A Novel Two Stage Evaluation Methodology for Word Segmentation Techniques

G. Louloudis[1,2], N. Stamatopoulos[1], B. Gatos[1]

[1] *Institute of Informatics and Telecommunications, Computational Intelligence Laboratory, National Center for Scientific Research "Demokritos", GR-153 10 Agia Paraskevi, Athens, Greece*
{nstam,bgat}@iit.demokritos.gr

[2] *Department of Informatics and Telecommunications, University of Athens, Greece*
http://www.di.uoa.gr
louloud@mm.di.uoa.gr

Abstract

Word segmentation is a critical stage towards word and character recognition as well as word spotting and mainly concerns two basic aspects, distance computation and gap classification. In this paper, we propose a robust evaluation methodology that treats the distance computation and the gap classification stages independently. The detection rate calculated for every distance metric corresponds to the maximum detection rate that we could have achieved if we had a perfect classifier for the gap classification stage. The proposed evaluation framework has been applied to several state-of-the-art techniques using a handwritten as well as a historical printed document set. The best combination of distance metric computation and gap classification state-of-the-art techniques is proposed.

1. Introduction

Segmentation of a text line image into words is still considered as an open problem in the document analysis research field. The reason for this is that there are several problems which may appear in a text line image. These include the appearance of skew and slant angle, the existence of punctuation marks that tends to reduce the distance of adjacent words as well as the non-uniform spacing of words. The last problem mostly appears in handwritten text lines.

A word segmentation methodology usually comprises three stages: i) preprocessing ii) distance computation and iii) gap classification. According to most existing practices, the evaluation of a given word segmentation methodology comes after the gap classification stage where the word hypotheses are generated. This evaluation schema cannot determine the performance of the distance metric and the gap classification methodology independently. Instead, it can only inform us on how well the combination of these two stages can perform.

In this paper, we propose a robust evaluation methodology that treats the distance computation and the gap classification stages independently. That is, given a number of state-of-the art distance metrics, the proposed methodology ranks them in terms of detection rate from best to worse. The detection rate calculated for every distance metric corresponds to the maximum detection rate that we could have achieved if we had a perfect classifier for the gap classification stage. The proposed evaluation framework has been applied to several state-of-the-art techniques using two different document image sets. The two sets comprise a) the test set of the ICDAR2007 handwriting segmentation competition [1] and b) a set of Greek historical typewritten documents [2].

The paper is organized as follows: in section 2, the related work is described. In section 3, the method to evaluate the two word segmentation stages is detailed. Experimental results are presented in section 4 and, finally, section 5 describes conclusions and future work.

2. Related work

Algorithms dealing with word segmentation in the literature are based primarily on analysis of geometric

relationship of adjacent components. Components are either connected components or overlapped components. An overlapped component is defined as a set of connected components whose projection profiles overlap in the vertical direction. The first stage of all methodologies described in the literature is the preprocessing stage which usually includes noise removal, skew and slant correction and calculation of either connected or overlapped components.

We can categorize the existing word segmentation evaluation methodologies into two groups. The first group contains evaluation methodologies that evaluate the overall procedure of word segmentation. That is, they only evaluate the final result and so they do not distinguish the different stages of the word segmentation procedure. Gatos et al. [1], present the results of the Handwriting Segmentation Contest that was organized in the context of ICDAR2007. The performance evaluation method is based on counting the number of matches between the text lines or words detected by the algorithms and the text line or words of the ground truth. Louloudis et al. [3] use the Euclidean distance between overlapped components as the distance metric and a threshold that is calculated making use of several characteristics on the whole document image. They report a word detection rate of 91.7%. The authors extend the previous work in [4], where they present the use of Gaussian mixture modeling for the gap classification stage and the combination of two different distance metrics for the distance computation stage. The evaluation methodology which is used in these works is the same as [1]. Kim and Govindaraju [5] present a methodology that is making use of neural networks. A similar technique is presented in [6] by Huang and Srihari. The authors report a 90.82% overall accuracy on the "Cedar Letter" documents while the method described in [5] presents an accuracy of 87.36%. Varga and Bunke [7], try to extend classical word extraction techniques by incorporating a tree structure in order to give the computed threshold flexibility. They evaluate their system at the end without separating the stages by measuring the accuracy of word extraction. A different approach is presented from Luthy et al. [8]. The problem of segmenting a text line into words is considered as a text line recognition task, adapted to the characteristics of segmentation. This methodology does not calculate distances between components so actually it merges the two stages into one stage. In the word segmentation methodologies [5, 6, 7, 8], the accuracy of the whole procedure is measured in terms of the percentage of correctly extracted words which is defined as the percentage of correctly detected words in relation to the total number of words.

The second group contains evaluation methodologies that evaluate the two stages of the word segmentation procedure independently. Seni et al. [9], present 8 different distance metrics. These include the bounding box distance, the minimum and average run-length distance, the Euclidean distance and different combinations of them which depend on several heuristics. These metrics are evaluated by measuring the number of text lines that have a correct word segmentation result. The authors also propose a gap classification technique which is based on an iterative procedure over the set of distances and the calculation of a ratio. Mahadevan et al. [10] define a different distance metric which is called convex-hull metric. The authors after comparing this metric with some of the metrics of [9] conclude that the convex hull - based metric performs better than the other distance metrics. The authors do not propose any methodology for gap classification. For evaluating a given distance metric the authors use the methodology described in [9]. Kim et al. [11], investigate the problem of word segmentation in handwritten Korean text lines. To this end, they use three well-known metrics in their experiments: the bounding box distance, the runlength/ Euclidean distance and the convex hull - based distance. For the classification of the distances, the authors consider three clustering techniques: the average linkage method, the modified Max method and the sequential clustering. The authors tried to evaluate the two stages independently. For the distance computation they use the evaluation scheme described in [9] whereas for the gap classification stage they calculate a cumulative accuracy up to the third hypothesis. They denote that the third hypothesis yields the best results. All existing methodologies that evaluate the distance computation and the gap classification methodology independently have the drawback that they count the number of text lines that have a correct word segmentation result. A text line is considered as correctly segmented only if all the distances of words are larger from the distances of characters. This assumption leads to non precise evaluation results.

3. Proposed evaluation methodology

A typical word segmentation system contains three stages: a) preprocessing, b) distance computation and c) gap classification. The starting point of a word segmenter is a text line image. The preprocessing step usually includes noise removal, dominant skew and

slant correction and computation of the components that constitute the text line image (either connected or overlapped [3, 4, 11]). At the second stage, the distances of adjacent components are computed using a gap metric. The final stage classifies the previous calculated distances as either inter-word gaps or inter-character gaps thus generating the final word hypotheses. In this section, we present the methodology to evaluate these two word segmentation stages independently. The flow chart in Fig. 1 summarizes the evaluation of the two stages.

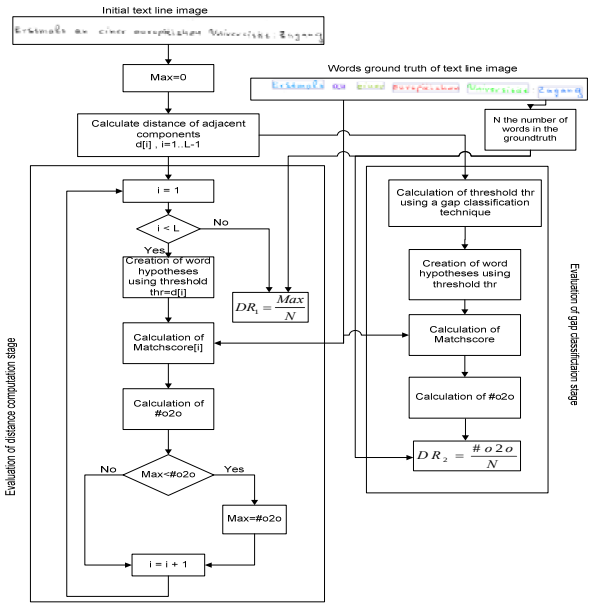

Figure 1. Flowchart summarizing the two evaluation stages.

3.1. Evaluation of distance computation stage

After the preprocessing stage, the overlapped components are calculated. For a given gap metric, we calculate the distances of adjacent overlapped components. If L is the number of the overlapped components then the total number of distances computed is L-1. We define these distances as d[i], i=1..L-1. In order to be able to evaluate the given gap metric, we assume that for every text line image involved in the procedure, the word segmentation groundtruth exists. That is, we consider that all words on all text line images are manually marked. Then, we produce L-1 possible word segmentation results, assuming for each result that the classification threshold thr equals to d[i] (Fig. 2).

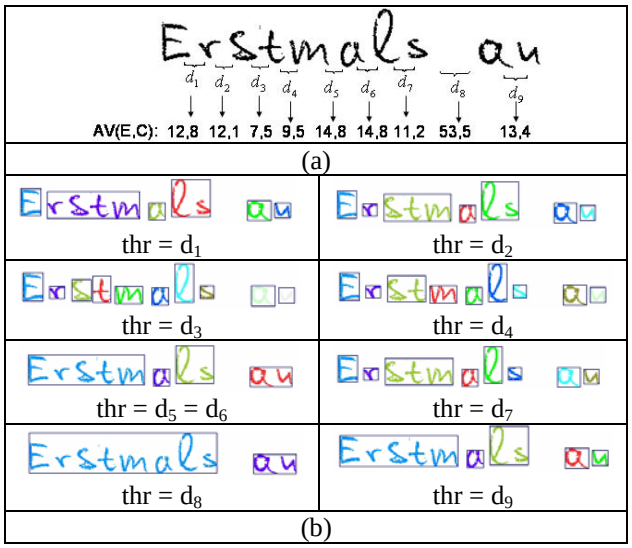

Figure 2. (a) Initial image with the calculated distances d_i. (b) Word hypotheses generated after considering each distance d_i as threshold. Threshold d_8 yields the correct result.

Let I be the set of all image points, G_i the set of all points inside the i word ground truth region, R_j the set of all points inside the j word result region, $T(s)$ a function that counts the elements of set s. Table *MatchScore(i,j)* represents the matching results of the i ground truth region and the j result region as follows:

$$MatchScore(i,j) = \frac{T(G_i \cap R_j \cap I)}{T((G_i \cup R_j) \cap I)} \quad (1)$$

The performance evaluator searches within the *MatchScore* Table for pairs of one-to-one matches. We call a pair one-to-one match (o2o) if the matching score for this pair is equal to or above the evaluator's acceptance threshold which is defined as 90% of the total word area.

Actually, assuming all possible word segmentation results, the maximum number of one-to-one matches corresponds to the number of words that could have been obtained if we had the perfect classifier for the gap classification stage. The metric for the distance computation stage is the detection rate DR_1 which is defined as

$$DR_1 = \frac{\max(\#o2o)}{N} \quad (2)$$

where max(#o2o) represents the maximum number of correct words in the text line image and N the total number of words in the text line. This metric differs from the one described in [1] in that we do not include the partial matches (one-to-many or many-to-one)

since a word segmentation result is considered correct only if the whole word is generated.
We extend this evaluator scheme to work on a set of document images. We only need to sum the maximum value of one-to-one matches for every text line. Thus, the detection rate is defined as the ratio of that number to the total number of words in the ground truth.

3.2. Evaluation of gap classification stage

After a gap metric is defined, the next step is to categorize the distances computed as either inter-word or inter-character gaps. The final result of such a procedure contains word hypotheses. The evaluation of the final stage is based on the the number of one-to-one matches as described in section 3 (Fig. 1). The metric that is used is the detection rate DR_2 which is defined as

$$DR_2 = \frac{\#o2o}{N} \quad (3)$$

It can be observed that for a given gap metric, the maximum value of detection rate that can be achieved is the detection rate value (DR_1) that the evaluation of the distance computation phase revealed. This is very important as the evaluation of the distance computation stage defines an upper threshold that we try to reach and we know that only the best classification methodology will reach.

4. Experimental results

The evaluation methodology presented in the previous section was tested on two different datasets. The first was the test set of the ICDAR 2007 Handwriting segmentation competition [1] while the second was a set of Greek historical typewritten documents [2].

The handwritten set consisted of 80 document images that contained 1773 text lines and 13315 words. The typewritten set consisted of 10 historical document images that contained 314 text lines and 3292 words.

For the distance computation stage we implemented 7 different gap metrics: the convex hull metric (Convex-Hull) [10], the Euclidean distance (Euclidean) [9], the bounding box distance (Bounding Box) [9], the average (Avg-Runlength) and minimum (Min-Runlength) runlength [9], the runlength with 2 heuristics (RLEH2) [9] and finally the average of the Euclidean distance and the convex hull distance (AV(E,C)) [4].

Concerning the gap classification stage, we tested 5 different gap classification methodologies: the sequential clustering (Sequential Clustering) [11], the modified_max (Modified_Max) [11] and the average linkage (Average_Linkage) [11] methodologies, Louloudis threshold [3] and the Gaussian mixtures methodology [4]. For all these we conducted two experiments. The first experiment defined one threshold on a whole document image whereas the second experiment (denoted as "local" in the tables) defined a threshold on a single text line image.

Table 1 summarizes the evaluator's results for the distance computation stage using the handwritten document set. From the detection rate we can see that the convex hull distance metric slightly outperformed the Euclidean distance metric. AV(E,C) metric yielded slightly better results compared to the convex hull metric. In order to evaluate the gap classification stage we used as distance computation metric AV(E,C).

Table 1. Comparative experimental results for distance computation stage on the handwritten set.

Gap Metric	N	#o2o	DR_1%
AV(E,C)	**13315**	**12983**	**97,51**
Convex-Hull	13315	12981	97,49
Euclidean	13315	12953	97,28
RLEH2	13315	12896	96,85
Bounding Box	13315	12876	96,70
Min-Runlength	13315	12538	94,16
Avg-Runlength	13315	11636	87,39

Table 2 summarizes the detection rate of the state-of-the-art gap classification techniques used. It is worth pointing out that the Gaussian Mixtures methodology achieved the maximum value. However, the detection rate achieved is far from the ideal classifier which is 97,51%.

Table 2. Comparative experimental results for gap classification stage using gap metric AV(E,C) on the handwritten set.

Classification methodology	N	#o2o	DR_2%
Gaussian mixtures	**13315**	**12381**	**92,9**
Sequential Clustering	13315	12352	92,7
Modified_max	13315	11997	90,1
Louloudis	13315	11765	88,3
Average_linkage	13315	10645	79,9
Sequential Clustering local	13315	12175	91,4
Louloudis local	13315	11979	89,9
Avearge linkage local	13315	11932	89,6
Gaussian mixtures local	13315	11835	85,8
Modified_Max local	13315	10957	82,3

Table 3 and 4 summarize the same results for the historical typewritten set. It is clear that the best gap metric was also AV(E,C). For the gap classification stage, Louloudis local threshold [3] yielded the maximum detection rate which was close to the best value that can be achieved (98,63%). It is observed that all local methodologies perform better than the global methodologies on the historical typewritten set.

Table 3. Comparative experimental results for distance computation stage on the typewritten set.

Gap Metric	N	#one-to-one	DR_1%
AV(E,C)	**3292**	**3247**	**98,63**
Euclidean	3292	3242	98,48
Convex-Hull	3292	3240	98,42
Bounding Box	3292	3240	98,42
RLEH2	3292	3235	98,26
Min-Runlength	3292	2815	85,51
Avg-Runlength	3292	2787	84,66

Table 4. Comparative experimental results for gap classification stage using gap metric AV(E,C) on the typewritten set.

Classification methodology	N	#o2o	DR_2%
Louloudis local	**3292**	**3178**	**96,5**
Gaussian mixtures local	3292	3126	95
Sequential clustering local	3292	3123	94,9
Average_linkage local	3292	3113	94,6
Modified_Max local	3292	2833	86,1
Louloudis global	3292	3121	94,8
Sequential clustering	3292	3127	95
Gaussian mixtures	3292	3067	93,2
Average_linkage	3292	3114	94,6
Modified_max	3292	2162	65,7

5. Conclusions

In this paper we propose a robust evaluation methodology that treats the distance computation and the gap classification stages independently. After thorough experimentation on two different document image sets, one handwritten and one typewritten we conclude that the best combination for distance computation metric and gap classification methodology is: i) AV(E,C) and Gaussian mixtures for the set of handwritten document images and ii) AV(E,C) and Louloudis local threshold for the set of typewritten document images.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT) and by the Greek Secretariat for Research and Development under the PENED 2003 framework.

References

[1] B. Gatos, A. Antonacopoulos, N. Stamatopoulos, "ICDAR2007 Handwriting Segmentation Contest", *9th International Conference on Document Analysis and Recognition (ICDAR'07)*, Curitiba, Brazil, September 2007.

[2] G. Vamvakas, B. Gatos, N. Stamatopoulos, S.J. Perantonis "A Complete Character Recognition methodology for Historical Documents", *8th IAPR International Workshop on Document Analysis Systems, (DAS 2008)*, pp. 525-532, Nara, Japan, September 2008.

[3] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Line and Word Segmentation of Handwritten Documents", *1st International Conference on Frontiers in Handwriting Recognition (ICFHR),* Montreal, Canada, pp. 247-252.

[4] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text Line and Word Segmentation of Handwritten Documents", to appear on *Pattern Recognition Journal, special issue on Handwriting recognition*, DOI information: 10.1016/j.patcog.2008.12.016

[5] G. Kim, V. Govindaraju, "Handwritten Phrase Recognition as Applied to Street Name Images", *Pattern Recognition*, 31(1), pp. 41-51, January, 1998.

[6] C. Huang, S. Srihari, "Word segmentation of off-line handwritten documents", *Proc. Document Recognition and Retrieval(DRR) XV, IST/SPIE Annual Symposium,* San Jose, CA, January 2008.

[7] T. Varga, H. Bunke, "Tree structure for word extraction from handwritten text lines", *8th International Conf. on Document Analysis and Recognition,* Seoul, Korea, 2005, pp. 352-356.

[8] F. Luthy, T. Varga, H. Bunke, "Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation", *Ninth International Conference on Document Analysis and Recognition,* Curitiba, Brazil, 2007, pp. 8-12.

[9] G. Seni, E. Cohen, "External Word Segmentation of Off-line Handwritten Text Lines", *Pattern Recognition,* 27(1): 41-52, 1994.

[10] U. Mahadevan, R. C. Nagabushnam, "Gap metrics for word separation in handwritten lines", *3rd International Conference on Document Analysis and Recognition,* Montreal, Canada, 1995, pp. 124-127.

[11] S.H. Kim, S. Jeong, Guee - Sang Lee, C. Y. Suen, "Word segmentation in handwritten Korean text lines based on gap clustering techniques", *6th International Conference on Document Analysis and Recognition,* Seattle, WA, USA, 2001, pp. 189-193.