An Adaptive Zoning Technique for Word Spotting Using Dynamic Time Warping

A. Papandreou[1], B. Gatos[1] and K. Zagoris[2]

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

[2] Department of Electrical & Computer Engineering,
Democritus University of Thrace,
67 100 Xanthi, Greece
kzagoris@ee.duth.gr

***Abstract* — Zoning features have been proved one of the most efficient statistical features which provide high speed and low complexity word matching. They are calculated by the density of pixels or pattern characteristics in several zones that the pattern frame is divided. In this paper, an adaptive zoning technique for efficient word spotting is introduced. The main idea is that the zoning features are extracted after cutting the query word in vertical zones, according to its length and pixel distribution along the horizontal axis, and adjusting these boundaries optimally with the corresponding zones in the candidate match-word using Dynamic Time Warping (DTW). This adjustment is performed by coupling every zone of the query word to the corresponding zone of each candidate match-word with the use of the corresponding warping matrix. This process absorbs the ambiguities between the query and the candidate match words and due to this fact it can be applied to both machine-printed and handwritten document images. The proposed word spotting technique is tested using the pixel density as a characteristic feature in every zone and an improvement is recorded compared to other state-of-the-art methods.**

Keywords - Zoning Features; Word Spotting; Dynamic Time Warping; Word Matching; Hanwritten Historical Documents;

I. INTRODUCTION

Word spotting applications are mainly used to locate all the occurrences of a given word image in a set of document images. Measuring the distance between two word images is an important step for all ‘‘query by example’’ word spotting applications that involve a word segmentation stage.

Existing approaches for segmentation-based word image matching incorporate statistical, structural and transformation-based features [1-4]. In [1], sets of 1-dimensional structural features are created from the segmented word images which are then compared using Dynamic Time Warping (DTW). In [2], the distance between two word images is calculated using an image dissimilarity measure based on curvature estimation using integral invariants and a windowed Hausdorff distance. Other descriptors used for word spotting are the SIFT key-points combined with a Bag of Visual Words (BoVW). Rodrguez and Perronnin in [5] extract features from a sliding window, based on the first gradient while Llados et. al. in [6] compare the performance of a BoVW, a pseudo-structural representation based on Loci Features and sequences of column features based on DTW. Finally, in [7] Zagoris et al. propose 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 fuzzy algorithm.

Zoning features have been proved one of the most efficient statistical features which provide high speed and low complexity word matching. They are calculated by the density of pixels or pattern characteristics in several zones that the pattern frame is divided. In particular, standard zoning methods are defined according to a $N \times M$ regular grid superimposed on the image body [8]. Zoning features based on pixel density have been combined with word profiles in a hybrid scheme for handwritten word recognition [3] as well for word spotting in historical printed documents [4]. In [9], features based on distances and angles of the skeleton pixels in each zone are used. Both neural networks and fuzzy logic techniques are then used for recognition. The methodology presented in [10] is based on the direction of the contour of the character by computing histograms of chain codes in each zone. In [11], it is observed that when the contour curve is close to zone borders, small variations in the contour curve can lead to large variations in the extracted features. For this reason, zones with fuzzy borders are introduced. Features detected near the zone borders are given fuzzy membership values to two or four zones. In [12], the role of feature membership functions in Voronoi-based zoning methods is investigated. Zoning is considered in [13] as the result of an optimization problem and a genetic algorithm is used to find the optimal zoning that minimizes the value of the cost function associated to the classification. In [14], the idea of adaptive zoning is introduced and the features are extracted after adjusting to the position of every zone based on local pattern information. This adjustment is performed by moving every zone towards the pattern body maximizing the local pixel density around each zone.

In this paper, an improved version of adapting zoning is proposed for word spotting. The novelty is mainly

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.79

introduced by the use of DTW for the adjustment of the horizontal boundaries of the zones of the two word images. According to the proposed approach, this adjustment is performed by coupling every zone of the query word to the corresponding zone of each candidate match-word with the corresponding warping matrix. This process absorbs the ambiguities between the query and the candidate handwritten word images.

In Section II, the proposed adaptive zoning technique is described. Furthermore, the proposed word spotting method is tested in handwritten historical documents. As it is shown in Section III, a significant improvement is recorded when the zoning features are used in the proposed adaptive way. Conclusions and future work plans are given in Section IV.

II. The Proposed Adaptive Zoning Technique

Binarization [15] and deslanting [16] are first applied in the query and candidate word images as pre-processing steps. In that way, the inclination of the handwriting is corrected, both words are normalized and this results in a more accurate matching between the query and the candidate word. Afterwards, points of interest are detected in the query word image in order to define its vertical zones. DTW is then used in order to couple these points of interest with the corresponding points of the candidate match-word. In that way, vertical zones are defined adaptively in respect with the query image swith the use of DTW. At a next step, the core-region of the two word images is computed and the horizontal zones of the images are defined in respect to that. Finally, the word images are normalized and their features are extracted. In detail:

A. DTW – Adjusting the Vertical Zones

As a first step, the vertical zones of the binary query word must be defined. In order to find the number and the width of these zones, the number $BW()$ of horizontal black to white transitions are calculated for each line of the image. For the binarized pattern image $\mathrm{I}(\mathrm{x},\mathrm{y}) \in \{0,1\}$ of size $I_x \times I_y$, $BW()$ is calculated as follows:

$$BW(y) = \sum_{\mathrm{i}=0}^{I_x-2} (\mathrm{I}(\mathrm{i},\mathrm{y}) - \mathrm{I}(\mathrm{i}+1,\mathrm{y}))^2 \quad (1)$$

Then, the line G that contains the maximum number of black and white transitions is the guide that defines the boundaries of the vertical zones.

$$G = \arg\max_{y}(BW(y)) \quad (2)$$

All black to white transition b, in line G of the query image, are denoted as points of interest b_n (see Figure 1).

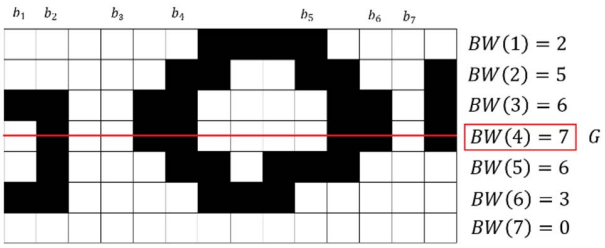


Figure 1. The $BW(y)$ calculated for each y and the b_n points of interest of the G line.

Afterwards, the vertical projection profile histogram of the query word image is calculated and its peaks (local maxima), p_n , are detected by scanning the histogram. Then, the local minima of the histogram are detected between consequent peaks and noted as a valleys v_n which, along with p_n , are also considered as points of interest as it is shown in Figure 2.

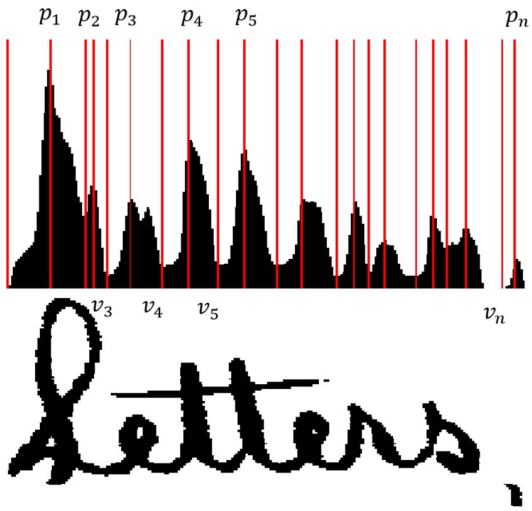


Figure 2. The vertical projection profile histogram of the query word image is calculated and the local minima of the histogram are detected between consequent peaks p_n and noted as a valley v_n .

At a next step, we consider b_n , p_n and v_n horizontal coordinates and a post processing step takes place in order to discard points of interest that are located less than 3 pixels from their previous neighbor. The remaining points compose the boundaries q_n^{x1}, q_n^{x2} of the N corresponding zones of the query image. In Figure 3, the partitioning in $N = 25$ vertical zones of a query image is presented.

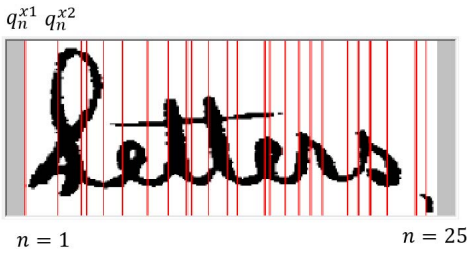


Figure 3. The 25 corresponding vertical zones of a query image.

Then, a set of features v_i (see [1]), where $i = 1 \ldots 4$, of the query word image is calculated. The features include i) the vertical projection profile, ii) the upper profile, iii) the derivative of the vertical projection profile and iv) the derivative of the upper profile. A linear scaling is applied in order to normalize the calculated features in $[0,1]$. Finally, these features are combined in a query characteristic feature vector sequence V. The same procedure is followed in order to calculate the features p_i of the candidate word image and its characteristic feature vector sequence P. Once both feature sequences V and P are calculated, the DTW algorithm is applied in order to compute the minimum cost and the corresponding alignment path.

The key-role of the DTW algorithm is to measure the similarity between two sequences which may have variable lengths. After applying the DTW algorithm, the feature vectors are aligned along a common, warped time axis. The cost of the alignment is given by the sum of distances $d(V,P)$ of each aligned vector pair along the corresponding alignment path. The distance similarity measure $d(V,P)$ employed is the squared Euclidean distance:

$$d(V,P) = \sum_{i=1}^{4}(v_i - p_i)^2 \quad (3)$$

The DTW cost and the corresponding warping matrix between a query word and a candidate word is defined as the minimum alignment cost which is calculated using the principles of dynamic programming [1]. This cost is normalized with respect to the length of the previously calculated warping path.

Consequently, according to the estimated warping path, the coordinates of every vertical zone q_n^{x1}, q_n^{x2} belonging to the query word image are matched with the coordinates of the candidate word image c_n^{x1}, c_n^{x2} (see Fig.4).

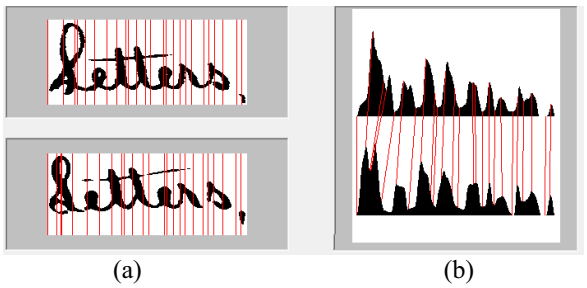


Figure 4. (a) The coordinates of the vertical zones of the query are matched with the corresponding coordinates of the candidate match word image. Their vertical vertical Histograms mapping in (b).

The adjustment of the vertical zones between a query word and a candidate word, results in vertical zones of variable width.

B. Size Normalization and Horizontal Zone Adjustment

For the horizontal adjustment of the query and the candidate word images, a normalization step is applied resulting respectively in images $Q[W \times H]$ and $C[W \times H]$ of width W and height H. The positioning of each word in its size-normalized image $W \times H$ is accomplished by placing the baseline areas of each word in the center of the matrix i.e. the upper word baseline at vertical offset $H/3$ and the lower word baseline at vertical offset $2 \times H/3$. In order to calculate the baseline the procedure in [16] is followed. A similar approach is also used in [14]. A word size-normalization example is presented in Figure 5 for the query and a candidate word.

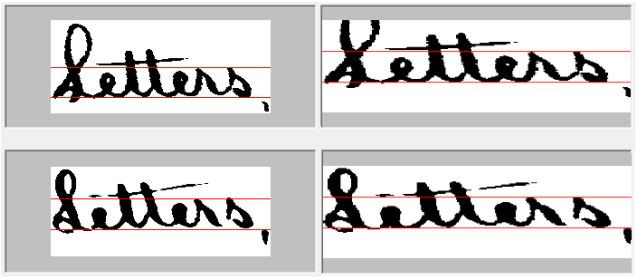


Figure 5. Word size normalization example: (a) orginal, (b) size-normalized image.

After normalizing the two words, the estimated coordinates of the vertical zones $q_n^{x1}, q_n^{x2}, cm_n^{x1}, cm_n^{x2}$ are also normalized to $q_n^{x1'}, q_n^{x2'}, cm_n^{x1'}, cm_n^{x2'}$ with the use of the corresponding transformation matrix, respectively to the new width, W, of the normalized images.

For the definition of the horizontal zones, a $1 \times M$ regular grid is superimposed on the normalized images. The coordinates of every horizontal zone are defined by the following equations:

$$q_m^{y1'} = (m-1)\Lambda \quad (4)$$

$$q_m^{y2'} = m\Lambda - 1 \quad (5)$$

where $q_m^{y1'}$ and $q_m^{y2'}$ correspond respectively to the beginning and the end of the horizontal zones, Λ is the height of the zones and $m = 1..M$. Concerning the coordinates of horizontal zones of the candidate word $c_m^{y1'}$ and $c_m^{y2'}$ it should be noted that they are equal to $q_m^{y1'}$ and $q_m^{y2'}$ since the normalization procedure results in images having the core region located in the same area. In Figure 6 the zones of the query and a candidate word for $M = 6$ are demonstrated.

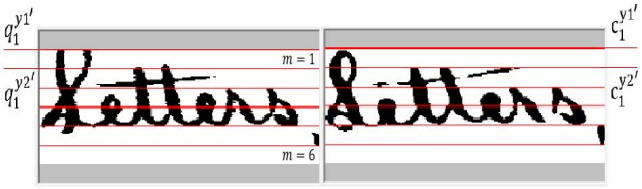
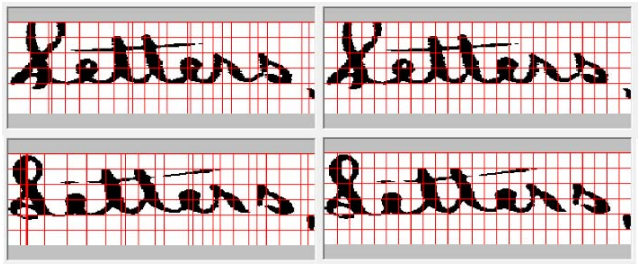


Figure 6. The zones of a query image for $M = 6$.

After the definition of both, horizontal and vertical zones a final $N \times M$ grid is defined which is superimposed on the query word as well as on the candidate word. Both query and candidate match grids are demonstrated in comparison in Figure 7.



(a) (b)

Figure 7. Comparison of the Query word and Candidate match word grids cut in (a) the proposed adaptive way and in (b) the classical way.

C. Feature Extraction

Features based on pixel density are calculated directly from the size-normalized images Q and C. The density, $\mathrm{dQ_{nm}}$, of the (n,m) window of Q is calculated as follows:

$$\mathrm{dQ_{nm}} = \frac{1}{(\mathrm{q_n^{x2'}} - \mathrm{q_n^{x1'}}) \times \Lambda} \sum_{\mathrm{x=q_n^{x1'}}}^{\mathrm{q_n^{x2'}}} \sum_{\mathrm{y=q_m^{y1'}}}^{\mathrm{q_m^{y2'}}} \mathrm{Q(x,y)} \tag{6}$$

Respectively, the density, $\mathrm{dC_{nm}}$, of the (n,m) window of C is computed as follows:

$$dC_{nm} = \frac{1}{(c_n^{x2'} - c_n^{x1'}) \times \Lambda} \sum_{x=c_n^{x1'}}^{c_n^{x2'}} \sum_{y=c_m^{y1'}}^{c_m^{y2'}} C(x,y) \tag{7}$$

The total number of features based on pixel density is $\mathrm{N \times M}$ and all features range between 0 and 1.

Additionally, due to the fact that the vertical zones of the query and candidate word images have variable width, the density of each window (n,m) will be weighted to their respective width. The following equation defines the final density:

$$dQ'_{nm} = dQ_{nm} \times \frac{(q_n^{x2'} - q_n^{x1'})}{W} \tag{8}$$

$$dC'_{nm} = dC_{nm} \times \frac{(c_n^{x2'} - c_n^{x1'})}{W} \tag{9}$$

The distance $Dist$, of the query and the candidate word image is defined as:

$$\mathrm{Dist} = \frac{1}{\mathrm{N \times M}} \sum_{\mathrm{x=1}}^{\mathrm{N}} \sum_{\mathrm{y=1}}^{\mathrm{M}} (\mathrm{dQ'_{nm}} - \mathrm{dC'_{nm}})^2 \tag{10}$$

Finally, the distance Dist of the two words is multiplied by the distance $\mathrm{d}(V,P)$ provided by the DTW algorithm, which is also an important similarity measure. In more detail, the final distance D of the proposed adaptive method is:

$$\mathrm{D} = \mathrm{Dist} \times \mathrm{d}(V,P) \tag{11}$$

III. Experimental Results

In this Section, we present the experimental results of the proposed adaptive zoning method in comparison with several state-of-the-art word matching algorithms. In order to test the proposed algorithm, two different historical handwritten corpuses were used. The first one was the Washington database [17] and the second one the manuscripts of Bentham [18] from tranScriptorium project [18]. The Washington dataset is written solely by a single writer, while the Bentham manuscripts are characterized by a more widespread ambiguity in the writing style.

The evaluation metrics used to measure the performance of the proposed segmentation-based algorithm are the Precision at the k Top Retrieved Words ($P@k$) and the Mean Average Precision (MAP). 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 defined as follows:

$$P@k = \frac{|\{relevant\ words\} \cap \{k\ retrieved\ words\}|}{|\{k\ retrieved\ words\}|} \tag{12}$$

In particular, in this evaluation, P@5 is used which is the precision at top 5 retrieved words. This metric defines how successfully the algorithms retrieves relevant results to the first 5 places of the ranking list.

The second evaluation metric used is the Mean Average Precision (MAP) which is a typical measure for the performance of information retrieval systems. 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 and is given by the equations that follow:

$$MAP = \sum_{q=1}^{Q} AP_q \tag{13}$$

where

$$AP = \frac{\sum_{k=1}^{n}(P@K \times rel(k))}{\{relevant\ words\}} \quad (14)$$

and

$$rel(k) = \begin{cases} 1, & if\ word\ at\ rank\ is\ k\ relevant \\ 0, & if\ word\ at\ rank\ is\ not\ relevant \end{cases} \quad (15)$$

It should also be mentioned that the proposed method is a segmentation based method and the segmentation outcome influences the performance of the algorithm. Though the scope of the proposed word is not the segmentation of handwritten documents and as a result the word image segmentation information is taken from the ground truth of the corpuses.

The total word image queries selected from the 20 document pages of the Washington dataset were 1570 while 1370 queries were selected from the 50 document images of the Bentham dataset. Both query sets contain words appearing in various frequencies and sizes. Some examples of the queries selected from the Washington and the Bentham dataset are presented in Figure 8 and Figure 9 respectively. Both datasets were binarized while the only preprocessing method applied to the word images was the slant correction described in [16].

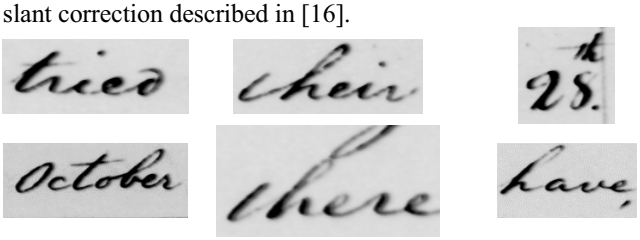


Figure 8. Samples of the query words selected from the Washington database.

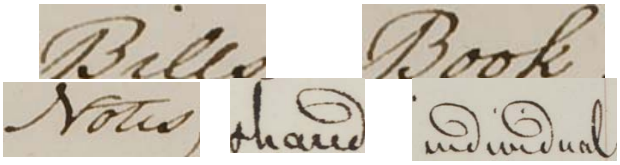


Figure 9. Samples of the query words selected from the Bentham manuscripts database.

In the following Tables, the word spotting performance achieved by the proposed adaptive method is presented in comparison with the performance of (a) Mamantha et al. as described in [20] and (b) the Zagoris et al. described in [7]. The performance of the proposed technique in the Washington and the Bentham databases are presented in Tables I and II respectively.

TABLE I. WASHINGTON DATABASE

Different Methodologies	*Evaluation Metrics*	
	***P*@5 (%)**	***MAP* (%)**
Mamantha et al. [19]	43.6	44.0
Zagoris et al. [7]	63.1	60.8
Proposed Method	**65.2**	**62.3**

TABLE II. BENTHAM MANUSCRIPTS DATABASE

Different Methodologies	*Evaluation Metrics*	
	***P*@5 (%)**	***MAP* (%)**
Mamantha et al. [19]	48,9	54,4
Zagoris et al. [7]	86,6	88,2
Proposed Method	**89,2**	**90,0**

As it can be derived from Tables I and II, the proposed method outperforms the state-of the-art algorithms. It achieved better results in both datasets of handwritten degraded historical documents. Moreover, it proved to be descriptive and yet flexible in handling the ambiguities among different instances due to handwriting style variations.

In order to perform a test in an even larger dataset another experiment took place. The 50 document images of the Bentham manuscripts were extended in the total 433 documents that are currently transcribed, 2 instances of 5 different queries (10 queries in total) were selected and the same three algorithms were evaluated in the queries word spotting. The results of this experiment are shown in Table III.

TABLE III. EXTENDED EXPERIMENT

Different Methodologies	*Evaluation Metrics*	
	***P*@5 (%)**	***MAP* (%)**
Mamantha et al. [19]	21.3	9.9
Zagoris et al. [14]	55.3	19.6
Proposed Method	**67.3**	**24.3**

The proposed algorithm proves to be way more efficient from [7] which is one of the latest descriptors for word images and has a $P@5$ above 67% in a corpus of hundreds of pages.

Another interesting finding is that due to the coupling of the DTW the proposed algorithm would be able to retrieve also words of the same root or different endings when the appropriate weights would be applied in the vertical zones that correspond to the endings of the query and the candidate words. If the windows of the right-most vertical zone contributed the least in the calculation of the distance between the query and the candidate word images, then "Letters," could be retrieved from "Letter". A visual, qualitative, estimation of this potential is demonstrated in Figure 10.

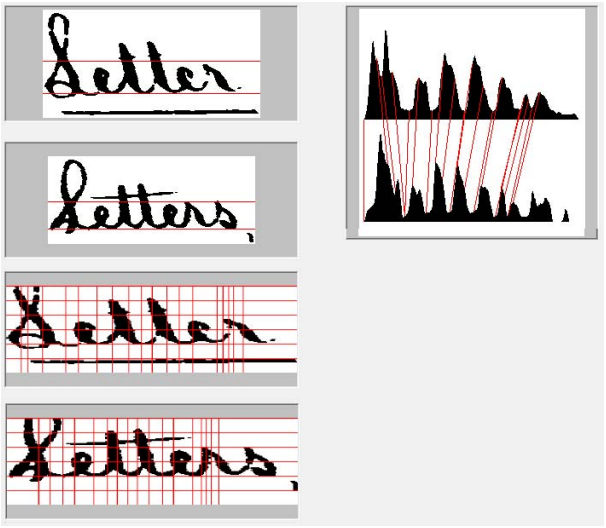


Figure 10. Qualitative estimation of potential use of the proposed algorithm to retrieve words of the same root.

IV. Conclusions and Future Work

In this paper, a novel and efficient adaptive zoning method was presented. The idea of adjusting the horizontal boundaries of the zones with the use of DTW was introduced and this was proved to result in significantly higher word retrieval performance. The proposed adaptive zoning method was tested in the difficult task of word spotting in historical degraded handwritten documents and achieved better results than the state-of-the-art. This demonstrates the fact that it is a descriptive method with the advantage of handling the ambiguities among different instances of the same word. In that way, it performs better than the classical zoning techniques that are unable to confront with handwritten word images since they don't absorb adequately the ambiguities but also DTW which is flexible in the variations but doesn't have the descriptive power of a zoning technique. As it is derived from Tables I and II it also outperforms the latest descriptors in the state-of-the-art (Zagoris [7]).

Concerning future work, it would be interesting to have a quantitative measurement on the success of this adaptive zoning method to retrieve words of the same conceptual root but in different gender or with different endings. For that purpose a different evaluation metric should be introduced.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600707 - tranScriptorium.

References

[1] T.M. Rath and R. Manmatha, "Word image matching using dynamic time warping", Proc. IEEE Computer Society Conference on Vol. 2, pp. 521-527, 2003.

[2] S. Colutto and B. Gatos, "Efficient Word Recognition Using A Pixel-Based Dissimilarity Measure", Proc. 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 1110-1114, 2011.

[3] B. Gatos, I. Pratikakis and S.J. Perantonis, "Hybrid Off-Line Cursive Handwritten Word Recognition", Proc. 18th International Conference on Pattern Recognition (ICPR), pp. 998-1001, 2006.

[4] 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", International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 167-177, 2007.

[5] 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.

[6] J. Llados, M. Rusinol, A. Fornes, D. Fernandez, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," International Journal of Pattern Recognition and Artificial Intelligence, vol. 26, no. 05, 2012.

[7] 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.

[8] M. Bokser , "Omnidocument Technologies", Proc. IEEE, Vol. 80, pp. 1066-1078, 1992.

[9] M. Hanmandlua, K.R. Murali Mohanb, S. Chakrabortyc, S. Goyald and D. Roy Choudhurye, "Unconstrained handwritten character recognition based on fuzzy logic", Pattern Recognition, Vol. 36, pp. 603–623, 2003.

[10] K. M. Mohiuddin and J. Mao, "A Comprehensive Study of Different Classifiers for Hand-printed Character Recognition", Pattern Recognition, Practice IV, pp. 437- 448, 1994.

[11] V.L. Lajish, "Handwritten Character Recognition using Perceptual Fuzzy-Zoning and Class Modular Neural Networks", Proc. 4th International Conference on Innovations in Information Technology, Innovations '07, DOI: 10.1109/IIT.2007.4430497, 2007.

[12] S. Impedovo, A. Ferrante, R. Modugno and G. Pirlo, "Feature Membership Functions in Voronoi-Based Zoning", AI*IA 2009: Emergent Perspectives in Artificial Intelligence Lecture Notes in Computer Science Volume 5883, pp 202-211, 2009.

[13] S. Impedovo, M. G. Lucchese, G. Pirlo, "Optimal zoning design by genetic algorithms", IEEE Transactions on Systems, Man, and Cybernetics, Part A 36(5), pp. 833-846, 2006.

[14] B. Gatos, A.L. Kesidis and A. Papandreou,"Adaptive Zoning Features for Character and Word Recognition", Proc. 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 1160-1164, 2011.

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

[16] A. Papandreou and B. Gatos, "Slant estimation and core-region detection for handwritten Latin words", Pattern Recognition Letters, Available online 29 August 2012, ISSN 0167-8655, 10.1016/j.patrec. 2012.08.005.

[17] 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.

[18] 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.

[19] http://transcriptorium.eu/

[20] 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, pages 631–637.IEEE, 1996.