2014 14th International Conference on Frontiers in Handwriting Recognition

Segmentation-based Historical Handwritten Word Spotting using Document-Specific Local Features

Konstantinos Zagoris
IIT, NCSR 'Demokritos'
Athens, Greece.
&
DUTH,Visual Computing Group
Dept. of Electrical and Computer Engineering
Xanthi, Greece
Email: kzagoris@ee.duth.gr

Ioannis Pratikakis
DUTH, Visual Computing Group
Dept. of Electrical and Computer Engineering
Xanthi, Greece
Email: ipratika@ee.duth.gr

Basilis Gatos
IIT, NCSR 'Demokritos'
Athens, Greece.
Email: bgat@iit.demokritos.gr

***Abstract*—Many word spotting strategies for the modern documents are not directly applicable to historical handwritten documents due to writing styles variety and intense degradation. In this paper, a new method that permits effective word spotting in handwritten documents is presented that relies upon document-specific local features which take into account texture information around representative keypoints. Experimental work on two historical handwritten datasets using standard evaluation measures shows the improved performance achieved by the proposed methodology.**

Keywords*—*Word Spotting, Handwritten Documents, Local Features

I. Introduction

Currently, many digitized historical manuscripts are not exploited due to lack of proper browsing and indexing tools. A valid strategy to deal with this kind of unindexed documents is a word matching procedure that relies upon a low-level pattern matching called word spotting [1]. It can be defined as the task of identifying locations on a document image which have high probability to contain an instance of a queried word, without explicitly recognizing it. Word spotting in document images is related to Content-Based Image Retrieval systems by searching a word image from a set of unindexed document images using the image content as the only information source. As final outcome, the system returns to the user a ranked list of document word images.

In the literature, word spotting appears under two distinct trends wherein the fundamental difference concerns the search space which could be either a set of segmented word images (segmentation-based approach) or the complete document image (segmentation-free approach). In this work, we address the word spotting problem with a segmentation-based approach.

Initial efforts in segmentation-based word spotting followed a methodological pipeline using as a first step, advanced procedures for binarization, pre-processing and text layout analysis towards word image segmentation. Then, analyzing the segmented word image, descriptors are extracted. Based on these descriptors, a distance measure is used to measure the similarity between the query word image and the segmented word image. Although there is an abundance of systems suitable for both modern [2], [3] and historical [4], [5], [6], [7] printed material, very few of these systems are suitable to handwritten documents due to noise sensitivity, character variation and text layout complexity.

Rath and Manmatha [8], [9], [10] calculate two families of feature sets. On the one hand, they have the scalar type features that include aspect ratio, area, etc. On the other hand, there are the profile-based features that are based on horizontal and vertical words projections and the upper and lower word profiles. Zagoris et. al. [11] created a similar set of profile-based features, encoded in a different way by Discrete Cosine Transformation, normalization by the first coefficients and quantization through the Gustafson - Kessel [12] fuzzy algorithm. The result was a very short stable-length descriptor, which has been tested on a Greek handwriting database from different writers, the Washington words database and the MPEG-7 CE1 Set B database. Rodriguez and Perronnin [13] extract features from a sliding window, based on the first gradient and inspired by the SIFT keypoint descriptor. Finally, Srihari et al. [14] present a system for searching handwritten Arabic documents based on a set of binary shape features suitable for Arabic script along with a correlation distance that performed best for matching those features.

Recently, there was an influx of works based on the local features in the form of the Bag-of-Visual Words model [15], [16], [17], [18]. Llados et. al. [16] evaluate the performance of various word descriptors, including a bag of visual words procedure (BoVW), a pseudo-structural representation based on Loci Features, a structural approach by using words as graphs, and sequences of column features based on DTW. They found that the statistical approach of the BoVW produces the best results, although the memory requirements to store the descriptors are significant.

Most works using local features are based on the Scale Invariant Feature Transform (SIFT) [19] in order to describe the local information. The original application of these local features are the natural images which they have many structural differences compared to document images. Firstly, the detection of the most powerful edges through pyramid scaling, creates local points between text lines. Secondly,we argue, that it is not beneficial in document images to incorporate

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.10

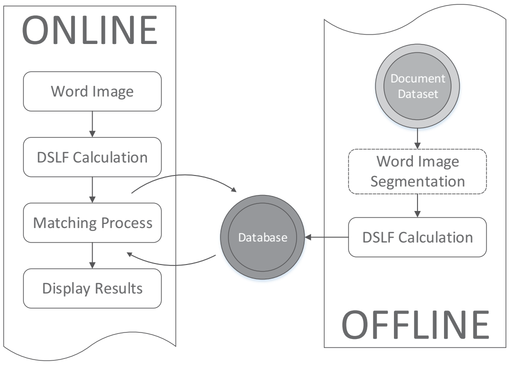


Fig. 1. Global diagram of the proposed word spotting framework

invariant properties in the descriptor of the local points as it results in noise amplification. This is further supported by the observation stated in [20] wherein the used features which are invariant to rotation have resulted in worse performance, when compared to features that are dependent on rotation. They adhere to the observation that the features that are invariant to rotation are more sensitive to the noise and the complex texture of the background. Moreover, it is worth noting that the features for word spotting which rely only on word shape characteristics are not effective in dealing with a document collection created by different writers, containing significant writing style variations. Although slant and skew preprocessing techniques can reduce the shape variations, they cannot eliminate the problem as the whole structure of the word is different in most of the cases. In this respect, we argue that although the shape information is meaningful, the texture information in a spatial context is more reliable.

Taking into account the aforementioned considerations, the proposed segmentation-based approach employ novel local features which are specific for documents, namely Document Specific Local Features (DSLF). Moreover, although the proposed features use the spatial information of the current points location they are based on texture information. For the sake of clarity, it is worth to note that since the focus of this work is on features extraction and matching, the segmented word images used in the proposed approach are achieved from the available ground truth dataset without involving any particular word image segmentation method.

The remainder of the paper is organized as follows: Section II describes the architecture of the word spotting framework, Section II-A details the keypoints detection and DSLF features and in Section II-B the matching procedure is described. Finally, Section III presents the experiments work while in Section IV conclusions are drawn.

II. Proposed Methodology

A global systematic diagram of the proposed word spotting framework is illustrated in Fig. 1. It consists of two distinct steps: the Offline and the Online. At the Offline step, which is executed once, the document images are segmented to the word images for which, the proposed local features are extracted and indexed to a database.

At the Online step, which is the only visible operation to the user, the DSLF are extracted for the query word image and a matching procedure is addressed between the features of the query and the feature set of each indexed word image. Finally, a ranking list of all the word images are presented to the user.

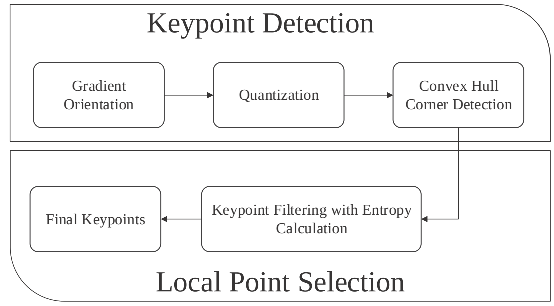


Fig. 2. The steps for the detection and selection of characteristic keypoints

A. *Document-Specific Local Features (DSLF)*

1) KeyPoints Detection: Fig. 2 shows the consecutive steps of the proposed methodology for the detection of characteristic local points (keypoints) in a document image.

Initially, the gradient vector G^k of the image k and its orientation (G^k) is calculated and they are defined as:

$$G^k = \begin{pmatrix} I_x^* \\ I_y^* \end{pmatrix}, \theta(G^k) = tan^{-1}\begin{pmatrix} I_x^* \\ I_y^* \end{pmatrix} \quad (1)$$

where I_x^* and I_y^* are calculated by the convolution of the 1-D kernels $[-1, 0, 1]$ and $[-1, 0, 1]^T$ to the grey-level image k, respectively. The resulting I_x^* and I_y^* is very sensitive to noise. In order to make them most robust, we assume that I_x^* and I_y^* contains two distinct clusters: noise and meaningful data. The estimation of the threshold that separates these two clusters is achieved by minimizing the intra-class variance between the clusters as in the Otsu approach [21]. Finally, the values that are below this dynamic threshold are rejected.

Then, the gradient orientation is calculated (Eq. 1). Fig. 3b shows an example of this feature using a grey-level representation. The grey values are not taken into consideration by the algorithm in the consecutive steps. The dark colours represent negative angles while the bright colours represent positive angles. The orientation of the gradient describes the uniformity in term of stroke orientation.

Gradient orientation features are also used in character recognition. Most of the authors use 4- or 8-direction histograms computed in zones [22], [23]. The first order feature is variant to image rotation, which it is an intent effect as we are argued before, that invariant properties results to noise amplification and poorer performance when dealing with document images.

The next step involves linear quantization of the $\theta(G^k)$ to n levels. The purpose of this step is to detect the changes to the writing direction as these points consist of important and descriptive information. The quantization levels n is a parameter of the proposed algorithm and controls the amount of the final local points. Increasing the quantization levels, the

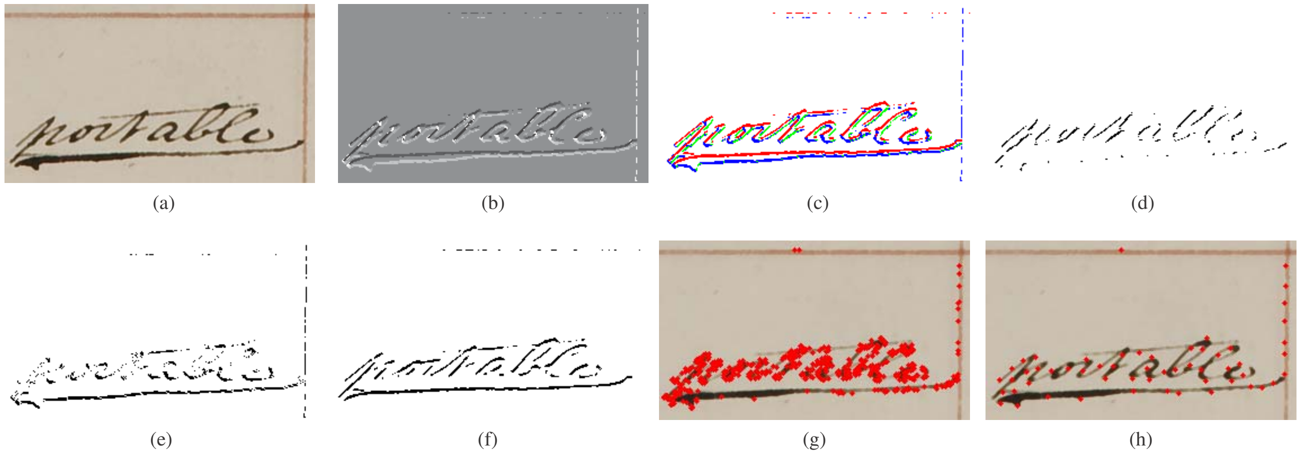


Fig. 3. The steps for the keypoint detection: (a) original document image, (b) orientation of the gradient vector, (c) quantization of the gradient vector orientation, (d-f) connected components of the distinct quantization levels, (g) the initial keypoints and (h) the final keypoints

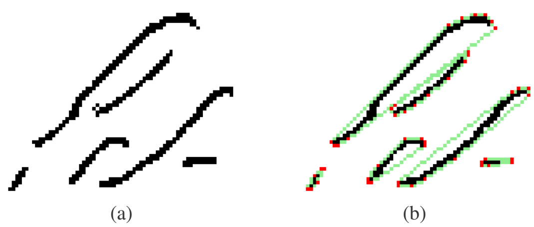


Fig. 4. Detection of keypoints : (a) Example connected components;(b) the corner points of the convex hull as the keypoints of the connected components

process becomes more sensitive to the writing directions. After extensive experimentation it is proven that 3 levels are enough in the context of word spotting retrieval. Fig. 3c shows the output for the quantization of the $\theta(G^k)$ values to 3 levels.

Example of connected components (CCs) attributed to each quantization level is shown at Fig. 3d-3f, respectively. These CCs represent chunks of strokes that have different writing directions between them. The most descriptive and important points on those CCs are the endpoints that appear along the edge. For the endpoint detection, the convex hull that contains one CC is taken into consideration wherein its corners are nominated as the initial keypoints kP. Fig. 4 shows a visual representation of the CCs convex hull (with green colour) and the red dots are the initial keypoints. The endpoint detection step is applied for each CC at each quantization level. It is worth mentioning that aiming to decrease the performance cost, small CCs, i.e. width and height less than 5 pixels, are rejected. Fig. 3g shows an example of the initial keypoints.

The next steps involve the selection of those kPs that relate to the most descriptive information of the word image. This is achieved, by initially calculating the entropy of the quantized gradient angles around the keypoint using the Shannon entropy equation:

$$E_W = -\frac{1}{N}\sum_{i\in W}\left(occ(\theta_i(G^k))\cdot ln\left(\frac{occ(\theta_i(G^k))}{N}\right)\right), \quad N = \sum_{i\in W} occ(\theta_i(G^k)) \tag{2}$$

where W denotes the pixels in the window and $occ(\theta_i(G^k))$ denotes the occurrence of the corresponding quantized gradient angle in W.

The kPs are finally selected starting from those that have the maximum entropy and if other kPs are found in their neighbourhood W, then those kPs are rejected. The remaining points are those that contain the maximum entropy in their neighbourhood and, consequently, are the most significant. The size of window W is a parameter that directly correlates to the final output numbers of the kPs detection algorithm. It is proposed to be W=18x18 which equals to the size of the neighbourhood that the local point descriptor is calculated upon as is detailed in the next section.

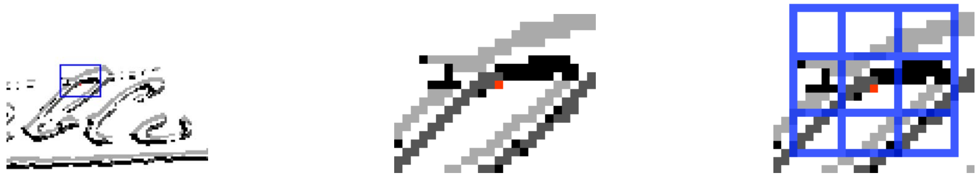
2) Feature Extraction: The feature for the local keypoint is calculated upon the quantized gradient angles. An area of 18x18 pixels around the kP, is divided into 9 cells with size 6x6 for each of them, as shown at Fig. 5. Each cell is represented by a 3-bin histogram (each bin corresponds to a quantization level) and each pixel accumulates a vote in the corresponding angle histogram bin. The strength of voting depends on the norm of the gradient vector and on the distance from the location of local point as shown at the following equation:

$$V_{x,y} = s_{x,y}\cdot \|G_{x,y}\| \tag{3}$$

where the $V_{x,y}$ is the contribution of pixel (x, y) to the distribution of the corresponding cell histogram, $\|G_{x,y}\|$ is the norm of the gradient vector G, and $s_{x,y}$ denotes a *linear weighting factor* of the distance between pixel (x, y) and the keypoint, based on the following equation:

$$s_{x,y} = 1 - \frac{2}{3}\cdot\frac{\sqrt{(x-x_{LP})^2 + (y-y_{LP})^2}}{9\sqrt{2}} \tag{4}$$

where (x_{LP}, y_{LP}) denotes the position of the LP. The weighting factor values range in the interval $[1/3, 1]$ for which the maximum value is achieved for $x = x_{LP}, y = y_{LP}$ while its minimum is reached at $x = x_{LP}\pm 9, y = y_{LP}\pm 9$. The



(a) (b) (c)

Fig. 5. (a) example keypoint at the global connected components level; (b) example keypoint at the local level; (c) the neighborhood used for the computation of the descriptor at the keypoint.

task of the variable $s_{x,y}$ is to weigh the pixel participation to the histogram taking into account its distance from the kP.

Finally, all nine (9) histograms are concatenated in one 27-bin histogram and normalized by its norm. In order to make the descriptor illumination independent all the values above 0.2 are fixed to 0.2 and the resulting values are re-normalized again [19].

B. Matching in a Segmentation-based Word Spotting context

In the case of segmentation-based word spotting, the aim is to match the query keypoints to the corresponding keypoints of any word image in the document. For this task, the descriptor that has been presented is taken into consideration along with a Local Proximity Nearest Neighbor (LPNN) search. The advantage of LPNN search is two-fold: (i) it enables a search in focused areas instead of searching in a brute force manner and (ii) it goes beyond the typical use of a descriptor by the incorporation of spatial context in the local search addressed. In the sequel, the complete matching step will be detailed.

The initial stage in the matching step is a normalization which is applied for any word image including the query word image. The aim of this stage is to alleviate any scale variability for the same word. The normalized procedure comprises the following steps:

Calculation of the mean center in the x- and y-axis of the keypoints set in a word image:

$$(c_x, c_y) = \left(\frac{\sum_{i=1}^{k} p_x^i}{k}, \frac{\sum_{i=1}^{k} p_y^i}{k} \right) \tag{5}$$

where p_x^i, p_y^i denote the location of the i^{th} keypoint and k denotes the total number of the keypoints in a word image.

Calculation of the mean distance of each keypoint from the mean center:

$$D_x = \frac{\sum_{i=0}^{k} \left| p_x^i - c_x \right|}{k}, D_y = \frac{\sum_{i=0}^{k} \left| p_y^i - c_y \right|}{k} \tag{6}$$

Calculation of the updated location for each keypoint addresses transformation to a new space wherein $[c_x, c_y]$ is the new coordinate origin:

$$p_x^{i\,\prime} = \frac{p_x^i - c_x}{D_x}, p_y^{i\,\prime} = \frac{p_y^i - c_y}{D_y} \tag{7}$$

After normalization, all word images are directly comparable due to the achieved scale invariance as seen in Fig. 6.

In the next stage, the LPNN for each keypoint that resides on the query image is addressed. LPNN is realized in a search area which is computed by taking into account a percentage (25%) of the already calculated distances Dx, Dy. During search, if there is one or more word keypoints in the proximity of the query keypoint under consideration, the Euclidean distance between their descriptors is calculated and the minimum distance is kept. This is repeated for each keypoint in the query image. The final similarity measure is the sum of all the minimal distances. If there is not a local point in its proximity then a penalty value is added to the similarity measure and it is equal to maximum Euclidean distance that can be calculated between the keypoint descriptors, which results in a value of $\sqrt{27}$.

As a final stage, the system presents to the user all the word images based on ascending sort order of the calculated similarity measure.

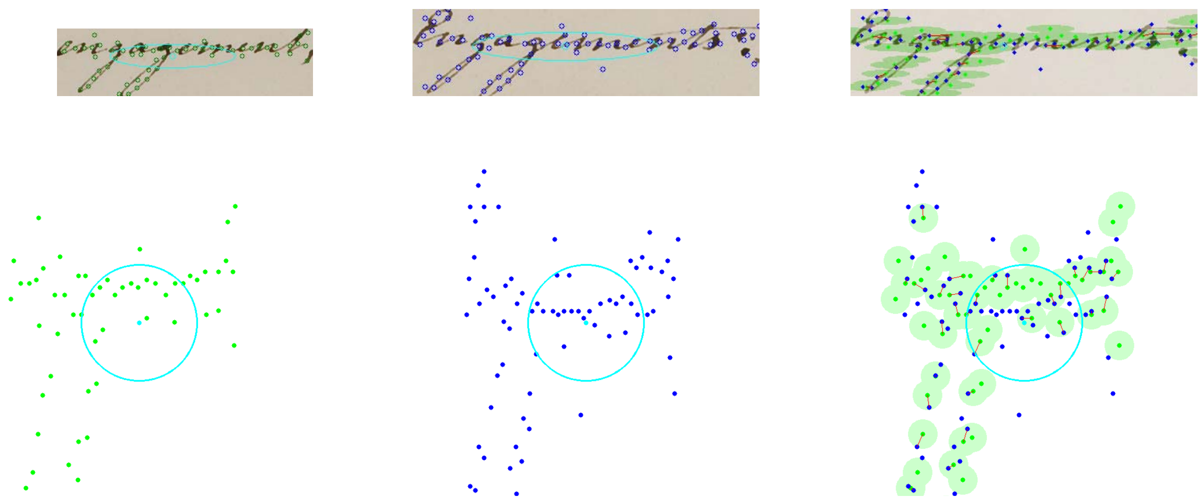
III. Experimental Results

The proposed segmentation-based word spotting approach is evaluated on two handwritten datasets:

- **Bentham Dataset [24]:** It consists of 50 high quality (approximately 3000 pixel width and 4000 pixel height) handwritten manuscripts written by Jeremy Bentham (1748-1832) himself over a period of sixty years, as well as fair copies written by Bentham's secretarial staff. It contains several very difficult problems, wherein the most difficult is the word variability. The variation of the same word is extreme and involves writing style, font size, noise as well as their combination. Fig. 7 shows some examples of these instances.
- **Washington Dataset [10]:** It consists of 20 document images from George Washington Collection of the Library of Congress [10]. The documents are were scanned from microfilm in 300 dpi resolution.

Fig. 8 shows some representative document images from both datasets.

The measures employed in the performance evaluation of the proposed segmentation-based algorithm are the Precision at the 5 Top Retrieved words (P@5) and the Mean Average Precision (MAP). To further detail the metrics, let define Precision and P@k as follows:



(a) (b) (c)

Fig. 6. (a) the query keypoints in the image and the normalized space, (b) the word image keypoints in the image and the normalized space, (c) the projection of query keypoints to the word image and the normalized space: the red lines are the matching local points, the green area is the local proximity area of the nearest neighbor search. The circle in the normalized space has radius 1.

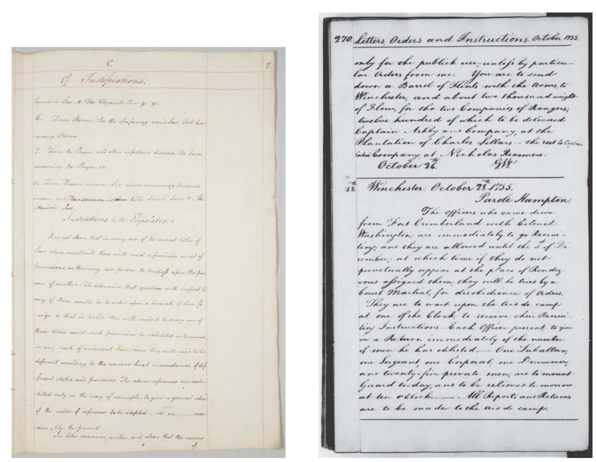
Fig. 7. Type of word variations met in the Bentham Dataset for the words 'England' and 'Embezzlement'

$$P@k = \frac{|\{relevant\ words\} \cap \{k\ retrieved\ words\}|}{|\{k\ retrieved\ words\}|} \quad (8)$$

Precision is the fraction of retrieved words that are relevant to the search, while in the case that precision should be determined for the k top retrieved words, P@k is computed. In particular, in the proposed evaluation, P@5 is used which is the precision at top 5 retrieved words. This metric defines how successfully the algorithms produce relevant results to the first 5 positions of the ranking list.

The second metric used in the proposed evaluation is the Mean Average Precision (MAP) which is a typical measure for the performance of information retrieval systems [25], [26]. It is implemented from the Text Retrieval Conference (TREC) community by the National Institute of Standards and Technology (NIST). The above metric is defined as the average of the precision value obtained after each relevant word is retrieved:

$$AP = \frac{\sum_{k=1}^{n} (P@k \times rel(k))}{\{relevant\ words\}} \quad (9)$$



(a) (b)

Fig. 8. Representative document images from (a) Bentham Dataset, (b) Washington Dataset

where:

$$rel(k) = \begin{cases} 1, & if\ word\ at\ rank\ k\ is\ relevant \\ 0, & if\ word\ at\ rank\ k\ is\ not\ relevant \end{cases} \quad (10)$$

At this point, it is worth to note that in our experimental work it is assumed that there is an outcome of a word image segmentation method. As the scope of the proposed word is in the local features and its accompanied matching process, there will be no discussion about any specific methodology used for the segmentation process. In particular, for the experiments, the word image segmentation information is taken from the ground truth corpora.

Initially the proposed segmentation-based word-spotting was evaluated against two previous segmentation-based

TABLE I. OVERALL PERFORMANCE EVALUATION RESULTS

Datasets	Methods	P@5	MAP
Washington [10]	WS [1]	0.436	0.440
	CSPD [11]	0.631	0.608
	SIFT [19]	0.600	0.577
	Proposed Method	0.660	0.637
Bentham [24]	WS [1]	0.528	0.506
	CSPD [11]	0.629	0.615
	SIFT [19]	0.642	0.630
	Proposed Method	0.701	0.680

profile-based strategies. Then, in order to highlight the advantage of the proposed DSLF, it was replaced by the SIFT but the proposed matching algorithm remained the same. The total word image queries for the Washington dataset was 1570 and for the Bentham dataset was 3668. Both query sets contain words appearing in various frequencies and sizes. Table I shows the performance evaluation results.

The proposed method outperformed both the profile-based strategies and the SIFT local features. It is worth to note, that the profile-based features were applied in words that were binarized, denoised, de-skew and de-slant as opposed to the local features that were applied to the original word images. Moreover, although the SIFT descriptor contains more information than the proposed local features (128 values against only 27), the latter performed better in both datasets

IV. CONCLUSION

In this work, novel local features are proposed driven by the challenges presented in historical handwritten word spotting scenarios. Moreover, a matching procedure was presented based on Local Proximity Nearest Neighbour, that augments performance in terms of effectiveness and efficiency incorporating spatial context. It is proven that the proposed framework achieves better performance after a consistent evaluation against two profile-based approaches as well as the proposed approach with the popular SIFT local features in two different handwritten datasets.

ACKNOWLEDGEMENT

Work supported by the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 600707 - tranScriptorium.

REFERENCES

[1] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*. IEEE, 1996, pp. 631–637.

[2] E. Hassan, S. Chaudhury, and M. Gopal, "Word shape descriptor-based document image indexing: a new dbh-based approach," *IJDAR*, pp. 1–20, 2012.

[3] K. Zagoris, K. Ergina, and N. Papamarkos, "A document image retrieval system," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 6, pp. 872 – 879, 2010.

[4] A. L. Kesidis, E. Galiotou, B. Gatos, and I. Pratikakis, "A word spotting framework for historical machine-printed documents," *IJDAR*, vol. 14, no. 2, pp. 131–144, 2011.

[5] T. Konidaris, B. Gatos, S. Perantonis, and A. Kesidis, "Keyword matching in historical machine-printed documents using synthetic data, word portions and dynamic time warping," in *Document Analysis Systems, 2008. DAS '08. The Eighth IAPR International Workshop on*, 2008, pp. 539–545.

[6] F. Zirari, A. Ennaji, S. Nicolas, and D. Mammass, "A methodology to spot words in historical arabic documents," in *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, 2013, pp. 1–4.

[7] K. Khurshid, C. Faure, and N. Vincent, "Word spotting in historical printed documents using shape and sequence comparisons," *Pattern Recognition*, vol. 45, no. 7, pp. 2598–2609, 2012.

[8] T. M. Rath and R. Manmatha, "Features for word spotting in historical manuscripts," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 2003, pp. 218–222.

[9] ——, "Word image matching using dynamic time warping," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–521.

[10] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic word recognition for handwritten historical documents," in *Document Image Analysis for Libraries, 2004. Proc. 1st International Workshop on*, pp. 278–287.

[11] K. Zagoris, K. Ergina, and N. Papamarkos, "Image retrieval systems based on compact shape descriptor and relevance feedback information," *Journal of Visual Communication and Image Representation*, vol. 22, no. 5, pp. 378 – 390, 2011.

[12] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, vol. 17. IEEE, 1978, pp. 761–766.

[13] J. A. Rodrıguez and F. Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," in *Int. Conf. on Frontiers in Handwriting Recognition*, 2008.

[14] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, "Handwritten arabic word spotting using the cedarabic document analysis system," in *Proc. Symposium on Document Image Understanding Technology (SDIUT-05)*, 2005, pp. 123–132.

[15] R. Shekhar and C. Jawahar, "Word image retrieval using bag of visual words," in *DAS 2012*, March 2012, pp. 297–301.

[16] J. Llados, M. Rusinol, A. Fornes, D. Fernandez, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *IJPRAI*, vol. 26, no. 05, 2012.

[17] M. Rusinol, D. Aldavert, R. Toledo, and J. Lladós, "Browsing heterogeneous document collections by a segmentation-free word spotting method," in *ICDAR 2011*. IEEE, 2011, pp. 63–67.

[18] K. Zagoris, I. Pratikakis, A. Antonacopoulos, B. Gatos, and N. Papamarkos, "Distinction between handwritten and machine-printed text based on the bag of visual words model," *Pattern Recognition*, vol. 47, no. 3, pp. 1051 – 1062, 2014.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[20] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009.

[21] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[22] H. Fujisawa and C.-L. Liu, "Directional pattern matching for character recognition revisited," *algorithms*, vol. 13, p. 14, 2010.

[23] A. L. Koerich, Y. Leydier, R. Sabourin, and C. Y. Suen, "A hybrid large vocabulary handwritten word recognition system using neural networks with hidden markov models," in *Frontiers in Handwriting Recognition, 2002. Proceedings. 8th International Workshop on*, 2002, pp. 99–104.

[24] D. G. Long *et al.*, *The manuscripts of Jeremy Bentham: a chronological index to the collection in the Library of University College, London: based on the catalogue by A. Taylor Milne*. The College, 1981.

[25] T. NIST. (2013) http://trec.nist.gov/pubs/trec16/appendices/measures.pdf.

[26] S. A. Chatzichristofis, K. Zagoris, and A. Arampatzis, "The trec files: the (ground) truth is out there," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, ser. SIGIR '11. New York, NY, USA: ACM, 2011, pp. 1289–1290.