document image analysis systems and refers to the conversion of the gray-scale image to a binary image. Since historical

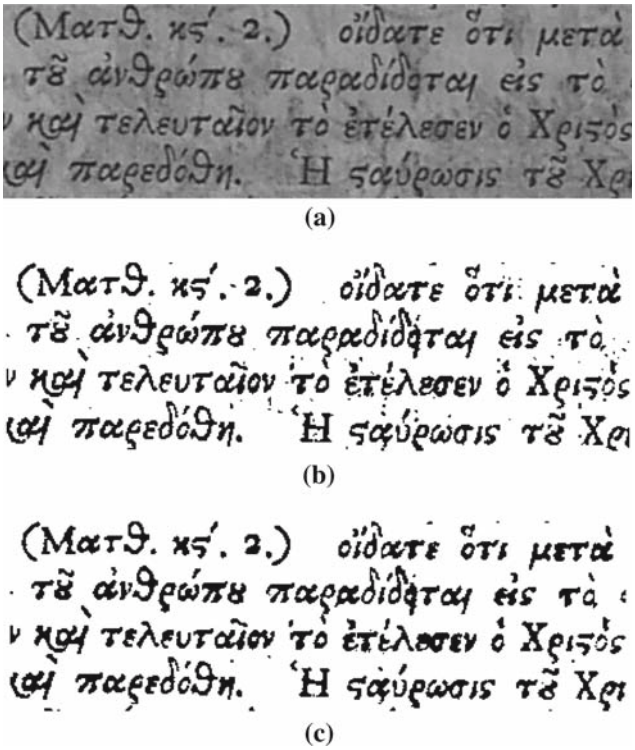

Fig. 3 Image binarization and enhancement example. **a** Original gray scale image; **b** Resulting image after binarization; **c** Resulting image after image enhancement

document collections are usually of very low quality, an image enhancement stage is also essential. The proposed scheme for image binarization and enhancement is described in [12,13] and consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach [24], a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. Figure 3 illustrates the steps for improving image quality and finally a post-processing step that preserves stroke connectivity as well as eliminates noise and improves the quality of text regions by isolated pixel removal and filling of possible breaks, gaps or holes. Figure 3 illustrates the steps for image binarization and enhancement.

2.2 Average character height estimation

The average character height estimation is required for the frame removal step described in Sect. 2.3 as well as for the segmentation phase described in Sect. 3. For the calculation of the average character height we take a random pixel (x_A, y_A) that has at least one background

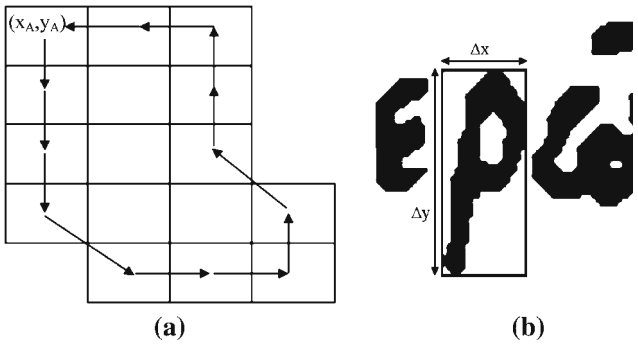

Fig. 4 **a** Contour following of the connected component; **b** Specifying the bounding box of the connected component

pixel in its four-connected neighborhood and we follow the contour of the connected component where pixel (x_A, y_A) belongs to. This procedure is repeated for a predefined number of random pixels that have at least one background pixel in their four-connected neighborhood. We then calculate the histogram with the heights of every connected component we have processed. The maximum value of the histogram corresponds to the average character height. Figure 4a illustrates the process of following the contour of the connected component that pixel (x_A, y_A) belongs to. Figure 4b illustrates how the bounding box of the connected component is formed. Δy corresponds to the height of the connected component.

2.3 Frame removal

To ease the segmentation process we remove potential frames around the text areas. The process of frame removal is based on the work of [10]. The line detection algorithm is based on processing horizontal and vertical black runs as well as on a set of morphological operations with suitable structuring elements in order to connect possible line breaks and to enhance line segments. All parameters used in this step depend on the average character height that has been calculated in Sect. 2.2. An example of the frame removal procedure is shown at Fig. 5.

3 Segmentation

The process involves the segmentation of the document images into words. Manmatha and Rothfeder [21] use a technique for automatically segmenting documents into words. In [29] different techniques for separating lines of unconstrained handwritten text into words is described. In the proposed methodology this is accomplished with the use of the Run Length Smoothing Algorithm (RLSA) [30,35] by using dynamic parameters which depend on the average character height as it is described in Sect. 2.2. RLSA examines the white runs existing in the horizontal and vertical directions. For

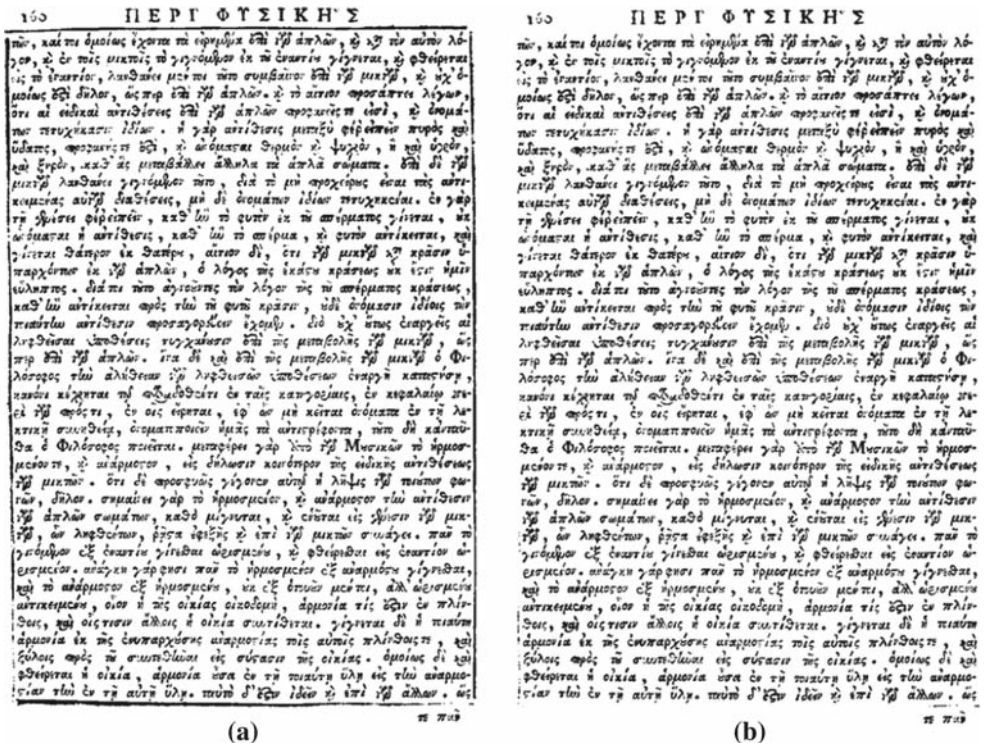

Fig. 5 **a** The original image surrounded by a frame; **b** The resulting image after applying frame removal procedure

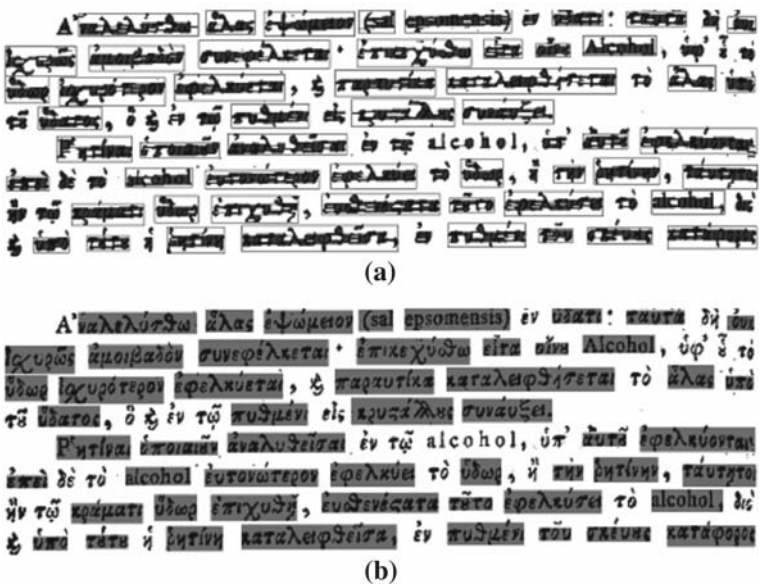

Fig. 6 Segmentation process: **a** Resulting image after RLSA; **b** Final word segmentation. The *words* that are not highlighted have a high probability to be stop-words due to their small size and are therefore rejected as keywords

each direction, white runs with length less than a threshold are eliminated. In the proposed method, the horizontal length threshold is experimentally defined as 50% of the average character height while the vertical length threshold is experimentally defined as 10% of the average character height. The application of RLSA results in a binary image where characters of the same word become connected to a single connected component (Fig. 6a). Although there should exist cases where the character spacing may exceed the RLSA threshold leading to incorrect segmentation (see, word "alcohol" at Fig. 6), this does not cause significant problems since words with wide character spacing are found rarely in the processed collections. In the sequel, a connected component analysis is applied using constraints which express the minimum expected word length (Fig. 6b). This will enable us to reject stop-words and therefore eliminating undesired word segmentation. More specifically, the minimum expected word length for words that are not stop-words has been experimentally defined to be twice the average character height.

4 Synthetic data creation

Synthetic data creation concerns the synthesis of the keyword images from their ASCII equivalent. Prior to the synthesis of the keyword image, the user selects one example image template for each character. This selection is performed "once-for-all" and can be used for entire books or collections. During manual character marking, adjustment of the baseline for each character image template is applied in order to minimize alignment problems. The baseline is manually adjusted by the user. Figure 7 demonstrates an example of the resulting synthetic data with and without adjustment of the baseline as well as the individual characters used for the creation of the synthetic keyword image. The spacing between the characters has been experimentally defined as 10% of the average character height that has been estimated over the complete document collection. The average character height may vary depending on the document collection under study.

5 Feature extraction

The feature extraction phase consists of two distinct steps; (1) normalization and (2) hybrid feature extraction.

5.1 Normalization

The normalization process is dedicated to preserve scale invariance for word images. In particular, for the normalization of the segmented words we use a bounding box with user-defined dimensions. The segmented words are resized to fit in the bounding box while preserving their aspect ratio. The size of the bounding box concerning both width and height is the same for all words. Thereafter, exact positioning of the word in the bounding box is achieved by placing the geometric center of the word in the center of the bounding box.

5.2 Hybrid feature scheme

For the word matching, feature extraction from the word images is required. Several features and methods have

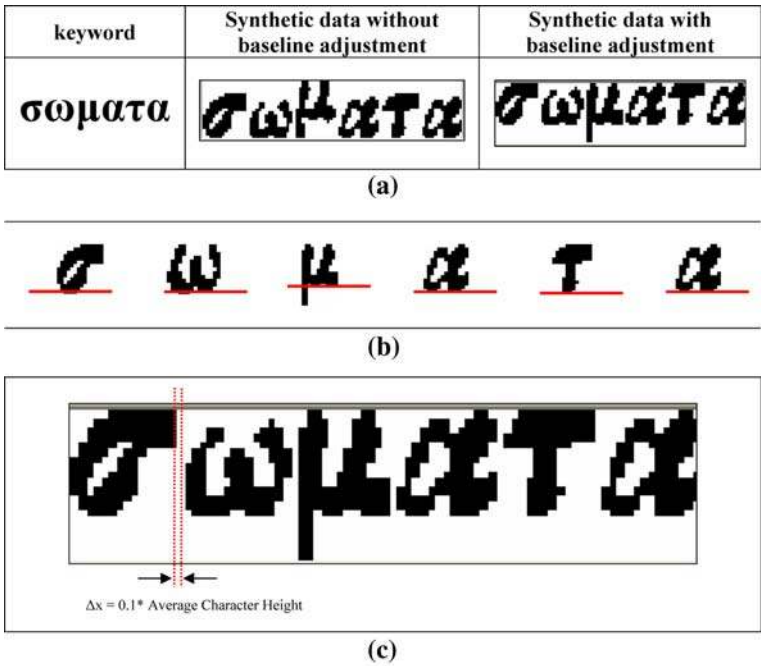

Fig. 7 An example of synthetic data. **a** The creation of the synthetic image from its ASCII representation; **b** baseline adjustment for character image templates; **c** Integration of single character image templates into a synthetic keyword. Spacing between the characters is defined to be Δx = 0.1 * Average Character Height

been proposed based on strokes, contour analysis, zones, projections, etc. [4–7,27]. In our approach, we employ two types of features in a hybrid fashion. The first one, which is based on [5], divides the word image into a set of zones and calculates the density of the character pixels in each zone. The second type of features is based on the work in [27], where we calculate the area that is formed from the projections of the upper and lower profile of the word.

In the case of features based on zones, the image is divided into horizontal and vertical zones. In each zone, we calculate the density of the character pixels (see Fig. 8). Let $\mathrm{im}(x,y)$ be the word image array having 1s for foreground and 0s for background pixels, $x_{\max}$ and $y_{\max}$ be the width and the height of the word image, respectively and Z_{H} and Z_{V} be the total number of zones formed in both horizontal and vertical direction. Then, features based on zones $f^z(i), i = 0,\ldots,Z_{\mathrm{H}}Z_{\mathrm{V}}-1$ are calculated as follows:

$$f^z(i) = \sum_{x=x_{\mathrm{s}}(i)}^{x_{\mathrm{e}}(i)} \sum_{y=y_{\mathrm{s}}(i)}^{y_{\mathrm{e}}(i)} \mathrm{im}(x,y) \tag{1}$$

where,

$$x_{\mathrm{s}}(i) = \left(i - \left\lfloor \frac{i}{Z_{\mathrm{H}}} \right\rfloor Z_{\mathrm{H}}\right) \frac{x_{\max}}{Z_{\mathrm{H}}}$$

$$x_{\mathrm{e}}(i) = \left(i - \left\lfloor \frac{i}{Z_{\mathrm{H}}} \right\rfloor Z_{\mathrm{H}} + 1\right) \frac{x_{\max}}{Z_{\mathrm{H}}}$$

$$y_{\mathrm{s}}(i) = \left\lfloor \frac{i}{Z_{\mathrm{H}}} \right\rfloor \frac{y_{\max}}{Z_{\mathrm{V}}}$$

$$y_{\mathrm{e}}(i) = \left(\left\lfloor \frac{i}{Z_{\mathrm{H}}} \right\rfloor + 1\right) \frac{y_{\max}}{Z_{\mathrm{V}}}$$

In the case of features based on word (upper/lower) profile projections, the word image is divided into two sections separated by the horizontal line $y = y_t$ which passes through the center of mass of the word image (x_t, y_t) (see 2).

$$y_t = \frac{\sum_x \sum_y \mathrm{im}(x,y) \cdot y}{\sum_x \sum_y \mathrm{im}(x,y)} \tag{2}$$

Upper/lower word profiles (3, 4) are computed by considering, for each image column, the distance between the horizontal line $y = y_{\mathrm{t}}$ and the closest character pixel to the upper/lower boundary of the word image (see Fig. 9):

$$y_{\mathrm{up}}(x) = y_t - y_0,$$

$$\text{where } y_0 = \begin{cases} y_t, \text{ if } \sum_{\mathrm{y}=0}^{y_{\mathrm{t}}} \mathrm{im}(x,y) = 0, \\ y: \ (\mathrm{im}(x,y) = 1 \ \& \ y = \min(y_i)), \\ \quad y_{\mathrm{i}} \in [0, y_t], \qquad \text{otherwise} \end{cases} \tag{3}$$

$$y_{lo}(x) = y_0 - y_t,$$

$$\text{where } y_0 = \begin{cases} y_t, \text{ if } \sum_{\mathrm{y}=y_t}^{y_{\max}} \mathrm{im}(x,y) = 0, \\ y: \ (\mathrm{im}(x,y) = 1 \ \& \ y = \max(y_i)), \\ \quad y_i \in [y_t, y_{\max}], \qquad \text{otherwise} \end{cases} \tag{4}$$

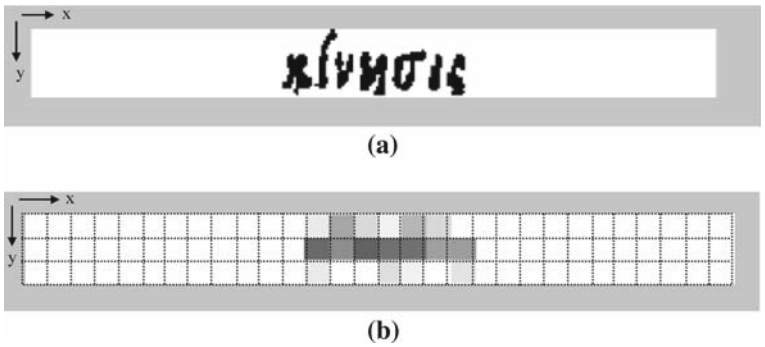

Fig. 8 Feature extraction of a word image based on zones. **a** the normalized word image; **b** features based on zones. *Darker squares* indicate higher density of character pixels

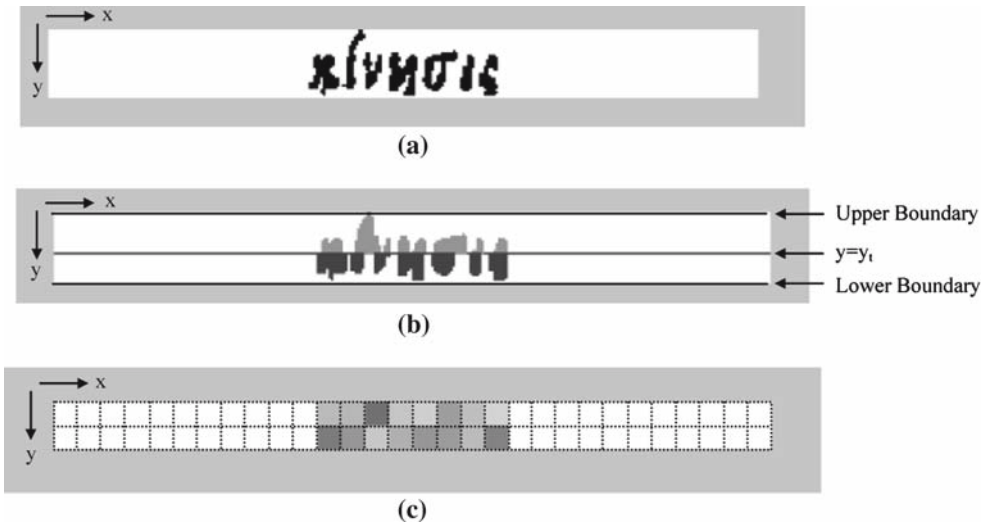

Fig. 9 Feature extraction of a word image based on word profile projections. **a** The normalized word image; **b** Upper and lower word profiles; **c** The extracted features. *Darker squares* indicate higher density of zone pixels

We define P_{V} as the total number of blocks that each produced zone (upper, lower) is divided. For each block, we calculate the area of the upper/lower word profiles denoted as in the following:

$$f^{p}_{\text{upper\_area}}(i) = \sum_{x=x_{\mathrm{s}}(i)}^{x_{\mathrm{e}}(i)} y_{\mathrm{up}}(x) \tag{5}$$

$$f^{p}_{\text{lower\_area}}(i) = \sum_{x=x_{\mathrm{s}}(i)}^{x_{\mathrm{e}}(i)} y_{\mathrm{lo}}(x) \tag{6}$$

where,

$$x_{\mathrm{s}}(i) = \left(i - \left\lfloor \frac{i}{P_{\mathrm{V}}} \right\rfloor P_{\mathrm{V}}\right) \frac{x_{\max}}{P_{\mathrm{V}}}$$

$$x_{\mathrm{e}}(i) = \left(i - \left\lfloor \frac{i}{P_{\mathrm{V}}} \right\rfloor P_{\mathrm{V}} + 1\right) \frac{x_{\max}}{P_{\mathrm{V}}}$$

where $i = 0, \ldots, P_V - 1$. Figure 9 illustrates the features extracted from a word image using projections of word profiles.

The overall calculation of the proposed hybrid feature vector is given in (7). The corresponding feature vector length equals to $Z_{\mathrm{H}}Z_{\mathrm{V}} + 2P_{\mathrm{V}}$.

$$f(i) = \begin{cases} f^{z}(i) = \sum_{x=x_{\mathrm{s}}(i)}^{x_{\mathrm{e}}(i)} \sum_{y=y_{\mathrm{s}}(i)}^{y_{\mathrm{e}}(i)} \mathrm{im}(x,y), i = 0 \ldots Z_{\mathrm{H}}Z_{\mathrm{V}} - 1 \\ f^{p}_{\text{upper\_area}}(i) = \sum_{x=x_{\mathrm{s}}(i-Z_{\mathrm{H}}Z_{\mathrm{V}})}^{x_{\mathrm{e}}(i-Z_{\mathrm{H}}Z_{\mathrm{V}})} y_{\mathrm{up}}(x), \\ \quad i = Z_{\mathrm{H}}Z_{\mathrm{V}} .. Z_{\mathrm{H}}Z_{\mathrm{V}} + P_{\mathrm{V}} - 1 \\ f^{p}_{\text{lower\_area}}(i) = \sum_{x=x_{\mathrm{s}}(i-Z_{\mathrm{H}}Z_{V}+P_{V})}^{x_{\mathrm{e}}(i-Z_{\mathrm{H}}Z_{\mathrm{V}}+P_{\mathrm{V}})} y_{\mathrm{lo}}(x), \\ \quad i = Z_{\mathrm{H}}Z_{\mathrm{V}} + P_{\mathrm{V}} .. Z_{\mathrm{H}}Z_{\mathrm{V}} + 2P_{\mathrm{V}} - 1 \end{cases} \tag{7}$$

6 Word image retrieval

6.1 Word matching

The process of word matching involves the comparison/matching between the query word (a synthetic keyword

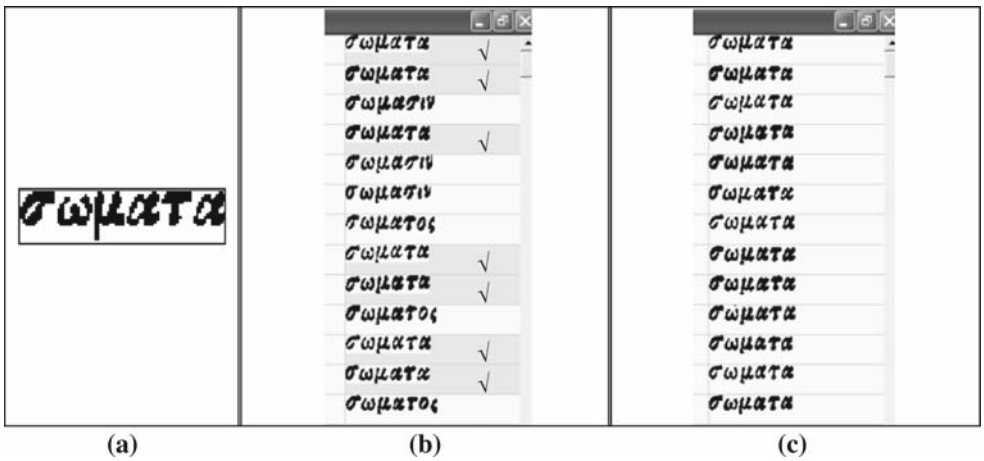

Fig. 10 Searching for word "$\sigma\omega\mu\alpha\tau\alpha$" using user feedback: **a** query synthetic keyword image; **b** initial results list using synthetic data. Marked results indicate the user selection of the correct words; **c** the final results of the word matching process using as query the selected word images from the list shown at **b**

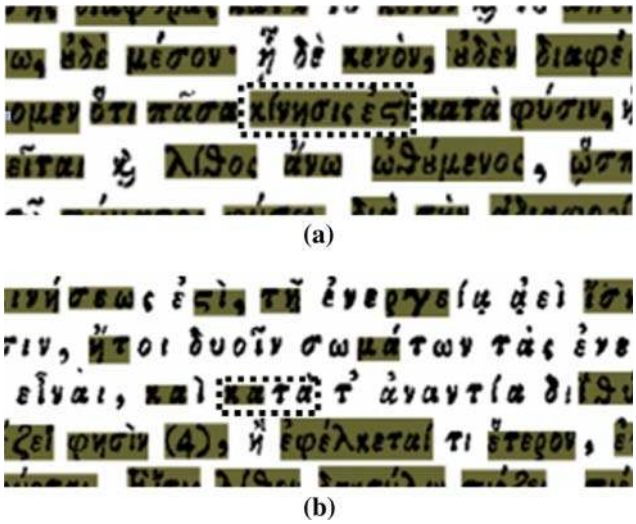

Fig. 11 Examples of wrongly segmented words: **a** word image merging; **b** single word image splitting

image) and all the indexed segmented words. Ranking of the comparison results is based on L_1 distance metric as in the following:

$$Dist(f_{\mathrm{q}}(i), f_{db}(i)) = \sum_i \left\| f_{\mathrm{q}}(i) - f_{db}(i) \right\| \tag{8}$$

where, $f_{\mathrm{q}}(i)$ concerns the features of the query word (synthetic image) and $f_{db}(i)$ concerns the features of the segmented images.

The aim is to produce an initial list of ranked results that will be improved with the user feedback.

6.2 User feedback

User feedback is an efficient mechanism for drastically improving the results of the matching process. Since the initial results are based on the comparison of the synthetic keyword with all the detected words, these results might not present high accuracy because a synthetic keyword cannot a priori perform a perfect match with a real word image. Motivated by this, we propose a user intervention where the user selects as query the correct results from the list produced after the initial word matching process as described in Sect. 6.1. Then, a new matching process is initiated. The critical impact of the user feedback in the word spotting process lies upon the transition from synthetic to real data. Furthermore, in our approach user interaction is supported by a simplified and user friendly graphical interface that makes the word selection procedure an easy task. Figure 10 illustrates the matching results before and after the user feedback process.

7 Experimental results

For the evaluation of the performance of the proposed method for keyword-guided word spotting in historical printed documents, we used the following methodology. We created a ground truth set by manually marking certain keywords on a subset of the available document collection. The performance evaluation method used is based on counting the number of matches between the words detected by the algorithm and the marked words in the ground truth. For the experiments we used a

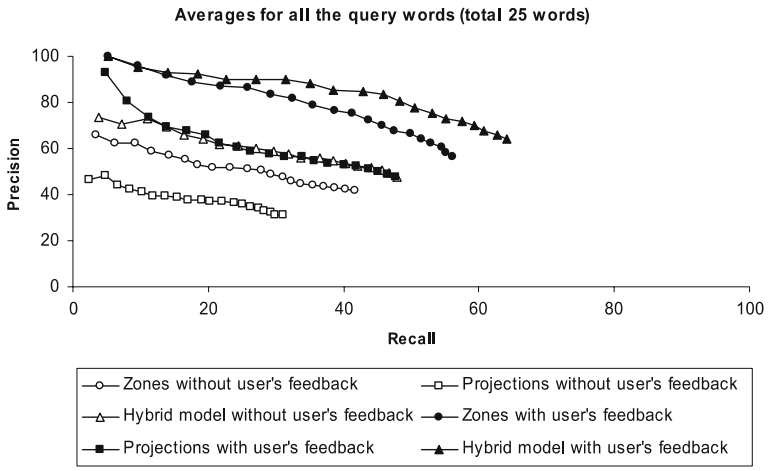

Fig. 12 Average Precision/Recall rates for all the query words

sample of 100 document pages. We have used in total 25 keywords that were randomly selected through a list of the most frequently appearing words in the sample document images. The total number of words detected in the sample images is 27,702. The task is to search for the 25 randomly selected keywords. When applying user feedback we use the first 5% of the correct words in the results list.

Due to segmentation errors (see Fig. 11), there are a number of word misses that affect the precision/recall rates. There is an average of 1.6% concerning segmentation errors.

Evaluation is performed using precision 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 images. They are defined as follows [28]:

$$\text{Precision}(A) = \frac{R_\text{a}}{A} \tag{9}$$

$$\text{Recall}(A) = \frac{R_\text{a}}{S} \tag{10}$$

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 10% of the total word instances in the dataset of the corresponding class.

The size of the normalized word images used is $x_\text{max} = 300$ and $y_\text{max} = 30$.

In the case of features based on zones, the word image is divided into three ($Z_\text{H} = 3$) horizontal and thirty ($Z_\text{V} = -30$) vertical zones forming a total of ninety (90) blocks with size 10×10 (see Fig. 8). Therefore, the total number of features is ninety (90).

In the case of features based on word (upper/lower) profile projections we keep the same size of the normalized image, while the image is divided into thirty (30) vertical zones ($P_\text{V} = 30$) (see Fig. 9). Consequently, the total number of features equals to sixty (60).

Combination of features based on zones and features based on word profile projections led to the proposed hybrid model (7) that uses a total of ninety (90) features.

The overall system performance given in Fig. 12 shows the average recall vs. average precision curves in the case of single features as well as in the case of the proposed hybrid scheme. Furthermore, it is demonstrated the performance achieved in the absence or presence of user feedback. It is clearly illustrated that the hybrid scheme outperforms the single feature approaches. Additionally, in all cases when the user feedback is applied, the precision/recall rates are improved by at least of 20%. Combination of synthetic data creation and user feedback leads to improved performance in terms of precision and recall.

8 Conclusions

This paper proposes a novel technique for keyword-guided word spotting in historical printed documents. It is based upon: (1) creation of synthetic image words; (2) word segmentation using dynamic parameters; (3) efficient hybrid feature extraction and (4) a retrieval procedure that is optimized by user feedback.

In this work, we propose a methodology which introduces a novel way to initialize the word retrieval mechanism through the creation of synthetic word data along with a robust hybrid feature extraction that supports meaningful representations of word images.

From our experimental results, it can be clearly stated that the combination of synthetic data and user feedback

in a hybrid fashion leads to an improved performance for keyword-guided word spotting in comparison with the other feature extraction schemes used.

Our future research will focus on exploiting new features as well as fusion methods to further improve the performance for keyword-guided word spotting in large historical collections. This will be applied for both printed typed and handwritten manuscripts.

Acknowledgments This research is carried out within the framework of the Greek GSRT-funded R&D project, KATOPTRON, which aims to develop an integrated system for integration, management and access to Greek historical material concerning Philosophy, Science and Education during the period between 1453–1821.

References

1. ABBYY FineReader®. http://www.abbyy.com/finereader_ocr/. (2005)
2. Baird, H.S.: The state of the art of document image degradation modeling. In: IAPR 2000 Workshop on Document Analysis Systems, December 2000, pp. 10–13 (2000)
3. Balasubramanian, A., Meshesha, M., Jawahar, C.V.: Retrieval form document image collections. In: DAS 2006, pp. 1–12 (2006)
4. Bhat, D.: An evolutionary measure for image matching. In: Proceedings of the Fourteenth International Conference on Pattern Recognition, ICPR'98, vol. I, pp. 850–852 (1998)
5. Bokser, M.: Omnidocument technologies. Proc. IEEE, **80**(7), 1066–1078 (1992)
6. Cha, S.-H., Shin, Y.-C., Srihari, S.N.: Approximate stroke sequence string matching algorithm for character recognition and analysis. In Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), pp. 53–56 (1999)
7. Doerman, D., Li, H., Kia, O.: The detection of duplicates in document image databases. In: Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), pp. 314–318 (1997)
8. Downton, A.C., Lucas, S.M., Patoulas, G., Beccaloni, G.W., Scoble, M.J., Robinson, G.S.: Computerising natural history cards. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), pp. 354–358 (2003)
9. Gatos, B., Papamarkos, N., Chamzas, C.: A binary tree based OCR technique for machine printed characters. Eng. Appl. Artif. Intell. **10**(4), 403–412 (1997)
10. Gatos, B., Danatsas, D., Pratikakis I., Perantonis, S.J.: Automatic table detection in document images. In: Proceedings of the Third International Conference on Advances in Pattern Recognition (ICAPR'05), Lecture Notes in Computer Science (3686), pp. 609–618, Path, UK (2005)
11. Gatos, B., Mantzaris, S.L., Chandrinos, K.V., Tsigris, A., Perantonis, S.J.: Integrated algorithms for newspaper page decomposition and article tracking. In: Proceedings Fifth International Conference on Document Analysis and Recognition (ICDAR'99), September 1999, pp. 559–562 (1999)
12. Gatos, B., Pratikakis, I., Perantonis, S.J.: An adaptive binarisation technique for low quality historical documents In: IAPR Workshop on Document Analysis Systems (DAS2004), Lecture Notes in Computer Science (3163), September 2004, pp. 102–113 (2004)
13. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recogn. **39**, 317–327 (2006)
14. Govindan, V.K., Shivaprasad, A.P.: Character recognition—a review. Pattern Recogn. **23**(7), 671–683 (1990)
15. Guillevic, D., Suen, C.Y.: HMM word recognition engine. In: Fourth International Conference on Document Analysis and Recognition (ICDAR'97), pp. 544–547 (1997)
16. Keaton, P., Greenspan, H., Goodman, R.: Keyword spotting for cursive document retrieval. In: Workshop on Document Image Analysis (DIA 1997), pp. 74–82 (1997)
17. Lavrenko, V., Rath, T.M., Manmatha, R.: Holistic word recognition for handwritten historical documents. In: Proceedings of the International Workshop on Document Image Analysis for Libraries. pp. 278–287 (2004)
18. Lu, Y., Tan, C., Weihua, H., Fan, L.: An approach to word image matching based on weighted Hausdorff distance. In: Sixth International Conference on Document Analysis and Recognition (ICDAR'01), September 2001, pp. 10–13 (2001)
19. Madhvanath, S., Govindaraju, V.: Local reference lines for handwritten word recognition. Pattern Recogn. **32**, 2021–2028 (1999)
20. Manmatha, R., Croft, W.B.: Word spotting: indexing handwritten manuscripts. In: Intelligent Multimedia Information Retrieval. MIT, Cambridge, MA, Maybury, pp. 43–64 (1997)
21. Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 2005
22. Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: Digital Libraries '96: First ACM International Conference on Digital Libraries, pp. 150–159 (1999)
23. Marcolino, A., Ramos, V., Ármalo, M., Pinto, J.C.: Line and Word matching in old documents. In: Proceedings of the Fifth IberoAmerican Sympsium on Pattern Recognition (SIAPR'00), September 2000, pp. 123–125 (2000)
24. Niblack, W.: An Introduction to Digital Image Processing. Prentice Hall, Englewood cliffs (1996)
25. Rath, T.M., Manmatha, R., Lavrenko, V.: A search engine for historical manuscript images. In: ACM SIGIR conference, pp. 369–376, (2004)
26. Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 521–527 (2003)
27. Rath, T.M., Manmatha, R.: Features for word spotting in historical documents. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), pp. 218–222 (2003)
28. Rijsbergen van K. Information Retrieval. http://www.dcs.gla.ac.uk/Keith/Preface.html
29. Seni, G., Cohen, E.: External word segmentation of off-line handwritten text lines. Pattern Recogn. **27**(1), 41–52 (1994)
30. Theodoridis, S., Koutroumbas, K.: Pattern recognition. Academic, New York (1997)
31. Vinciarelli, A., Bengio, S., Bunke, H.: Offline recognition of unconstrained handwritten texts using hmms and statistical language models. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 709–720 (2004)

32. Waked, B., Suen, C. Y., Bergler, S.: Segmenting document images using diagonal white runs and vertical edges. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pp. 194–199 (2001)
33. Wang, J., Leung, K.H., Hui, S.C.: Cursive word reference line detection. Pattern Recogn. **30**(3), pp. 503–511 (1997)
34. Weihua, H., Tan, C.L., Sung, S.Y., Xu, Y.: Word shape recognition for image-based document retrieval. In: International Conference on Image Processing, ICIP'2001, October 2001, pp. 8–11 (2001)
35. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block segmentation and text extraction in mixed text/image documents. Comput. Graph. Image Process. **20**, 375–390 (1982)