Word Segmentation using the Student's-t Distribution

Georgios Louloudis, Giorgos Sfikas, Nikolaos Stamatopoulos and Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos", GR-153 10 Agia Paraskevi, Athens, Greece
{louloud, sfikas, nstam, bgat}@iit.demokritos.gr

***Abstract* — Word segmentation refers to the process of defining the word regions of a text line. It is a critical stage towards word and character recognition as well as word spotting and mainly concerns three basic stages, namely preprocessing, distance computation and gap classification. In this paper, we propose a novel word segmentation method which uses the Student's-t distribution for the gap classification stage. The main advantage of the Student's-t distribution concerns its robustness to the existence of outliers. In order to test the efficiency of the proposed method we used the four benchmarking datasets of the ICDAR/ICFHR Handwriting Segmentation Contests as well as a historical typewritten dataset of Greek polytonic text. It is observed that the use of mixtures of Student's-t distributions for word segmentation outperforms other gap classification methods in terms of Recognition Accuracy and F-Measure. Also, in terms of all examined benchmarks, the Student's-t is shown to produce a perfect segmentation result in significantly more cases than the state-of-the-art Gaussian mixture model.**

Keywords- Word Segmentation;Student's-t Distribution;Finite mixture models; Robust models.

I. INTRODUCTION

Segmentation of a text line image into words is still considered an open problem in the document analysis research field. Potential challenges include but are not limited to the appearance of skew and slant angle (even with different direction) in a text line image, the existence of punctuation marks that tends to reduce the distance of adjacent words and the non-uniform spacing of words.

A word segmentation methodology usually comprises three stages: i) preprocessing ii) distance computation and iii) gap classification. The preprocessing stage mainly includes noise removal, skew and slant correction. Distance computation concerns the selection and application of a distance measure in order to calculate the distance of adjacent components. Finally, the gap classification stage is responsible for the classification of the previously calculated distances as either between-word gaps or within-word gaps.

In this paper, we propose a novel word segmentation method which uses the Student's-t distribution for the gap classification stage. The main advantage of the Student's-t distribution concerns its robustness to the existence of outliers. In order to test the efficiency of the proposed method we used the four benchmarking datasets of the ICDAR/ICFHR Handwriting Segmentation Contests series as well as a historical typewritten dataset of Greek polytonic text. It is observed that the use of the Student's-t distribution for the gap classification step outperforms state-of-the-art word segmentation methods in terms of F-Measure. Also, we show that segmentations using the Student's-t model are significantly more likely to be 100% accurate than its Gaussian counterpart with respect to all benchmarks.

The remainder of the paper is organized as follows. In Section II, related work is presented. In Section III the proposed method is described. Section IV provides the details of the metrics used for the experiments as well as comparative experimental results. Finally, conclusions are drawn in Section V.

II. RELATED WORK

Word segmentation methods presented in the literature mainly differ either in the metric used in the distance computation stage or in the classification/clustering procedure which is considered in the final (gap classification) stage.

Several distance metrics are described in the literature. Seni et al. [1] presented eight 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. A thorough evaluation of the proposed metrics was described. A different distance metric was defined by Mahadevan et al. [2] called convex hull-based metric. The author after comparing this metric with some of the metrics of [1] concludes that the convex hull-based metric performs better than the others. Kim et al. [3], investigated the problem of word segmentation in handwritten Korean text lines. To this end, they used three well-known metrics in their experiments: the bounding box distance, the run-length/Euclidean distance and the convex hull-based distance. For the classification of the distances, the authors considered three clustering techniques: the average linkage method, the modified Max method and the sequential clustering. Their experimental results showed that the best performance was obtained by the sequential clustering technique using all three gap metrics. Varga and Bunke [4], tried to extend classical word extraction techniques by incorporating a tree structure. Since classical word segmentation techniques depend solely on a single threshold value, they tried to improve the existent theory by letting the decision about a gap to be taken not only in terms of a threshold, but also in terms of its context i.e. considering the relatives sizes of the surrounding gaps. Experiments conducted with different gap metrics as well as threshold types showed that their methodology yielded slight improvements over conventional word extraction methods.

In all the aforementioned methodologies, the gap classification threshold used derives: (i) from the processing of

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.35

the calculated distances, (ii) from the processing of the whole text line image or (iii) after the application of a clustering technique over the estimated distances. There also exist methodologies in the literature that make use of classifiers for the final decision of whether a gap is a between-word gap or a within-word gap [5 - 7].

One of the main drawbacks of several methods reported in the literature is the influence of their accuracy to the existence of aberrant, atypical or extreme values in the training set generally referred to as *outliers*. Outliers correspond to realizations of values that are related to low-probability mass areas of the true underlying data distribution. The presence of even a small number of outliers in the (finite) training dataset may lead model inference towards an inexact estimate of the true distribution. Models capable of identifying outliers in input data have been otherwise put to good use in anomaly and novelty detection as well as in one-class classification [22].

Training of the Gaussian distribution, as well as the Gaussian mixture model (GMM), is known to be sensitive to small numbers of aberrant data points. The Student's-t distribution is a heavy-tailed alternative to the Gaussian that is robust to extreme values [8, 21]. Modeling with the Student's-t has been successfully used in numerous applications including clustering, image segmentation [9], restoration [10] and sparse parameter estimation [11]. Also, training of both the Student's-t and the Student's-t mixture model (SMM) is possible using an efficient implementation of the EM algorithm [12, 13]. This leads to an iterative model parameter estimation scheme akin to the solution of the GMM.

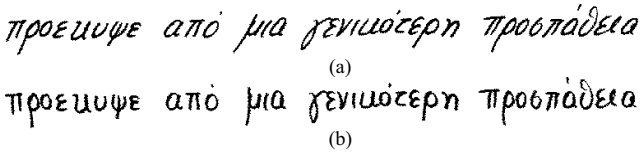
In order to overcome this problem we propose to use a 2-kernel SMM for modeling the two distributions (intra vs inter word gaps). In the current problem context, outliers correspond to intra- or inter-word gaps that are much larger or much smaller than their respective cluster mean. One extremely large value for an inter-word gap for example (i.e. a large blank area between words) will affect the variance of the respective mixture kernel, leading to a possibly inexact fit and classification. While a solution is to use a threshold to prune extreme values, its main drawback is that the ideal threshold value cannot be known a priori. We show that the proposed Student's-t mixture model can cope with outliers with no need of manual parameter fine-tuning, allowing robust estimation of the gap sizes statistics. Extensive quantitative evaluation shows that the proposed approach leads to superior results when compared with several state-of-the art methods.

III. Proposed Method

The method for the segmentation of a document image into words is an extension of the method described in [14] which is based on the use of Gaussian mixture modeling for the gap classification state. It includes three stages: (A) preprocessing, (B) distance computation and (C) gap classification. The novelty of the proposed method concerns the gap classification stage.

A. Pre-processing

Before we proceed with the word segmentation technique we apply a pre-processing procedure which concerns the removal of small connected components considered as noise as well as the correction of the dominant slant angle of the text line image [15]. Fig. 1 shows the resulting images after applying the pre-processing steps.



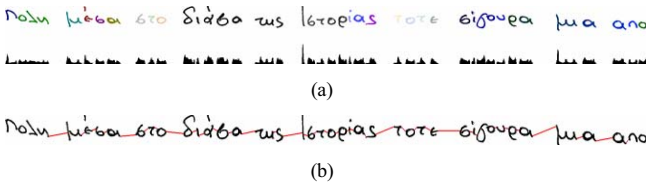
(a)

(b)

Fig. 1. Pre-processing stage: (a) original text line image, (b) after noise removal and slant correction.

B. Distance Computation

This step deals with the computation of the distances of adjacent components in the text line image [14]. The computation of the distance metric is considered not on the connected components (CCs) but on the overlapped components (OCs), where an OC is defined as a set of CCs whose projection profiles overlap in the horizontal direction. We define as distance of two adjacent OCs their Euclidean distance. The Euclidean distance between two adjacent overlapped components is defined as the minimum Euclidean distance of all pairs of points of the two adjacent overlapped components (see fig. 2).



(a)

(b)

Fig. 2. Graphical representation of (a) ovelapped components using different color per component together with the corresponding projection profiles and (b) the Euclidean distance metric (red lines) for adjacent overlapped components of a text line image.

C. Gap Classification

The gap classification stage is based on the use of a 2-kernel Student's-t model in order to describe the two underlying distributions in which the distances belong (i.e. inter word and intra word gaps). A detailed description of the finite mixture model is provided in the following subsection.

1. The Student's-t mixture model and training with EM

The univariate Student's-t distribution is defined using parameters mean μ, standard deviation σ > 0 and degrees of freedom ν > 0 as follows:

$$p(x;\ \mu,\sigma^2,\nu)\ = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{|\sigma|(\pi\nu)^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)[1+\frac{(x-\mu)^2}{\nu\sigma^2}]} \quad (1)$$

where Γ stands for the Gamma function. It can be shown that the Student's-t can be analyzed as a two-stage generative model by introducing a Gamma-distributed latent variable, on which the variance of normally distributed data depends. More specifically, according to this scheme data X are assumed to be distributed as

$$X \mid \mu,\sigma^2,\nu,u \sim N(\mu,\ \sigma^2/u)$$

where u is a random variable distributed as

$$u \sim Gamma(\nu/2, \nu/2)$$

Integrating out the weights u leads to (eq. 1) [8]. Therefore, the Student's-t can be seen as an infinite sum of Gaussians with the same mean but different variance. This leads in effect to a distribution that can integrate extreme values more robustly than the Gaussian. Also, the Gaussian can be seen as a special case of the Student's-t, as for $\nu \to +\infty$ the Student's-t distribution tends to a Gaussian with standard deviation σ. Conversely, for $\nu \to 0$ the distribution's tails become heavier. Note that the Cauchy distribution [21] can also be seen as a special case of the Student's-t, obtained specifically for $\nu = 1$.

The Student's-t mixture model (SMM) is defined as:

$$\varphi(x;\ \theta)\ =\ \sum_{i=1}^{K} \pi_i p(x; \mu_i, \sigma_i^2, \nu_i) \tag{2}$$

where $x = [x^1, \dots, x^N]^T$ denotes the observed-data vector and $\theta = [\pi^1, \dots, \pi^N, \mu^1, \dots, \mu^N, \sigma^1, \dots, \sigma^N, \nu^1, \dots, \nu^N]^T$ correspond to the parameters of the components of the mixture.

In order to estimate Maximum Likelihood (ML) values for the parameters of the SMM we use the EM algorithm [8, 12]. We consider the complete data vector $X^c = [x^1, \dots, x^N, z^1, \dots, z^N, u^1, \dots, u^N]^T$ which includes the observed data vector plus the model latent random variables. Latent random variables $z^1, \dots, z^N$ are component-labeled vectors and z_{ij} is either one or zero, according to whether the observation x_j is generated or not by the i^{th} component. The E-step of the EM algorithm requires the calculation of the posterior probability of the model latent variables given the observations. Thus, on the $(t+1)^{th}$ iteration of the E- step we calculate the posterior probability that the datum x^j belongs to the i^{th} component of the mixture:

$$z_{ij}^{(t+1)} = \frac{\pi_i^{(t)} p(x^j; \mu_i^{(t)}, \sigma_i^{2(t)}, \nu_i^{(t)})}{\sum_{m=1}^{K} \pi_m^{(t)} p(x^j; \mu_m^{(t)}, \sigma_m^{2(t)}, \nu_m^{(t)})} \tag{3}$$

where in our case the number of kernels K is fixed to 2. Kernel i = 1 corresponds to the class of distances between letters of the same word; kernel i = 2 corresponds to the class of distances between letters of different words. In the E- step the expectation of the variance factors u_{ij} for each observation is also computed:

$$u_{ij}^{(t+1)} = \frac{\nu_i^{(t)} + 1}{\nu_i^{(t)} + (x^j - \mu_i^{(t)})^2 / \sigma_i^{2(t)}} \tag{4}$$

M- step update equations are obtained by maximizing the log-likelihood of the complete data with respect to the Student's-t parameters of both kernels and their weights:

$$\pi_i^{(t+1)} = \frac{1}{N}\sum_{j=1}^{N} z_{ij}^{(t)}, \quad \mu_i^{(t+1)} = \frac{\sum_{j=1}^{N} z_{ij}^{(t)} u_{ij}^{(t)} x_j}{\sum_{j=1}^{N} z_{ij}^{(t)} u_{ij}^{(t)}},$$
$$\sigma_i^{2(t+1)} = \frac{\sum_{j=1}^{N} z_{ij}^{(t)} u_{ij}^{(t)} (x_j - \mu_i^{(t+1)})^2}{\sum_{j=1}^{N} z_{ij}^{(t)}} \tag{5}$$

The update for the degrees of freedom parameters ν cannot be computed in closed form; it is given by the solution to the following equation:

$$\log\left(\frac{\nu_i^{(t+1)}}{2}\right) - \psi\left(\frac{\nu_i^{(t+1)}}{2}\right) + 1 - \log\left(\frac{\nu_i^{(t)}+1}{2}\right) + \\ + \frac{\sum_{j=1}^{N} z_{ij}^{(t)} (log u_{ij}^{(t)} - u_{ij}^{(t)})}{\sum_{j=1}^{N} z_{ij}^{(t)}} + \psi\left(\frac{\nu_i^{(t)}+1}{2}\right) = 0 \tag{6}$$

where ψ stands for the digamma function. A detailed derivation of the EM algorithm for Student's-t mixtures is presented in [13].

2. Assigning distances to classes

The distances calculated for all text lines of a document image are assigned to one of K=2 classes. In effect, we convert the soft classification result expressed as a probability $p(z^1, \dots, z^N | x^1, \dots, x^N)$ and obtained by the ML parameters into a hard assignment of each distance to a single class. A distance j can be thus classified by comparing its posterior probability of belonging to the one or the class given the observation. Formally:

$$p(z_{1j} = 1 | x^j) \lesseqgtr p(z_{2j} = 1 | x^j) \Leftrightarrow$$
$$\pi_1\ p(x^j; \mu_1, \sigma_1^2, \nu_1) \lesseqgtr \pi_2\ p(x^j; \mu_2, \sigma_2^2, \nu_2)$$

IV. Evaluation Results

We evaluated the performance of the proposed word segmentation method against four state-of-the-art word segmentation methods which follow the same protocol for the first stages of the word segmentation procedure while differentiate from the proposed method at the gap classification stage. These methods comprise: a) Gaussian mixture modeling [14], b) sequential clustering [3], c) average linkage clustering [3] and d) modified max clustering [3]. MATLAB code for the proposed Student's-t model is publicly available [23].

The evaluation method we followed is robust and well established since it corresponds to the protocol of the ICDAR 2009 Handwriting Segmentation Contest [17]. The accuracy was measured in terms of Detection Rate (*DR*), Recognition Accuracy (*RA*) and the final performance metric F-Measure (*FM*). The abovementioned metrics use the number of ground truth words (*N*), the number of result words (*M*) and the number of one-to-one matches (*o2o*). The evaluation metrics depend only on the selection of the acceptance threshold T_a. A more detailed description of these metrics can be found in [17].

In order to check the stability of the proposed method, we experimented on five different datasets[1]. Table I summarizes information of the datasets in terms of the total number of document images and the total number of words.

[1] The dataset named here "Appian" is referred to as GRPOLY-DB-MachinePrinted-C in its original publication [20].

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS. RESULT COLUMNS CORRESPOND TO S (SCENARIO, SEE TEXT), M (NUMBER OF RESULTING SEGMENTED WORDS, COMPARE WITH ACTUAL NUMBER OF WORDS SHOWN IN THE FIRST TABLE), O2O (NUMBER OF EXACT MATCHES OF SEGMENTED AND ACTUAL WORDS), DR (DETECTION RATE), RA (RECOGNITION ACCURACY), FM (F-MEASURE). BEST VALUES PER METHOD/BENCHMARK/SCENARIO ARE SHOWN IN BOLD.

Dataset	Number of Document Images	Number of Words
ICDAR 2007 [16]	80	13311
ICDAR 2009 [17]	200	29717
ICFHR 2010 [18]	100	15130
ICDAR 2013 [19]	150	23525
APPIAN [20]	315	65875

ICDAR07 Method	*S*	*M*	*o2o*	*DR*	*RA*	*FM*
Student's-t	S1	13086	12171	91.44	**93.01**	**92.22**
	S2	13051	12137	91.18	**93.00**	**92.08**
Gaussian	S1	13618	**12267**	**92.16**	90.08	91.11
	S2	13528	**12331**	**92.64**	91.15	91.89
Sequential Clustering	S1	12843	11853	89.05	92.29	90.64
	S2	13288	12121	91.06	91.22	91.14
Average Linkage	S1	11224	9951	74.76	88.66	81.12
	S2	13032	11888	89.31	91.22	90.26
Modified Max	S1	2832	1115	8.38	39.37	13.81
	S2	11698	9508	71.43	81.28	76.04

ICDAR09 Method	*S*	*M*	*o2o*	*DR*	*RA*	*FM*
Student's-t	S1	30131	28389	95.53	**94.22**	**94.87**
	S2	29968	28255	95.08	**94.28**	**94.68**
Gaussian	S1	31109	**28405**	**95.59**	91.31	93.40
	S2	30841	**28501**	**95.91**	92.41	94.13
Sequential Clustering	S1	27962	26063	87.70	93.21	90.37
	S2	29585	27767	93.44	93.85	93.65
Average Linkage	S1	23638	21180	71.27	89.60	79.39
	S2	29396	27255	91.72	92.72	92.21
Modified Max	S1	10969	7576	25.49	69.07	37.24
	S2	26722	24473	82.35	91.58	86.72

ICFHR10 Method	*S*	*M*	*o2o*	*DR*	*RA*	*FM*
Student's-t	S1	15351	13851	91.55	**90.23**	**90.88**
	S2	15240	13805	91.24	**90.58**	**90.91**
Gaussian	S1	16169	**13935**	**92.10**	86.18	89.04
	S2	15975	**13970**	**92.33**	87.45	89.82
Sequential Clustering	S1	15049	13537	89.47	89.95	89.71
	S2	15332	13691	90.49	89.30	89.89
Average Linkage	S1	13296	10887	71.96	81.88	76.60
	S2	14836	13236	87.48	89.22	88.34
Modified Max	S1	4033	2511	16.60	62.26	26.21
	S2	13097	10848	71.70	82.83	76.86

ICDAR13 Method	*S*	*M*	*o2o*	*DR*	*RA*	*FM*
Student's-t	S1	23153	20790	88.37	**89.79**	**89.08**
	S2	23150	20791	88.38	**89.81**	**89.09**
Gaussian	S1	23870	**20960**	**89.10**	87.81	88.45
	S2	23718	**21034**	**89.41**	88.68	89.05
Sequential Clustering	S1	21460	18943	80.52	88.27	84.22
	S2	22458	19844	84.35	88.36	86.31
Average Linkage	S1	16501	13840	58.83	83.87	69.16
	S2	22568	19893	84.56	88.15	86.32
Modified Max	S1	5355	2487	10.57	46.44	17.22
	S2	17854	14199	60.36	79.53	68.63

APPIAN Method	*S*	*M*	*o2o*	*DR*	*RA*	*FM*
Student's-t	S1	66795	**65375**	**99.24**	**97.87**	**98.55**
	S2	66688	**65342**	**99.19**	**97.98**	**98.58**
Gaussian	S1	68101	64795	98.36	95.15	96.73
	S2	67581	65153	98.90	96.41	97.64
Sequential Clustering	S1	64052	62079	94.24	96.92	95.56
	S2	65390	63812	96.87	97.59	97.23
Average Linkage	S1	32060	24930	37.84	77.76	50.91
	S2	64712	62781	95.30	97.02	96.15
Modified Max	S1	18131	9383	14.24	51.75	22.34
	S2	56653	53419	81.09	94.29	87.19

For all the above datasets and methods, two different scenarios (*S*) were defined. According to the first scenario (S1), all the distances appearing in the document image were used for the gap classification stage. For the second scenario (S2), the largest distances appearing in each document image which correspond to the 2% of the total number of distances, were excluded from the classification. The idea for the definition of two scenarios was to check the effectiveness of methods into a scenario were outliers are defined using an empirical criterion and are 'manually' discarded. We assume that 2% of the distances correspond to outliers.

We used the text line segmentation ground truth as input to the word segmentation algorithms. Furthermore, the acceptance threshold used was T_a=90%. It should be stressed that a direct comparison with the evaluation results of the participating methods of the competitions is not fair since in the case of the competitions the input to the word segmentation algorithm corresponds to the output of an automatic text line segmentation method whereas in our case the input corresponds to the text line segmentation ground truth.

Table I presents comparative experimental results for all abovementioned datasets and for the two previously described scenarios (S1 and S2). In Figure 3, comparative experimental results in terms of FM for all benchmarking datasets are presented. It is clear that the Student's-t method outperforms all state-of-the-art methods when we do not attempt to prune extreme values manually as in S2, proving its ability to better model data without making any supplementary assumptions about which values should be considered aberrant and pruned.

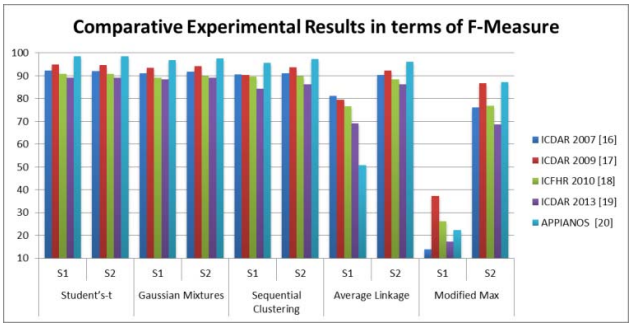


Fig. 3. Comparative experimental results for all benchmarking datasets and both scenarios in terms of F-Measure. All methods except the Student's-t (proposed) necessitate pruning using a manually selected threshold for optimal performance.

Summarizing the results shown in table I, we can draw the following conclusions. In all datasets and given any benchmark, the top two methods are the Student's-t mixture-based model and the Gaussian mixture-based model. In terms of FM and RA, the Student's-t always outperforms the Gaussian model. In terms of o2o and RA, the results are more ambiguous, with the Gaussian taking the lead in most cases. When pruning extreme value inputs (S2), the Student's-t result remain largely the same, with only a very slight difference upwards or downwards at most. On the other hand, all other methods show improved results. This means that pruning values using a manually chosen threshold is required for them to obtain optimal performance, unlike the Student's-t which in

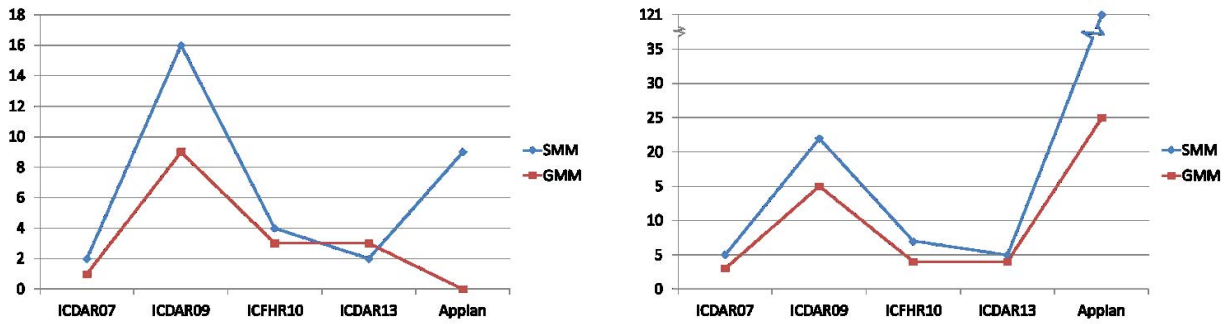


Fig. 4. Comparison of SMM (proposed) versus GMM in terms of number of perfectly segmented pages per dataset. Segmentation correctness is evaluated in terms of Recognition Accuracy (left) and Detection Rate (right). F-measure -based results coincide with results based on Recognition Accuracy.

a sense has a built-in mechanism of dealing with extreme values / outliers. Selecting an appropriate threshold manually may not be always successful and may result in pruning useful values. This problem can become a serious issue especially for methods that seem to suffer in performance when no extreme values are beforehand pruned, like the Average Linkage and Modified Max methods. Perhaps most importantly, all results retain their relative rankings, e.g. the Student's-t is still winning always in terms of RA and FM in both S1 and S2, even if absolute numbers have changed.

We compare the Student's-t with the Gaussian in figure 4, where we show numbers of pages that have been *perfectly* segmented with respect to a number of benchmarks. We consider segmentation perfect when the respective benchmark has attained a 100% value. Note that the Student's-t consistently outperforms the Gaussian model (save for only a single case).

Figure 5 shows the segmentation result using the proposed SMM versus the result using the fore-running GMM model for a representative example. Additionally, a graphical illustration of the 2 kernels of the Student's-t mixture modeling produced by the proposed method (intra vs inter word gaps) superimposed on the calculated distances is presented (Fig.5c). In this figure, it can be observed that the Student's-t is much less affected by the presence of extreme values (note the few points in fig.5c around and over x=90 for example). This is reflected here as a much more conservative variance value of the inter-word mixture kernel for the Student's-t (90.1 vs 225.9 for the GMM corresponding kernel), which eventually leads to a much more accurate classification (fig. 5d,e).

Concerning the machine-printed dataset (Appian Dataset), we would expect a very high accuracy for all state-of-the-art methods. However, only the Student's-t method achieves accuracy very close to 100% with respect to all benchmarks. Also interestingly, the Student's-t model achieves a significantly high number of perfect segmentations (fig. 4) with respect to the Gaussian model. This result seems to suggest that when a Euclidean distance-based partition of data does exist (more likely in typewritten data, and is essentially the underlying premise of the finite mixture-based methods), the Student's-t model is the most appropriate to identify the partition correctly.

V. CONCLUSIONS

A novel word segmentation method is presented. It is based on the use of the Student's-t distribution for modeling the two class distributions (intra vs inter word gaps). The Student'-t distribution overcomes one of the main drawbacks of several methods reported in the literature which is the influence of their accuracy to the existence of extreme input values. Extensive experimentation on several publicly available handwritten datasets, a historical typewritten dataset of Greek polytonic text and comparing to several state-of-the-art methods proves the efficiency of the method.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 600707 (project tranScriptorium). This work has been also supported by the OldDocPro project (ID 4717) funded by the GSRT.

REFERENCES

[1] G. Seni, and E. Cohen, "External Word Segmentation of Off-line Handwritten Text Lines", Pattern Recognition, vol. 27, no. 1, pp. 41-52, Jan. 1994.

[2] U. Mahadevan, and R. C. Nagabushnam, "Gap metrics for word separation in handwritten lines", Proc. 3rd Int'l Conf. on Document Analysis and Recognition (ICDAR'95), pp. 124-127, 1995.

[3] S.H. Kim, S. Jeong, G.- S. Lee, C.Y. Suen, "Word segmentation in handwritten Korean text lines based on gap clustering techniques", Proc. 6th Int'l Conf. on Document Analysis and Recognition (ICDAR'01), pp. 189-193, 2001.

[4] T. Varga, and H. Bunke, "Tree structure for word extraction from handwritten text lines", Proc. 8th Int'l Conf. on Document Analysis and Recognition (ICDAR'05), pp. 352-356, 2005.

[5] G. Kim, and V. Govindaraju, "Handwritten Phrase Recognition as Applied to Street Name Images", Pattern Recognition, vol. 31, no. 1, pp. 41-51, Jan. 1998.

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

[7] F. Luthy, T. Varga, and H. Bunke, "Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation", Proc. 9th Int'l Conf. on Document Analysis and Recognition (ICDAR'07), pp. 8-12, 2007.

[8] G. Sfikas, C. Nikou, and N. Galatsanos. "Robust Image Segmentation with Mixtures of Student's-t Distributions", IEEE International Conference on Image Processing, 2007. ICIP 2007. Vol. 1, pp. 273-276, 2007.

[9] T. N. Minh, and Q. M. J. Wu, "Robust Student's-t Mixture Model with Spatial Constraints and its Application in Medical Image Segmentation", IEEE Transactions on Medical Imaging 31(1), pp. 103-116, 2012.

[10] G. Sfikas, C. Heinrich, J. Zallat, C. Nikou, and N. Galatsanos, "Recovery of Polarimetric Stokes Images by Spatial Mixture Models", Journal of the Optical Society of America A 28.3, pp. 465-474, 2011.

[11] K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Group-sparse Adaptive Variational Bayes Estimation", Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), pp.1342-1346, September 2014.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood Estimation via the EM Algorithm", Journal of the Statistical Royal Society B 39.1 pp. 1-38, 1977.

[13] D. Peel, and G. J. McLachlan, "Robust Mixture Modelling using the t Distribution", Statistics and computing 10(4), pp. 339-348, 2000.

[14] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Text line and word segmentation of handwritten documents", Pattern Recognition, 42 (12), pp. 3169-3183, 2009.

[15] A. Vinciarelli and J. Luettin, "A new normalization technique for cursive handwritten words", Pattern Recognition Letters, vol. 22, no. 9, pp. 1043-1050, 2001.

[16] B. Gatos, A. Antonacopoulos and N. Stamatopoulos, "ICDAR2007 Handwriting Segmentation Contest", 9th International Conference on Document Analysis and Recognition (ICDAR'07), pp. 1284-1288, Curitiba, Brazil, September 2007.

[17] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICDAR2009 Handwriting Segmentation Contest", 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp. 1393-1397, Barcelona, Spain, July 2009.

[18] B. Gatos, N. Stamatopoulos and G. Louloudis, "ICFHR 2010 Handwriting Segmentation Contest", 12th International Conference on Frontiers in Handwriting Recognition (ICFHR'10), pp. 737-742, Kolkata, India, November 2010.

[19] N. Stamatopoulos, G. Louloudis, B. Gatos, U. Pal and A. Alaei, "ICDAR2013 Handwriting Segmentation Contest", 12th International Conference on Document Analysis and Recognition (ICDAR'13), pp. 1402-1406, Washington DC, USA, August 2013.

[20] B. Gatos, N. Stamatopoulos, G. Louloudis, G. Sfikas, G. Retsinas, F. Simistira, V. Papavassiliou, V. Katsouros, "GRPOLY-DB: An old Greek polytonic document image database", 13th International Conference on Document Analysis and Recognition (ICDAR'15), Nancy, France, Ausust 2015.

[21] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[22] G. Sfikas, "Modèles statistiques non linéaires et non gaussiens compacts pour l'analyse de structures cérébrales", PhD Thesis, Université de Strasbourg, 2012

[23] http://www.cs.uoi.gr/~sfikas

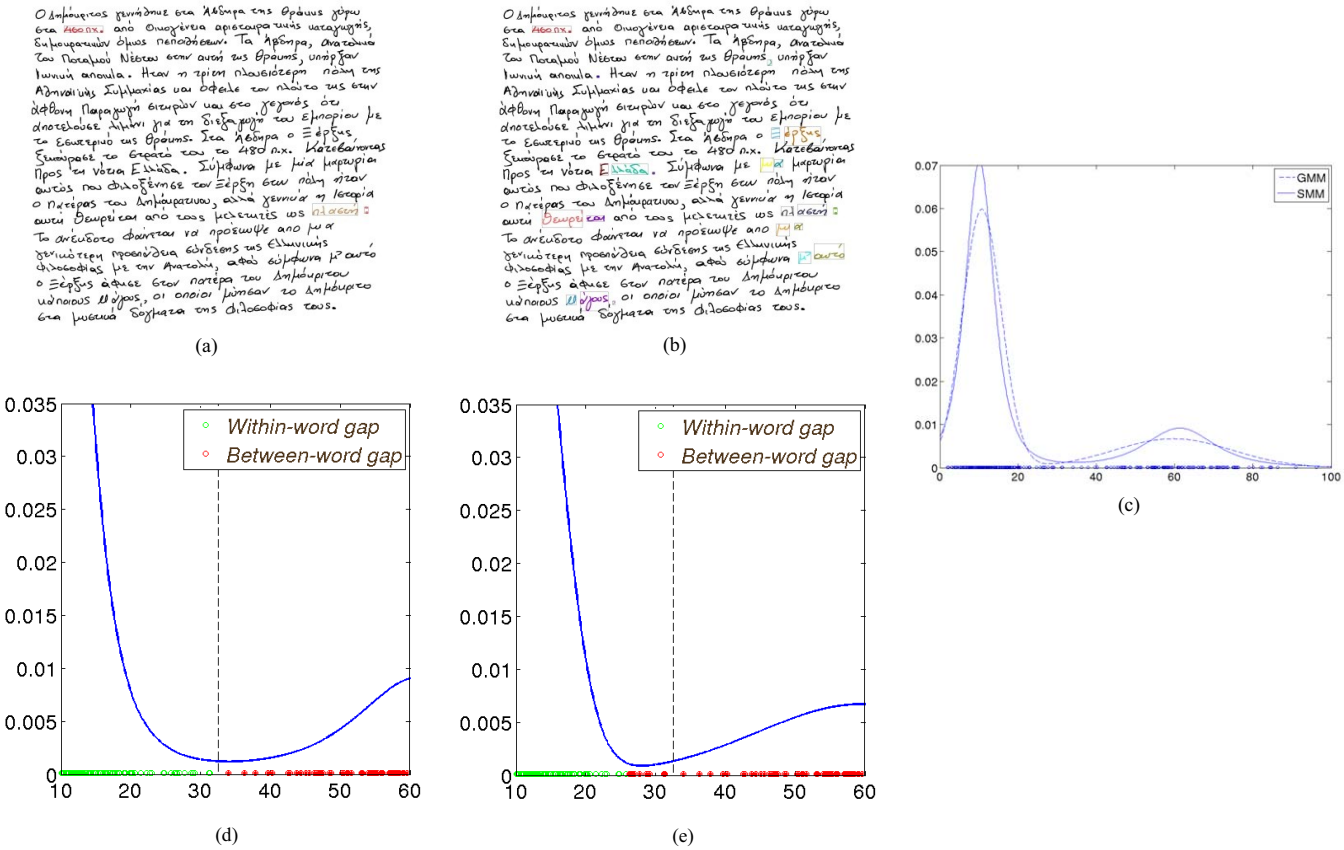


Fig. 5: Analysis of results on document image with id 137 of the ICDAR 2009 dataset (a) Word segmentation result using the SMM (proposed); all words are segmented correctly except words in colour. (b) Segmentation and errors committed by the GMM. (c) SMM fit (solid curve) versus GMM fit (dashed curve) on distance data. Intra (inter)-word gaps correspond to the mixture component on the left (right). (d), (e) Detail of data and fit shown in fig. 4c, with the optimal separating threshold overlaid as a black dashed line. The SMM fit (proposed) is shown on the left (fig.5d) and the GMM fit is shown on the right (fig.5e). Data point colour corresponds to class assignment. The Student's-t between-word component has a much smaller variance (σ^2=90.1) than its Gaussian counterpart (σ^2=225.9). This is related to the distribution's robustness to extreme values, which leads to a more conservative estimate of data variance and eventually better separation of the data around the actual threshold.