2011 International Conference on Document Analysis and Recognition

Enhancing Handwritten Word Segmentation by Employing Local Spatial Features

Fotini Simistira
School of Electrical and Computer Engineers
National Technical University of Athens
Athens, Greece
e-mail: fotini@mail.ntua.gr

Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros
Institute for Language and Speech Processing
R.C. "Athena"
Maroussi, Greece
e-mail: {vpapa, themosst, vsk}@ilsp.gr

***Abstract*—This paper proposes an enhancement of our previously presented word segmentation method (ILSP-LWseg) [1] by exploiting local spatial features. ILSP-LWseg is based on a gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components (CCs). Then a global threshold for the gap metrics is estimated and used to classify the candidate gaps in "within" or "between" words classes. In the proposed enhancement the initial categorization is examined against the local features (i.e. margin and slope of the linear classifier for every pair of CCs in each text line) and a refined classification is applied for each text line. The method was tested on the benchmarking datasets of ICDAR07, ICDAR09 and ICFHR10 handwriting segmentation contests and performs better than the winning algorithm.**

Keywords: handwritten word segmentation; document image processing; support vector machines

I. INTRODUCTION

The segmentation of a document image into words is a critical stage in the workflow of a system for retrieving unconstrained handwritten documents. If a document is segmented into words then further tasks such as word recognition and character segmentation and recognition may be developed. Therefore, the efficiency of document image analysis methods is often affected by the precision of the word segmentation process. Even though word segmentation could be considered a solved problem in machine-printed documents, the same task in handwritten documents remains an open issue. The main reason is that the format of a handwritten manuscript and the writing style depend solely on the author's choices. Due to high variability of writing styles and scripts, methods that adapt to the properties of the document image, would be more robust.

The main assumptions that most word segmentation approaches adopt are that: i) the document is already segmented into text lines, ii) each CC belongs to only one word and iii) gaps between words are greater than gaps between consecutive segments belonging to the same word. These techniques consider a spatial measure for the gap between consecutive CCs and employ a proper threshold to classify the gaps as "within" or "between" words. The ILSP-LWseg algorithm utilizes a gap measure which results from the optimal value of the objective function of a soft-margin linear SVM that separates successive CCs [1].

The adoption of this metric allows us to get a tolerant measure. It is known that the SVM classifier's separation plane is located properly in order to maximize the margin between two classes. Considering the text pixels on either side of the gap under consideration instances of the two classes, the classifier is adjusted properly in order to "achieve" the maximum distance between the successive CCs (see fig. 1). Therefore, the spatial measure is adjusted to the local topology of the text and is equivalent (see fig. 2) to the Bounding Box Distance (BBD), or the Minimum Euclidean Distance (MED) or the Convex Hull Distance (CHD). In addition the "soft nature" of the SVM classifier allows penetration of pixels from either side into the margin zone and therefore results to a gap metric similar to the Minimum Run-Length Distance (MRLD).

Another critical issue is the slope of the linear SVM classifier which provides information about the writing style. For example, an almost vertical separator denotes non-cursive handwriting (fig. 1a), while significant slope implies cursive handwriting (fig. 1b).

The organization of the rest of the paper is as follows: In Section II, we refer to recent related work. In Section III, we describe in detail the proposed algorithm. Evaluation results and conclusions are discussed in Sections IV and V, respectively.

II. RELATED WORK

This section surveys recent work in word segmentation of handwritten document images. The subsequent techniques either achieved remarkable results in the corresponding test datasets, or are incorporated in the workflows of integrated systems for specific tasks. As mentioned, the majority of word segmentation algorithms consider that the documents to be processed are segmented firstly into text lines properly.

Marti and Bunke [2] employ the CHD (fig. 2c) to estimate the gap metric between successive CCs. Based on the horizontal distance between the leftmost and rightmost black pixel in each text line and the median stroke width, a threshold for each text line is calculated. Then the threshold is used to classify the candidate gaps to "inter" or "intra" words. The algorithm tested on 541 text lines containing 3899 words of IAM [3] and performed a correct segmentation rate of 95.56%.

Seni and Cohen [4] propose a similar method which combines the MRLD (fig.2 d) with the vertical overlapping

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.264

of two successive CCs. Based on two predefined thresholds (the first for MRLD values and the second for vertical overlapping) the gaps are categorized into "within" and "between" words. Then the results are enhanced by utilizing the results of a punctuation mark detection algorithm. The method tested on nearly 3000 handwritten text lines and performed an error rate of about 10%.

Manmatha and Rothfeder [5] introduce a scale space approach based on filtering the document image by an anisotropic Laplacian filter at different scales. The produced blobs correspond to portions of characters at small scales and to words at larger scales. It was shown experimentally that the optimum scale is equal to the 10% of the text-line height. The method applied on a sample of 100 manuscripts of George Washington and a total error rate of 17%.

Lemaitre et al. [6] propose a segmentation of text lines into words based on the cooperation among digital data and symbolic knowledge. The digital data are obtained from distances inside a Delaunay graph at the pixel level. Structural knowledge is taken into account in order to group small isolated CCs with words. The method tested on the ICDAR09 Handwriting Segmentation Contest test set and achieved a correct segmentation rate of 94.20%.

Geraud [7] suggests a fast technique based on mathematical morphology. The main steps include morphological closing, application of the distance transform, area closing and application of the watershed transform. The method participated in the above-mentioned contest and performed a correct segmentation rate of 83.92%. Even though the performance is not high enough, it is worth mentioning that this technique is the only one which does not require text-line segmentation as a pre-processing step.

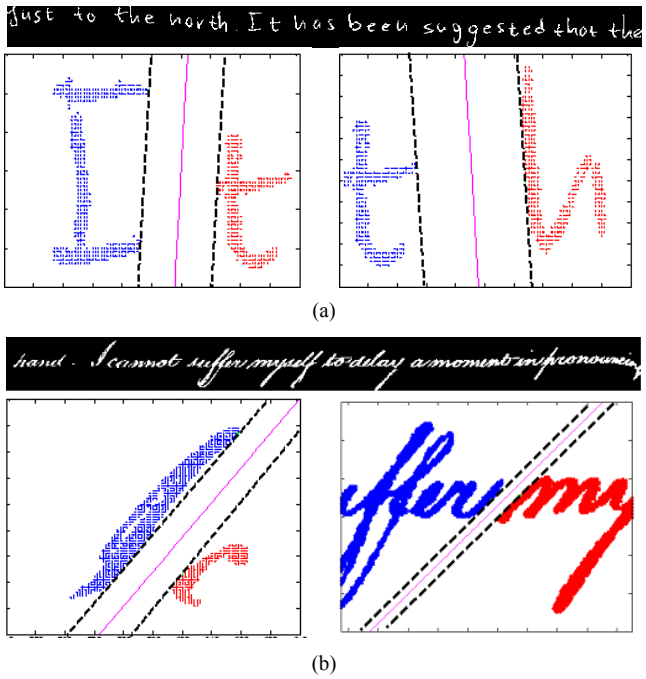


Figure 1. Examples of estimation of the gap metrics for candidate word separators in the:(a) 10th text line of image 022.tif, (b) 2nd text line of image 009.tif from ICFHR2010 Handwriting segmentation Contest.

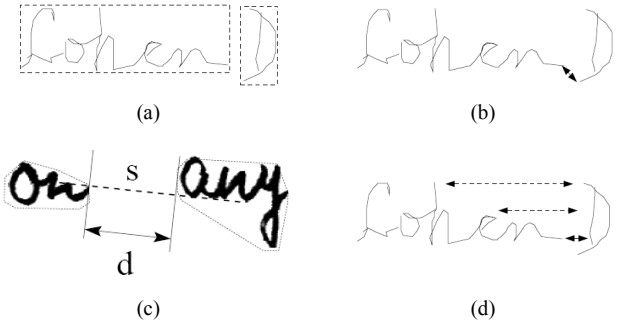


Figure 2. Four metrics used in word segmentation: (a) Bounding Box Distance, (b) Minimum Euclidean Distance, (c) Convex Hull Distance, and (d) Run-Length Distance.

III. PROPOSED METHOD

In the ILSP-LWseg method, the gap metric for every pair of successive CCs in the whole document is calculated. Let g_k^{ℓ} be the gap metric between the *k*-th and the *(k+1)*-th CCs of the ℓ -th text line. We introduce the variables $x_m \in X_k \subseteq \mathbb{R}^2$ that correspond to the 2-d coordinates of the *m*-th foreground pixel and $y_m \in Y_k = \{-1, 1\}$ to denote the CC that the *m*-th pixel belongs to (i.e. –1 refers to pixels of the left CC while 1 refers to pixels of the right CC). The primary objective function for the soft margin SVM for the dataset $Z_k = (X_k, Y_k)$ is given by

$$L(\mathbf{w}, b, a, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{|Z_k|}\xi_i$$

$$-\sum_{i=1}^{|Z_k|} a_i\left\{y_i\left[(x_i \cdot \mathbf{w}) + b\right] - 1 + \xi_i\right\} - \sum_{i=1}^{|Z_k|}\mu_i\xi_i \qquad (1)$$

where $(\mathbf{w}, b)$ define the hyperplane, ξ_i are the slack variables, a_i and μ_i are the Lagrange multipliers, C is a non-negative constant used to penalize classification errors, x_i are the feature space data points (i.e. the 2-d coordinates of the foreground pixels) and $|Z_k|$ is the cardinality of the dataset.

The optimal classifier for the two CCs results from the minimization of L, i.e. the lowest value of L corresponds to the smallest $\|\mathbf{w}\|$ and consequently to the largest margin. Therefore, we define the gap metric between the *k*-th and the *(k+1)*-th CCs of the ℓ -th text line as

$$g_k^{\ell} = -\log\left\{\min_{0<a_i\le C}(L)\right\} \qquad (2)$$

It is worth mentioning that the transformation in the log domain is introduced to enhance small size differences in

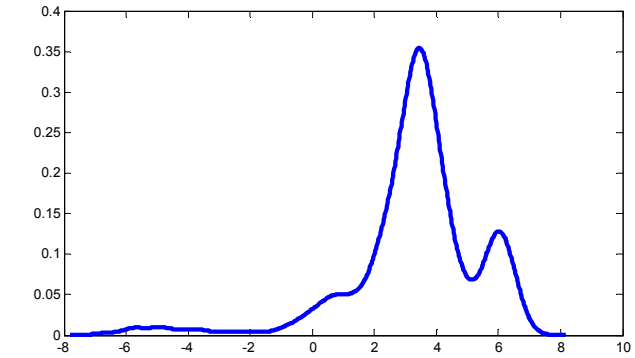
L and the minus sign so that the gap metric increases with respect to the margin.

Then, a nonparametric approach [11] is employed to estimate the probability density function of the gap metrics (fig. 3a) using the following formulae

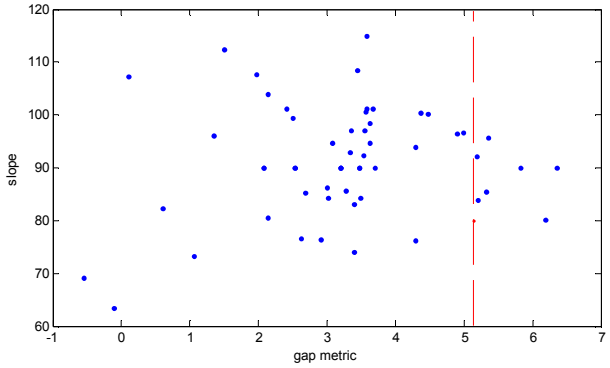
$$p(x) = \frac{1}{Mh}\sum_{t=1}^{M} K\left(\frac{x - x^t}{h}\right) \quad (3)$$

where M is the number of all gap metrics within a document page and $K(\cdot)$ denotes the normal kernel.

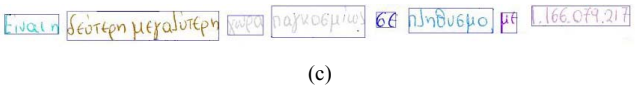
Since high (low) values correspond to gaps between (within) words, two main lobes can be identified. Therefore, a proper threshold for classification is the value that is equal to the minimum between the two lobes.



(a)



(b)



(c)

Figure 3. An example of word segmentation based on global threshold in document image 354.tif from ICFHR10: (a) the estimated pdf (vertical axis) of gap metrics (horizontal axis), from which the global threshold is found equal to 5.169, (b) the scatter diagramm of candidate gaps in the second text line of the image (the red line stands for the global threshold), (c) the resulting segmentation of the text line.

However, in case that the two lobes could not be distinguished clearly (i.e. the valley is not deep enough) the gaps around the threshold might be misclassified. Fig. 3b and 3c illustrate such a case. For example, the candidate gaps of the second text line in image 354.tif from ICFHR10 dataset would be classified as shown in the scatter diagram (fig. 3b) by adopting the global threshold. Even though the initial classification (ILSP-LWseg method) is based on the gap metrics only, we plot the slope values in order to show that the distribution of the slope values for the "real" gaps (i.e. the "between" words gaps) is low. One could observe (fig. 3b) that only seven gaps between successive CCs have been classified as inter-word gaps. In addition, two more gap metrics lie near the global threshold. The resulting segmentation is presented in fig 3c. Since two pairs of words (e.g. "Είναι" and "η", and "δεύτερη" and "μεγαλύτερη") have been merged, a refined categorization is required.

In order to reclassify the candidate gaps, which lie in the proximity of the global threshold, we introduce a post-processing stage that aims to deal with such instances and enhance further the ILSP-LWseg method. The proposed enhancement consists of the following steps which are applied on each text line:

i) Based on the global threshold we classify the gaps as "candidate between" and "candidate within" words, denoted as CB and CW respectively.

ii) Given the gap metrics, the slope values and the in initial labels, we calculate the mean values $\mathbf{m}_{CB}$ and $\mathbf{m}_{CW}$, and the covariance matrices $\mathbf{\Sigma}_{CB}$ and $\mathbf{\Sigma}_{CW}$ for each initial candidate class.

iii) The candidate gaps that should be re-examined, lie near the global threshold and their slopes are "similar" to the slopes of the gaps that have been labeled as CB in step (i). Therefore, we define an area of "ambiguity" which includes such gaps (fig. 4a). The borders (GL, SL, GH, SH) of this area are calculated as follows:

$$GL = m_{CB}^{G} - 3\cdot\sigma_{CB}^{G} \quad (4)$$

$$SL = m_{CB}^{S} - 3\cdot\sigma_{CB}^{S} \quad (5)$$

$$GH = m_{CB}^{G} \quad (6)$$

$$SH = m_{CB}^{S} + 3\cdot\sigma_{CB}^{S} \quad (7)$$

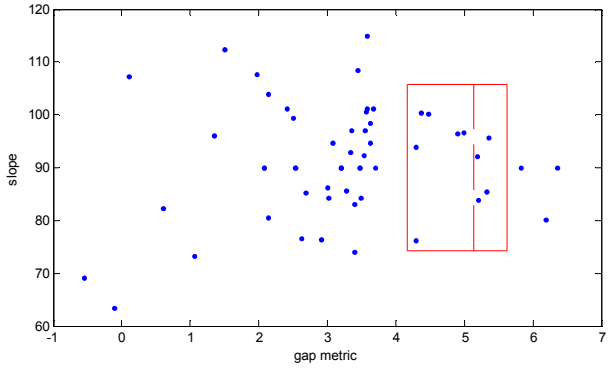
where m_{CB}^{G}, m_{CB}^{S}, σ_{CB}^{G} and σ_{CB}^{S} are the mean values and the standard deviations of the gap metrics and the slope values in the CB class, respectively. Assuming that the distributions of the slopes and the gap metrics of CB gaps are approximately normal, equations 4-7 imply that the "between word" gaps are within three standard deviations. The only difference is that

the right boundary of the area is set to the mean value (Eq. 6), since greater values correspond to "between word" gaps and there is no need to re-examine them.

iv) We calculate the corresponding probabilities for each gap $\mathbf{x}$ in the area of "ambiguity" as follows:

$$p(\mathbf{x};\mathbf{m}_c,\boldsymbol{\Sigma}_c) \sim N(\mathbf{m}_c,\boldsymbol{\Sigma}_c), \quad c = CB, CW \tag{8}$$

v) By comparing the probabilities for each $\mathbf{x}$, the gap is classified as the "between" and "within" words properly. Finally, the gaps, which have been classified as CB or CW and are not in the area of "ambiguity", are considered "between" and "within" word gaps respectively.



(a)

(b)

Figure 4. An example of the refined classification: (a) the scatter diagramm of the candidate gaps in the second text line (the red rectangle denotes the area of "ambiguity"), (b) the resulting segmentation.

From fig. 3b one could observe that seven candidate gaps have been classified as "between" words during the initial phase of the process. Actually, this would be the result of the ILSP-LWseg method. By following the proposed procedure, two more candidate gaps (the gaps represented by the two nearest to the threshold points lying on the left side of the threshold) have been classified as "between" words and the resulting segmentation is illustrated in fig. 4b.

IV. EVALUATION RESULTS

In order to test the enhancement of the algorithm we used the datasets from three Handwriting Segmentation Contests: ICDAR07 [8], ICDAR09 [9] and ICFHR10 [10]. The test datasets of the contests consist of 80 (13311 words), 200 (29717 words) and 100 (15130 words) binary handwritten document images respectively. Further details for the datasets and the organization of the contests are included in the related papers. The performance evaluation is based on the detection rate (*DR*) and recognition accuracy (*RA*). If *X* denotes the number of correctly segmented words, *T* is the number of the words in the ground truth and *M* is the number of detected words, then *DR* and *RA* are calculated as *DR=X/T* and *RA=X/M*.

The evaluation results of the three contests are illustrated in Tables I, II and III, where *FM* is the *F*-measure of *DR* and *RA*. The proposed technique improves the performance of the winning algorithm on all datasets. Specifically, *FM* achieved on the first and the third datasets an increase of 2.69 and 0.93 respectively. Regarding the second dataset the performance remained almost the same.

TABLE I. EVALUATION RESULTS OF ICDAR 2007 DATASET

	M	***X***	***DR (%)***	***RA (%)***	***FM (%)***
BESUS	19091	9114	68,47	47,74	56,26
DUTH-ARLSA	16220	9100	68,36	56.10	61,63
ILSP-LWSEG	**13027**	**11732**	**88,14**	**90,06**	**89,09**
PARC	14965	10246	76,97	68,47	72,47
UoA-HT	13824	11794	88,60	85,32	86,93
RLSA	13792	9566	71,87	69,36	70,59
PROJECTIONS	17820	8048	60,46	45,16	51,70
Proposed	**13142**	**12140**	**91,19**	**92,38**	**91,78**

TABLE II. EVALUATION RESULTS OF ICDAR 2009 DATASET

	M	***X***	***DR (%)***	***RA (%)***	***FM (%)***
CASIAMSTSeg	31421	25938	87,28	82,55	84,85
CMM	31197	27078	91,12	86,80	88,91
CUBS	31533	26631	89,62	84,45	86,96
ETS	30848	25720	86,55	83,38	84,93
ILSP-LWSEG	**29962**	**28279**	**95,16**	**94,38**	**94,77**
Jadavpur	27596	23710	79,79	85,92	82,74
LRDE	33006	26318	88,56	79,74	83,92
PAIS	30560	27288	91,83	89,29	90,54
Proposed	**29742**	**28183**	**94,84**	**94,76**	**94,80**

TABLE III. EVALUATION RESULTS OF ICFHR 2010 DATASET

	M	***X***	***DR (%)***	***RA (%)***	***FM (%)***
NifiSoft-a	15192	13796	91,18	90,81	91,00
NifiSoft-b	15145	13707	90,59	90,51	90,55
IRISA	14314	12911	85,33	90,20	87,70
CUBS	15012	13454	88,92	89,62	89,27
TEI	14667	13406	88,61	91,40	89,98
ILSP-a	**14796**	**13642**	**90,17**	**92,20**	**91,17**
Proposed	**15012**	**13885**	**91,77**	**92,49**	**92,13**

By observing carefully the results for each document, we conclude that in most cases the incorrectly segmented words occur in documents, which have been segmented in text lines erroneously. Obviously, the proposed enhancement cannot handle such cases. Another common mistake concerns the punctuation signs. Since the punctuation marks are associated with previous words in our ground truth, the assignment of a punctuation mark to the following word or the consideration of such a symbol as an individual word is penalized.

V. CONCLUSIONS

We have presented a technique to enhance an already existing efficient method for handwritten word segmentation. The existing method exploits the objective function of a soft-margin linear SVM to formulate the distance between two successive CCs, and calculates a threshold to classify the candidate gaps of the whole document image as "between" or "within" words. Since the global use of this threshold results to misclassifications, we introduce a method which aims to eliminate these errors. In particular, based on the initial classification we formulate a normal distribution for each class. Then we reclassify the candidate gaps, which lie around the threshold, by employing the maximum likelihood criterion. The adoption of this procedure improves the performance of the method as concluded by testing on three well-know handwriting segmentation contest datasets. We believe that a combination of the algorithm with a punctuation mark detection algorithm would achieve further improvement.

REFERENCES

[1] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, "Handwritten document image segmentation into text lines and words", Pattern Recognition, vol. 43, Jan. 2010, pp. 369-377, doi:10.1016/j.patcog.2009.05.007.

[2] U.V. Marti and H. Bunke, "Text line segmentation and word recognition in a system for general writer independent handwriting recognition", Proc. International Conference on Document Analysis and Recognition (ICDAR 01), 2001, pp. 159–163.

[3] U.V. Marti and H. Bunke, "The IAM-Database: an English sentence database for off-line handwriting recognition", International Journal on Document Analysis and Recognition, 2002, vol. 5, pp. 39-46, doi:10.1007/s100320200071.

[4] G. Seni and E. Cohen, "External word segmentation of off-line handwritten text lines", Pattern Recognition, vol. 27, 1994, pp. 41–52, doi:10.1016/0031-3203(94)90016-7.

[5] R. Manmatha and J.L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, 2005, pp. 1212–1225, doi:10.1109/TPAMI.2005.150.

[6] A. Lemaitre, J. Camillerapp, and B. Coüasnon "A perceptive method for handwritten text segmentation", Document Recognition and Retrieval XVIII, 2011.

[7] http://www.lrde.epita.fr/cgi-bin/twiki/view/Olena/ModuleIcdar

[8] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos, "ICDAR2007 handwriting segmentation contest", Proc. of International Conference on Document Analysis and Recognition (ICDAR 07), 2007, pp. 1284–1288.

[9] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR2009 Handwriting Segmentation Contest", Proc. 10th International Conference on Document Analysis and Recognition (ICDAR'09), July 2009, pp. 1393-1397, doi: 10.1109/ICDAR.2009.245.

[10] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICFHR 2010 Handwriting Segmentation Contest", Proc. 12th International Conference on Frontiers in Handwriting Recognition (ICFHR'10), Nov. 2010, pp. 737-742, doi: 10.1109/ICFHR.2010.120.

[11] E. Alpaydin, Introduction to Machine Learning, The MIT Press, Cambridge USA, 2004.