2009 10th International Conference on Document Analysis and Recognition

Segmentation-free Word Spotting in Historical Printed Documents

B. Gatos and I. Pratikakis

Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
{bgat, ipratika}@iit.demokritos.gr

Abstract

In this paper, a new efficient word spotting methodology is presented that can be applied to historical printed documents without requiring any previous block or word segmentation step. Our aim is to address a methodology which is segmentation-free since in many cases of historical documents, the segmentation process does not produce meaningful results due to unconstraint layout, several degradations or typesetting imperfections. The proposed method is based on block-based document image descriptors that are used at a template matching process satisfying invariance in terms of translation, rotation and scaling. Improvement in terms of time expense is obtained by applying the matching process only on salient regions of the image. Experimental results on a database with representative historical printed documents prove the efficiency of the proposed approach.

1. Introduction

Effective historical document indexing and retrieval poses a great challenge due to the vast amount of information that is available in libraries all over the world in the form of printed or handwritten manuscripts. The challenge is amplified by the variability of documents due to the multi-linguality and the wide range of historical periods that available collections are built, as well as by the poor quality of existing historical documents.

Word spotting is a content-based retrieval procedure which results in a ranked list of word images that are similar to a query word image. The query comprises either an actual example from the collection of interest or it is artificially generated from an ASCII keyword. A crucial aspect in the retrieval procedure is the word image representation which relies upon robust features. The word spotting procedure is mostly used in an unsupervised manner and the lack of dependencies like training along with the ease to use several different feature variations make it as a very appealing alternative to Optical Character Recognition (OCR) which is a difficult problem to solve, especially for historical documents.

In the literature, word spotting appears under two distinct trends: the segmentation-based approach and the segmentation-free approach. In the former approach, there is a tremendous effort towards solving the word segmentation problem [1-4].

In the latter approach, the query word image is fitted to the corresponding word images in the document without any segmentation involved, mostly seen the underlying problem as a template matching. Representative work is reported in [5], which uses differential features that are compared using a cohesive elastic matching method, based on zones of interest in order to match only the informative parts of the words.

In the same spirit with the aforementioned approach, this paper concerns a segmentation-free word spotting methodology which permits a fast and effective retrieval based on block-based document image descriptors that are used at a template matching process satisfying invariance in terms of translation, rotation and scaling.

The remainder of the paper will be structured as follows. The proposed methodology is detailed in Section 2. In Section 3, the evaluation results on representative historical documents are presented, and in Section 4, conclusions are drawn.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.236

2. Proposed methodology

The proposed word spotting methodology receives as input the word image query ***W*** and the document image ***I*** and produces as output the word spotting results which correspond to a set of rectangular areas of image ***I*** which delimit the word images that match the word image query. It consists of several distinct steps: (a) image pre-processing and size normalization of both images ***I*** and ***W***; (b) detection of salient regions in image ***I*** based on a RLSA smoothing [6]; (c) block-based feature extraction of image ***I*** as well of image ***W***. For the word image query ***W***, it is not only considered the original image but also more instances are considered which are produced after a set of stepwise transforms with respect to translation, rotation and scaling; (d) efficient and fast word matching which is not based on any segmentation step and also involves a procedure to combine several matching results. The proposed methodology is detailed in this section while a flowchart is presented in Fig. 1.

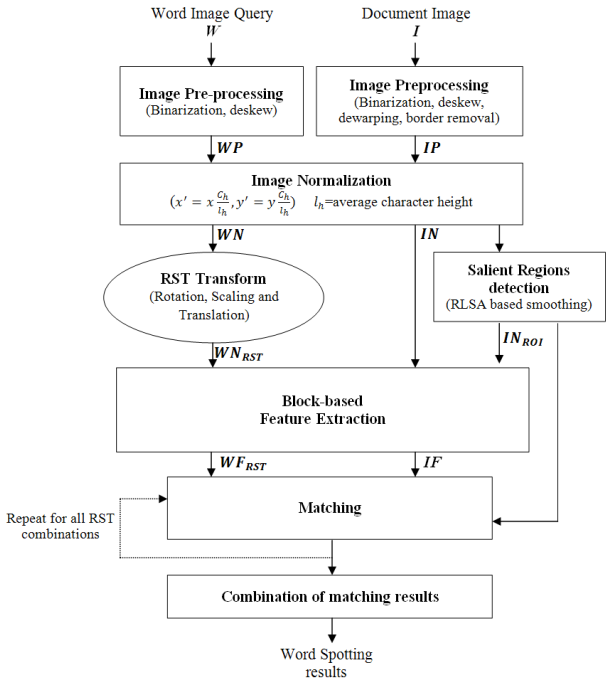


Figure 1. The flowchart of the proposed segmentation-free word spotting methodology.

2.1. Image pre-processing and normalization

First, we proceed to a pre-processing of document image ***I*** which involves document image binarization [7], skew and warping correction [8] as well as border detection [9]. We also proceed to the pre-processing of word image query ***W*** which involves document image binarization and deskew. In this way we produce images ***IP*** and ***WP***, respectively. Then, the average character height of images ***IP*** and ***WP*** is computed based on their connected component histograms [3]. This computation is used for an image normalization of both ***IP*** and ***WP*** using a common average character height equal to C_H. Parameter C_H is defined so as to have a meaningful feature vector representation (in our experiments we use C_H=20).

2.2. Detection of Salient Regions

For a fast and efficient word spotting procedure, it is desired to constrain the applied matching process only on certain regions of interest. These regions should correspond to the text regions of image ***I***. For this purpose, we apply to ***IN*** a horizontal RLSA [6] with threshold equal to 2* C_H which corresponds to a rough text line estimation outcome. After this process, we produce image $\boldsymbol{IN_{ROI}} \in \{0,1\}$ (see Fig. 2c).

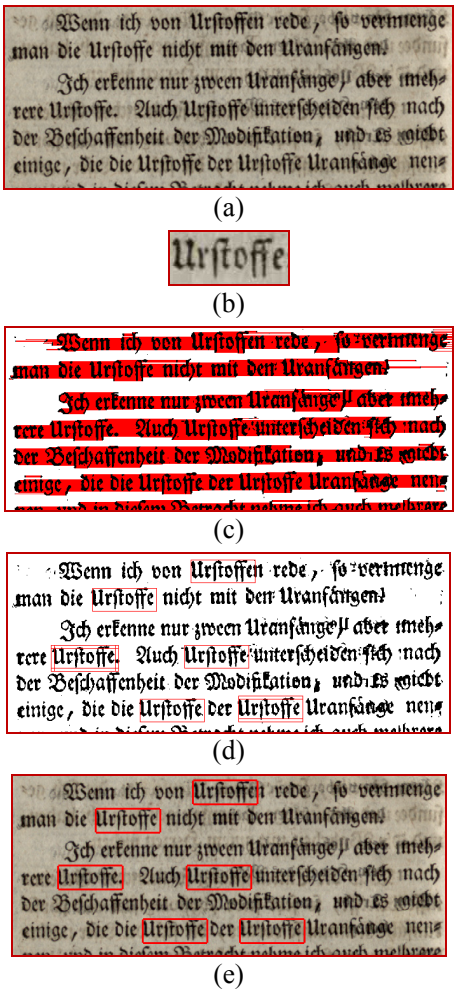


Figure 2. (a) Original image ***I***; (b) word image query ***W***; (c) salient regions; (d) matching results using several word instances; (e) final word spotting results.

2.3. Block-based feature extraction

At this step, we first produce several word instances of image ***WN*** in order to capture transformation variations in terms of rotation and scaling. We apply rotation which range from -1^0 to 1^0 by a step of 1^0 and resizing with factor which ranges from 0.8 to 1.2 by a step of 0.1. In this way, we produce 3x5=15 different word instances $\boldsymbol{WN_{RS}}$, R=1..3 and S=1..5. Then, for every word instance $\boldsymbol{WN_{RS}}(x,y) \in \{0,1\}$, $x=1..w_x$, $y=1..w_y$, we calculate 5 different set of feature vectors based on calculating all 5x5 non-overlapping window pixel density and applying word image translation at (-2,2), (2,2), (2,-2) and (-2,-2). More specifically, the 5 feature vectors are given by the following equations:

$$\boldsymbol{WF_{RS1}}(k,s) = \sum_{x=k*5+1,\ y=s*5+1}^{x=k*5+5,\ y=s*5+5} \boldsymbol{WN_{RS}}(x,y) \quad [1]$$

$$\boldsymbol{WF_{RS2}}(k,s) = \sum_{x=k*5+1,\ y=s*5+1}^{x=k*5+5,\ y=s*5+5} \boldsymbol{WN_{RS}}(x-2,y-2) \quad [2]$$

$$\boldsymbol{WF_{RS3}}(k,s) = \sum_{x=k*5+1,\ y=s*5+1}^{x=k*5+5,\ y=s*5+5} \boldsymbol{WN_{RS}}(x-2,y+2) \quad [3]$$

$$\boldsymbol{WF_{RS4}}(k,s) = \sum_{x=k*5+1,\ y=s*5+1}^{x=k*5+5,\ y=s*5+5} \boldsymbol{WN_{RS}}(x+2,y-2) \quad [4]$$

$$\boldsymbol{WF_{RS5}}(k,s) = \sum_{x=k*5+1,\ y=s*5+1}^{x=k*5+5,\ y=s*5+5} \boldsymbol{WN_{RS}}(x+2,y+2) \quad [5]$$

where $k \in [0..k_m], s \in [0..s_m], s_m = \left\lfloor \frac{w_x}{5} \right\rfloor, k_m = \left\lfloor \frac{w_y}{5} \right\rfloor$.

In eq. [1]-[5] we assume $\boldsymbol{WN_{RS}}(x,y)$=0 if $x<1$ or $x>w_x$ or $y<1$ or $y>w_y$.

It is worth noting that we calculate the feature vector of document image $\boldsymbol{IN}(x,y) \in \{0,1\}$, $x=1..i_x$, $y=1..i_y$ which has non-zero values only at the salient regions (see section 2.2):

$$\boldsymbol{IF}(\lambda,\mu) = \boldsymbol{IN_{ROI}}(\lambda*5+3, \mu*5+3) \sum_{x=\lambda*5+1,\ y=\mu*5+1}^{x=\lambda*5+5,\ y=\mu*5+5} \boldsymbol{IN}(x,y) \quad [6]$$

where $\lambda \in [0..\lambda_m], \mu \in [0..\mu_m], \lambda_m = \left\lfloor \frac{i_x}{5} \right\rfloor, \mu_m = \left\lfloor \frac{i_y}{5} \right\rfloor$,

$\boldsymbol{IN}(x,y)$=0 if $x<1$ or $x>i_x$ or $y<1$ or $y>i_y$. Involving $\boldsymbol{IN_{ROI}}$ in eq.6 we provide an important acceleration in the feature extraction procedure.

2.4. Word matching

At the word matching step, all word feature vectors $\boldsymbol{WF_{RST}}$ are compared with the corresponding feature vectors of image ***IF*** by applying a matching which is constrained by the regions of interest $\boldsymbol{IN_{ROI}}$. More specifically, we consider a successful match of word instance $\boldsymbol{WN_{RST}}(x,y)$, $x=1..w_x$, $y=1..w_y$ at the rectangular area (λ*5+1, μ*5+1) - (($\lambda+k_m+1$)*5, ($\mu+s_m+1$)*5) of image ***IN*** only if:

$$\boldsymbol{IN_{ROI}}(\lambda*5+3, \mu*5+3) == 1 \quad [7]$$

and

$$\prescript{\lambda}{\mu}{\boldsymbol{D}}_{\boldsymbol{RST}} = \sum_{k=0,s=0}^{k=k_m,s=s_m} (\boldsymbol{WF_{RST}}(k,s) - \boldsymbol{IF}(\lambda+\kappa, \mu+s))^2 < th \quad [8]$$

where $k_m = \left\lfloor \frac{w_x}{5} \right\rfloor, s_m = \left\lfloor \frac{w_y}{5} \right\rfloor$ and $th = 2(k_m+1)(s_m+1)$.

Since $\prescript{\lambda}{\mu}{\boldsymbol{D}}_{\boldsymbol{RST}}$ can take values from 0 (best match) to $25(k_m+1)(s_m+1)$ (worse match) we select threshold th assuming an ~90% perfect match. An example of matching results using several word instances is given at Fig. 2d.

The last step concerns the combination of all matching results in order to produce the final word spotting result. According to our approach, if we have several intersecting rectangular areas that correspond to successful matching results then we select only the one rectangular area that corresponds to the lower $\prescript{\lambda}{\mu}{\boldsymbol{D}}_{\boldsymbol{RST}}$ value (see Fig. 2e).

3. Experimental results

We tested our methodology on a historical book from Eckartshausen which was published on 1788 and is owned by the Bavarian State Library [10] (see Fig.3a). The images of the book suffer from several problems such as degradations, typesetting imperfections and non-uniform spacing between words which do not permit the application of any segmentation task. In Fig. 4 we demonstrate the results of applying binarization using [7] and word segmentation tasks using OCRopus [11] and FineReader Engine 8.1 [12] to a page from our test set.

We selected 100 pages from this book and manually marked 5 keywords (see Fig. 3b). These keywords are semantically significant and frequently repeated in the

book. In the selected pages we marked 207 instances of all keywords.

Then, we applied the proposed word spotting methodology. The time needed for searching a keyword on a document page image is about 2 sec using a Core Duo PC at 2.0 GHz. A word spotting result is demonstrated in Fig. 5.

Let N 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:

$$RC = \frac{Corr}{N} 100\% \quad [9]$$

$$PR = \frac{Corr}{M} 100\% \quad [10]$$

$$FM = \frac{2 * RC * PR}{RC + PR} \quad [11]$$

Table 1 presents the word spotting results for the 5 keywords in terms of recall and precision as well the F-measure. As it can be observed, we can achieve high recall rates (93.2% on average) while keeping the precision on acceptable levels (75.1% on average) resulting to an F-measure equal to 83.2% on average.

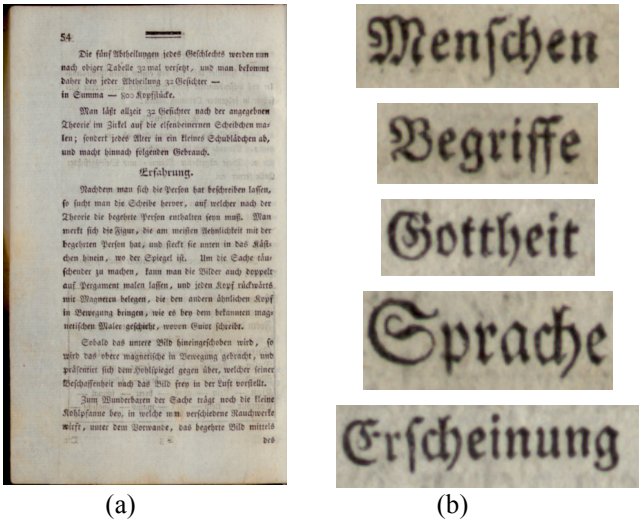
4. Conclusions

A new efficient word spotting methodology is presented that can be applied to historical printed documents without requiring any previous block or word segmentation step. It is based on block-based document image descriptors that are used at a template matching process satisfying invariance in terms of translation, rotation and scaling. For a fast and efficient word spotting procedure, we constrain the applied matching process only on certain regions of interest.

Taking into account the low quality of the documents considered in this experimentation as well as the fast application of the underlying matching procedure, we believe that the proposed approach is very promising for the historical document image indexing and retrieval.

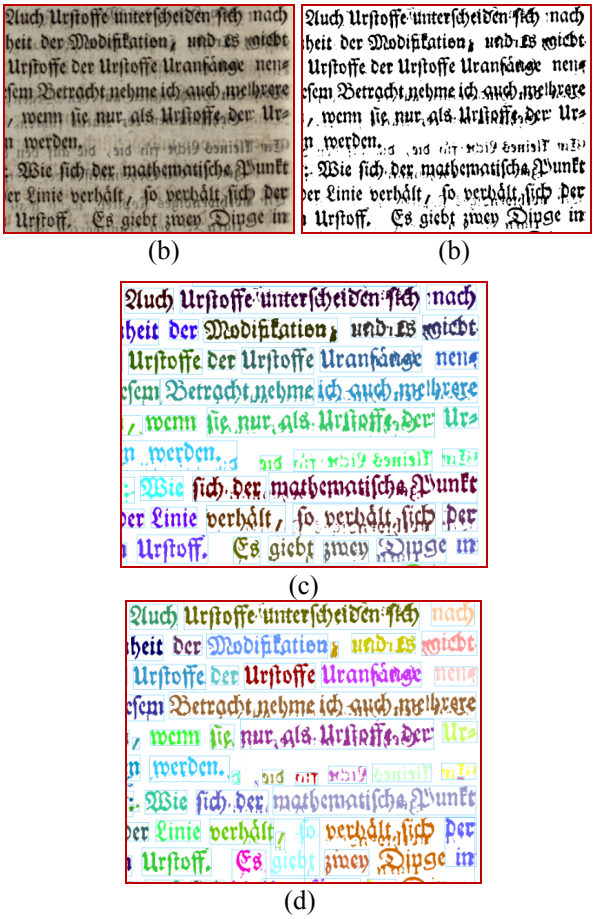
Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).



(a) (b)

Figure 3. (a) an image sample from Eckartshausen book (1788, Bavarian State Library); (b) the 5 keywords selected for testing.

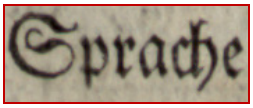


(b) (b)

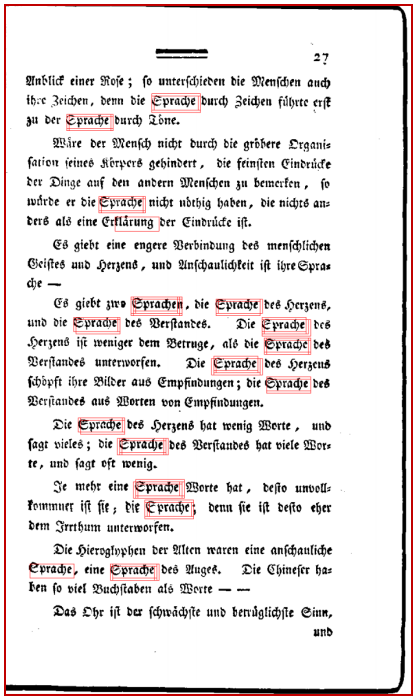
(c)

(d)

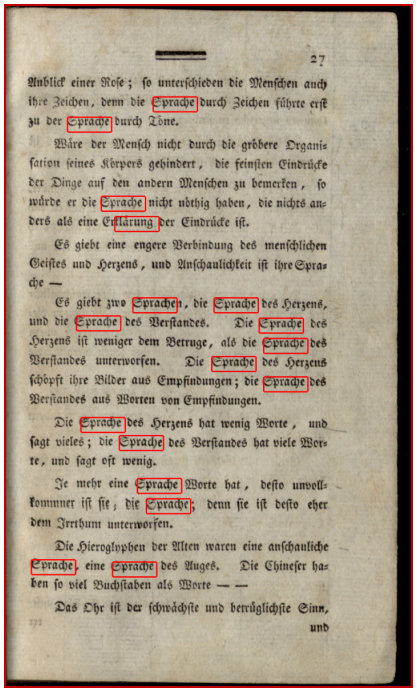
Figure 4. (a) Image portion from our test set; (b) binarization result [7]; (c) word segmentation result using OCROpus [11]; (d) word segmentation result using FineReader engine 8.0 [12].



(a)



(b)



(c)

Figure 5. Word spotting result demonstration: (a) keyword image; (b) word matching result; (c) final word spotting result.

Table 1. Word spotting results in terms of recall and precision for the 5 keywords used in the experiments.

	N	M	Corr	RC(%)	PR(%)	FM(%)
Menſchen	81	95	72	88.9	75.8	81.8
Begriffe	29	38	29	100	76.3	86.6
Gottheit	32	43	31	96.9	72.1	82.7
Sprache	30	42	30	100	71.4	83.3
Erſcheinung	35	39	31	88.6	79.5	83.8
TOTAL	**207**	**257**	**193**	**93.2**	**75.1**	**83.2**

References

[1] T. M. Rath, and R. Manmatha, "Features for word spotting in historical documents" *In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)* , 2003, pp 218-222.

[2] A. Balasubramanian, M. Meshesha, and C. V. Jawahar "Retrieval form Document Image Collections", *DAS 2006*, 2006, pp 1-12.

[3] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback", *International Journal of Document Analysis and Recognition*, vol. 9, 2007, pp. 167-177.

[4] A. Bhardwaj, D. Jose and V. Govindaraju, "Script Independent word spotting in multilingual documents", *In Proceedings of the 2nd International Workshop on Cross Lingual Information Access*, 2008, pp. 48-54.

[5] Y. Leydier, F. Lebourgeois and H. Emptoz, "Text search for medieval manuscript images", *Pattern Recognition*, vol. 40, 2007, pp. 3552-3567.

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

[7] B. Gatos, I. Pratikakis and S. J. Perantonis, "Adaptive Degraded Document Image Binarization", *Pattern Recognition*, Vol. 39, 2006, pp. 317-327.

[8] N. Stamatopoulos, B. Gatos, I. Pratikakis and S. J. Perantonis, "A Two-Step Dewarping of Camera Document Images", *8th International Workshop on Document Analysis Systems (DAS'08),* 2008.

[9] N. Stamatopoulos, B. Gatos and A. Kesidis, "Automatic Borders Detection of Camera Document Images", *2nd International Workshop on Camera-Based Document Analysis and Recognition (CBDAR'07)*, 2007, pp. 71-78.

[10] Carl von Eckartshausen, "Aufschlüsse zur Magie aus geprüften Erfahrungen über verborgene philosophische Wissenschaften und verdeckte Geheimnisse der Natur", Bavarian State Library, 1778.

[11] OCRopus, http://sites.google.com/site/ocropus/

[12] FineReader Engine 8.1, http://www.abbyy.com/