Pattern Recognition ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Text line and word segmentation of handwritten documents

G. Louloudis[a,*], B. Gatos[b,1], I. Pratikakis[b,1], C. Halatsis[a]

[a]Department of Informatics and Telecommunications, University of Athens, Greece
[b]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", 153 10 Athens, Greece

ARTICLE INFO

Article history:
Received 8 August 2008
Received in revised form 12 November 2008
Accepted 21 December 2008

Keywords:
Handwritten document image analysis
Hough transform
Text line segmentation
Word segmentation
Gaussian mixture modeling

ABSTRACT

In this paper, we present a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words. Text line segmentation is achieved by applying Hough transform on a subset of the document image connected components. A post-processing step includes the correction of possible false alarms, the detection of text lines that Hough transform failed to create and finally the efficient separation of vertically connected characters using a novel method based on skeletonization. Word segmentation is addressed as a two class problem. The distances between adjacent overlapped components in a text line are calculated using the combination of two distance metrics and each of them is categorized either as an inter- or an intra-word distance in a Gaussian mixture modeling framework. The performance of the proposed methodology is based on a consistent and concrete evaluation methodology that uses suitable performance measures in order to compare the text line segmentation and word segmentation results against the corresponding ground truth annotation. The efficiency of the proposed methodology is demonstrated by experimentation conducted on two different datasets: (a) on the test set of the ICDAR2007 handwriting segmentation competition and (b) on a set of historical handwritten documents.

1. Introduction

Segmentation of a document image into its basic entities, namely, text lines and words, is considered as a non trivial problem to solve in the field of handwritten document recognition. The difficulties that arise in handwritten documents make the segmentation procedure a challenging task. Different types of difficulties are encountered in the text line segmentation and the word segmentation procedure. In the case of text line segmentation procedure, major difficulties include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching. Furthermore, the frequent appearance of accents in many languages (e.g. French, Greek) makes the text line segmentation a challenging task. In the case of word segmentation, difficulties that arise include the appearance of skew and slant in the text line, the existence of punctuation marks along the text line and the non-uniform spacing of words which is a common residual in handwritten documents.

In this paper, we present a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words. The main novelties of the proposed approach consist of (i) the extension of a previously published work for text line segmentation [1] taking into account an improved methodology for the separation of vertically connected text lines and (ii) a new word segmentation technique based on an efficient distinction of inter and intra-word gaps using the combination of two different distance metrics. The distance metrics that we use comprise the Euclidean distance metric and the convex hull-based metric. The distinction of the two classes is considered as an unsupervised clustering problem for which we make use of the Gaussian mixture theory in order to model the two classes. Extensive experimentation proves the efficiency of the proposed methodology against different combinations of gap metrics on two different datasets including (a) modern and (b) historical handwritten documents. The test set of ICDAR2007 handwriting segmentation competition is used as the set of modern handwritten documents. A consistent benchmarking enables the comparison of the proposed methodology against the state-of-the-art.

The paper is organized as follows: Section 2 is dedicated to the related work for text line and word segmentation of handwritten documents. In Section 3, the proposed methodology to segment text lines is detailed. Section 4 deals with the proposed word segmentation method. In Section 5, we present the experimental results and, finally, Section 6 describes conclusions and future work.

* Corresponding author. Tel.: +30 210 7275241; fax: +30 210 7275201.
E-mail addresses: louloud@mm.di.uoa.gr (G. Louloudis), bgat@iit.demokritos.gr (B. Gatos), ipratika@iit.demokritos.gr (I. Pratikakis), halatsis@di.uoa.gr (C. Halatsis)
URLs: http://www.di.uoa.gr, http://www.iit.demokritos.gr/cil.
[1] Tel.: +30 2106503183; fax: +30 2106532175.

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.12.016

2. Related work

A wide variety of segmentation methods for handwritten documents has been reported in the literature. We categorize these methods depending on whether they refer to text line segmentation, word segmentation or both text line and word segmentation. To this end, an analytical description of the related work on the text line segmentation problem is presented in Section 2.1. Section 2.2 refers to the word segmentation methodologies that exist in the literature. Finally, Section 2.3 presents description of works that encounter both text line and word segmentation problems.

2.1. Text line segmentation

There are mainly three basic categories that these methods lie in: methods making use of the projection profiles, methods that are based on the Hough transform and, finally, smearing methods. Also, several methods exist that cannot be included in a certain category since they do not share a common guideline.

Methods that make use of the projection profiles include partitioning into vertical strips and horizontal run calculation [2] as well as calculation of the projection profiles of every vertical strip (chunk) and traverse around any obstructing handwritten connected component (CC) by associating it to the text line above or below [3]. According to methods that make use of the Hough transform, a set of points of the initial image comprise the input while the lines that fit best to this set of points are calculated. The set of points considered in the voting procedure of the Hough transform is usually either the gravity centers [4] or minima points [5] of the CCs. A recent block based Hough transform approach [1] takes into account the gravity centers of parts of CCs. Smearing methods include the fuzzy RLSA [6] and the adaptive RLSA [7]. The fuzzy RLSA measure is calculated for every pixel on the initial image and describes "how far one can see when standing at a pixel along horizontal direction". The adaptive RLSA is an extension of the classical RLSA [8] in the sense that additional smoothing constraints are set in regard to the geometrical properties of neighboring CCs. Other methodologies include the use of the adaptive local connectivity map [9] by summing the intensities of each pixel's neighbors in the horizontal direction, the use of foreground/background transitions counting combined with a min-cut/max-flow graph cut algorithm [10] as well as of the theory of Kalman filtering to detect text lines on low resolution images [11]. An overview of text line segmentation methods is presented in Table 1. For a more detailed description of methodologies for text line segmentation, interested readers should consult [12].

2.2. Word segmentation

Algorithms dealing with word segmentation in the literature are based primarily on analysis of geometric relationship of adjacent components. Components are either CCs or overlapped components (OCs). An OC is defined as a set of CCs whose projection profiles overlap in the vertical direction. Related work for the problem of word segmentation differs in two aspects. The first aspect is the way the distance of adjacent components is calculated while the second aspect concerns the approach used to classify the previously calculated distances as either between word gaps or within word gaps. Most of the methodologies described in the literature have a preprocessing stage which includes noise removal, skew and slant correction.

Many distance metrics are defined in the literature. Seni et al. [22] presented eight different distance metrics. These included 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 [23] which was called convex hull-based metric. After comparing this metric with some of the metrics of [22] concluded that the convex hull-based metric performs better than the other ones. Kim et al. [24] 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 [25] extended 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. the relative sizes of the surrounding gaps. Experiments that have been conducted with different gap metrics as well as threshold types showed that their methodology yielded improvements over conventional word extraction methods.

In all the aforementioned methodologies the gap classification threshold used derives: (i) from the processing of the calculated distances, (ii) from the processing of the whole text line image or (iii) after the implementation 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 [26–28]. An early published work making use of classifiers for the word segmentation problem is the work of Kim and Govindaraju [26]. After a character segmentation methodology which was performed in the whole text line image to determine the possible segmentation points and the calculation of a feature vector for every possible segmentation point, a simple neural network was adopted to determine the segmentation points. The neural network used had eight input units, four hidden units and one output unit. A similar work was presented in [27] by Huang and Srihari. This approach claimed two differences from previous methods: (i) the gap metric was computed by combining three different distance measures, which avoided the weakness of each of the individual one and thus provided a more reliable distance measure and (ii) besides the local features, such as the current gap, a new set of global features were also extracted to help the classifier make a better decision. The classification was done by using a three-layer neural network. The feature vector contained eleven features. The neural network had eleven input units, four hidden units and two output units. The authors report a 90.82% overall accuracy on the "Cedar Letter" documents while the method described in [26] presented an accuracy of 87.36%. Finally, a different approach was presented from Luthy et al. [28]. The problem of segmenting a text line into words was considered as a text line recognition task, adapted to the characteristics of segmentation. That is, at a certain position of a text line, it had to be decided whether the considered position belonged to a letter of a word, or to a space between two words. For this purpose, three different recognizers based on hidden Markov models were designed, and results of writer-dependent as well as writer-independent experiments were reported.

2.3. Text line and word segmentation

Apart from papers that deal only with the text line or the word segmentation problem, there exist works that consider both the text line and word segmentation problem. In these works, the input to the methodology is a handwritten page image (in contrast to the methodologies presented in Section 2.2 where a text line image

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

Table 1
Text line segmentation methods of handwritten documents.

Authors	Category	Short description
Bruzzone and Coffetti [2]	Projection profiles method	The algorithm is based on the analysis of horizontal run projections and connected components grouping and splitting on a partition of the input image into vertical strips, in order to deal with undulate or skewed text. Goal of the algorithm is to preserve the ascending and descending characters from been corrupted by arbitrary cuts. The algorithm has been designed for cursive text and it can be applied also to hand-printed one
Arivazhagan et al. [3]	Projection profiles method	The projection profile of every vertical strip (chunk) is calculated. The first candidate lines are extracted among the first chunks. These lines traverse around any obstructing handwritten connected component by associating it to the text line above or below. This decision is made by either (i) modeling the text lines as bivariate Gaussian densities and evaluating the probability of the component for each Gaussian or (ii) the probability obtained from a distance metric
Likforman et al. [4]	Hough transform method	Potential alignments are hypothesized in the Hough domain and validated in the image domain. The gravity centers of the connected components are the units for the Hough transform
Pu and Shi [5]	Hough transform method	The Hough transform is first applied to minima points (units) in a vertical strip on the left of the image. The alignments in the Hough domain are searched starting from a main direction, by grouping cells in an exhaustive search in six directions. Then a moving window, associated with a clustering scheme in the image domain, assigns the remaining units to alignments. The clustering scheme (*natural learning algorithm*) allows the creation of new lines starting from the middle of the pages
Louloudis et al. [1]	Hough transform method	The methodology incorporates a block based Hough transform approach which takes into account the gravity centers of parts of connected components. After the first candidate text line extraction, a post-processing step is used to correct possible splitting as well as to detect text lines that the previous step did not reveal. A key idea in the whole procedure is the partitioning of the connected component domain into three distinct sub-domains each of which is treated in a different manner
Shi and Govindaraju [6]	Smearing method	The fuzzy RLSA measure is calculated for every pixel on the initial image and describes "how far one can see when standing at a pixel along horizontal direction". By applying this measure, a new grayscale image is created which is binarized and the lines of text are extracted from the new image
Gatos et al. [7]	Smearing method	The adaptive RLSA is an extension of the classical RLSA in the sense that additional smoothing constraints are set in regard to the geometrical properties of neighboring connected components. The replacement of background pixels with foreground pixels is performed when these constraints are satisfied
Shi et al. [9]	Other	A methodology that makes use of the adaptive local connectivity map. The input to the method is a grayscale image. A new image is calculated by summing the intensities of each pixel's neighbors in the horizontal direction. Since the new image is also a grayscale image, a thresholding technique is applied and the connected components are grouped into location maps by using a grouping method
Douglas et al. [10]	Other	A technique that uses the count of foreground/background transitions in a binarized image to determine areas of the document that are likely to be text lines. Also, a min-cut/max-flow graph cut algorithm is used to split up text areas that appear to encompass more than one line of text. A merging of text lines containing relatively little text information to nearby text lines is then applied
Lemaitre and Camillerapp [11]	Other	A methodology which is based on a notion of perceptive vision: at a certain distance, text lines can be seen as line segments. This method is based on the theory of Kalman filtering to detect text lines on low resolution images
Nicolas et al. [13]	Other	A method that confronts the problem from an Artificial Intelligence perspective. The aim is to cluster the connected components of the document into homogeneous sets that correspond to the text lines of the document. To solve this problem, a search over the graph that is defined by the connected components as vertices and the distances among them as edges is applied
Li et al. [14]	Other	A technique that models text line detection as an image segmentation problem by enhancing text line structures using a Gaussian window and adopting the level set method to evolve text line boundaries
Weliwitage et al. [15]	Other	A method that involves cut text minimization for segmentation of text lines from handwritten English documents. In doing so, an optimization technique is applied which varies the cutting angle and start location to minimize the text pixels cut while tracking between two text lines
Basu et al. [16]	Other	A technique for multi-skewed document of handwritten English or Bengali text. It is assumed that hypothetical water flows, from both left and right sides of the image frame, face obstruction from characters of text lines. The stripes of areas left unwetted on the image frame are finally labeled for extraction of text lines
Yin and Liu [17]	Other	An approach based on minimum spanning tree (MST) clustering with new distance measures. First, the connected components of the document image are grouped into a tree by MST clustering with a new distance measure. The edges of the tree are then dynamically cut to form text lines by using a new objective function for finding the number of clusters. This approach is totally parameter-free and can apply to various documents with multi-skewed and curved lines
Yin and Liu [18]	Other	A methodology based on minimum spanning tree (MST) clustering with distance learning. Given a distance metric, the connected components of the document image are grouped into a tree structure. Text lines are extracted by dynamically cutting the edges of the tree using a new objective function. For avoiding artificial parameters and improving the segmentation accuracy, the distance metric is designed by supervised learning
Stamatopoulos et al. [19]	Other	A combination method of different segmentation techniques is presented. The goal is to exploit the segmentation results of complementary techniques and specific features of the initial image so as to generate improved segmentation results
Roy et al. [20]	Other	A technique that is based on morphological operation and run-length smearing. At first RLSA is applied to get individual word as a component. Next, the foreground portion of this smoothed image is eroded to get some seed components from the individual words of the document. Erosion is also done on background portions to find some boundary information of text lines. Finally, using the positional information of the seed components and the boundary information, the lines are segmented
Du et al. [21]	Other	A method that is based on the Mumford–Shah model is described. The algorithm is script independent. In addition, morphing is used to remove overlaps between neighbouring text lines and connect broken ones

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

was used as input). Manmatha [29] presented a methodology to extract words from handwritten historical documents. The input was a grayscale handwritten page image. The text line segmentation procedure was based on the calculation of the vertical projection profile. The vertical projection profile was smoothed with a Gaussian filter in order to eliminate false alarms and to reduce sensitivity to noise. The local maxima of the smoothed profile defined the space between text lines. The methodology for word segmentation was based on the creation of blobs. A blob can be regarded as a connected region in space. In order to calculate the blobs, a differential expression based on second order partial Gaussian derivatives was used. The experimental results reported a word error rate of 17.4% based on the assumption that a hit was counted if only 60% of the area of the word was found in the result. Ataer et al. [30] presented a methodology in which both text line segmentation and word segmentation procedures made use of projection profiles. The experimental results showed 100% accuracy on the text line segmentation procedure and 82% accuracy on word segmentation. The work of Marti and Bunke [31] was based on the projection profiles for text line segmentation. They used the convex hull-based metric for the computation of the distances of adjacent components and a threshold to classify the calculated distances which was based on characteristics of the text line image to be segmented. The threshold was defined per text line. Kim et al. [32] presented a system that performed word recognition starting from a handwritten document image. The text line segmentation module was based on the notion that the baseline of a handwritten text line is well-defined, as people seem to write on an imaginary line on which the core of each word of the text line resides. This imaginary baseline was approximated by the local minima points of each CC. The word segmentation module used was the one described in [26]. Experimental results showed an overall 95.3% accuracy on 20 document images of different writing styles (discrete, cursive, mixed) for text line segmentation and 94.5% accuracy for word segmentation. Recently, Stafylakis et al. [33] presented a novel methodology for both text line and word segmentation. The text line segmentation module used a Viterbi algorithm while an SVM-based metric was adopted to locate words in each text line.

3. Text line segmentation

The proposed methodology for text line segmentation in handwritten document images deals with the following challenges: (i) each text line that appears in the document may have an arbitrary skew angle and converse skew angle along the text line, (ii) text lines may have different skew directions, (iii) accents may be cited either above or below the text line and (iv) parts of neighboring text lines may be connected.

The text line segmentation methodology is based on [1] and includes three stages: (i) pre-processing, (ii) Hough transform mapping and (iii) post-processing, wherein a novel methodology is presented for the efficient separation of vertically connected characters.

3.1. Pre-processing

The pre-processing stage consists of three steps. In the first step, the CCs of the binary image are extracted. Then, the average character height *AH* for the whole document image is calculated based on the average height of all CCs. We assume that the average character height equals to the average character width *AW*. The final step includes the partitioning of the CCs domain into three sub-domains which are denoted as "Subset 1", "Subset 2" and "Subset 3". These sub-domains are treated in a different manner by the methodology.

"Subset 1" is expected to contain all components which correspond to the majority of the characters with size which satisfies the following constraints:

$$(0.5 * AH \leqslant H < 3 * AH) \quad \text{and} \quad (0.5 * AW \leqslant W) \tag{1}$$

where H, W denote the component's height and width, respectively, and AH, AW denote the average character height and the average character width, respectively.

"Subset 2" is expected to contain all large CCs. Large components are either capital letters or characters from adjacent text lines which touch each other. The height of these components is defined by the following equation:

$$H \geqslant 3 * AH \tag{2}$$

Finally, "Subset 3" should contain characters like accents, punctuation marks and small characters. The equation that defines this set is

$$((H < 3 * AH) \quad \text{and} \quad (0.5 * AW > W)) \quad \text{or}$$
$$((H < 0.5 * AH) \quad \text{and} \quad (0.5 * AW < W)) \tag{3}$$

3.2. Hough transform mapping

In this stage, Hough transform takes into consideration only the CCs' "Subset 1". In our approach, instead of having only one representative point for every CC, a partitioning is applied for each CC lying in "Subset 1", to equally sized blocks, so as to have more representative points voting in the Hough domain. An exception might be applied on the rightmost block. The width of each block is defined by the average character width *AW*. An example is shown in Fig. 1. After the creation of blocks, we calculate the gravity center of the CC contained in each block. The set of all these points contributes to the Hough transform.

The Hough transform is a line to point transformation from the Cartesian space to the Polar coordinate space. A line in the Cartesian coordinate space is described by the equation

$$x \cos(\theta) + y \sin(\theta) = p \tag{4}$$

It is easily observed that the line in the Cartesian space is represented by a point in the Polar coordinate space whose coordinates are p and θ. Every gravity center in the subset corresponds to a set of cells in the accumulator array of the (p,θ) domain. To construct the Hough domain, the resolution along θ direction was set to 1° letting θ take values in the range of 85–95° and the resolution along p direction was set to $0.2*AH$ (as in [4]).

After the computation of the accumulator array we proceed to the following procedure: We detect the cell (p_i, θ_i) having the maximum contribution and we assign to the text line (p_i, θ_i) all points that vote in the area $(p_i - 5, \theta_i) \ldots (p_i + 5, \theta_i)$. To decide whether a CC belongs to a text line, at least half of the points representing the corresponding blocks must be assigned to this area. After the assignment of a CC to a text line, all votes that correspond to this particular CC are removed from the Hough transform accumulator array. This procedure is repeated until cell (p_i, θ_i) having the maximum contribution contains less than n_1 votes in order to avoid false alarms. During the evolution of the procedure, the dominant skew angle of currently detected lines is calculated. In the case that the cell (p_i, θ_i) has a maximum contribution less than n_2 $(n_2 > n_1)$, an additional constraint is applied upon which, a text line is valid only if the corresponding skew angle of the line deviates from the dominant skew angle less than 2°.

3.3. Post-processing

The post-processing stage consists of two steps. At the first step, (i) a merging technique over the result of the Hough transform is

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

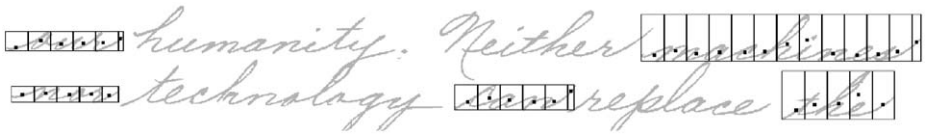

Fig. 1. An example showing the partitioning of connected components to blocks of width *AW* and the corresponding gravity centers. All connected components not placed in a bounding box correspond to either "Subset 2" or "Subset 3".

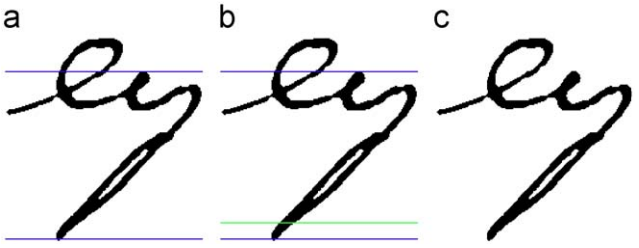

Fig. 2. Example of a connected component that is not split. (a) lines y_i; (b) area under green 'dotted' line checked in Eq. (5); (c) final result.

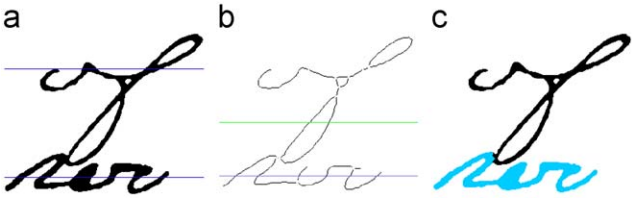

Fig. 3. Example of a connected component that is split into two parts. (a) lines y_i; (b) zone Z_1 (starting from green 'dotted' line to blue 'solid' line); (c) final result.

applied to correct some false alarms and (ii) CCs of "Subset 1" that were not clustered to any line are examined to determine whether a new line is detected (see [1]). After the detection of the final set of lines, all components lying in "Subset 3" as well as those unclassified components of "Subset 1" become grouped to the closest line.

The second step deals with large components lying in the sub-domain "Subset 2". All components of this subset mainly belong to n detected text lines ($n > 1$). Our novel methodology for splitting these components consists of the following:

(A) Calculate y_i, which are the average y values of the intersection of detected line i and the CCs' bounding box ($i = 1 \dots n$) (see Figs. 2–4(a)).

(B) Exclude from the procedure the last line n if the condition at Eq. (5) is not satisfied.

$$\frac{\sum_{y=y_n-(y_n-y_{n-1})/10}^{y_e}\sum_{x=x_s}^{x_e} I(x,y)}{\sum_{y=y_{n-1}}^{y_e}\sum_{x=x_s}^{x_e} I(x,y)} > 0.08 \tag{5}$$

where (x_s,y_s), (x_e,y_e) are the coordinates of the bounding box of the component (see Fig. 2(b) and (c)) and I the image of the component (value 1 for foreground and 0 for background pixels). Eq. (5) verifies that the component area near line n is due to a vertical character merging and not due to a long character descender from text line $n-1$ (see Fig. 2(c)).

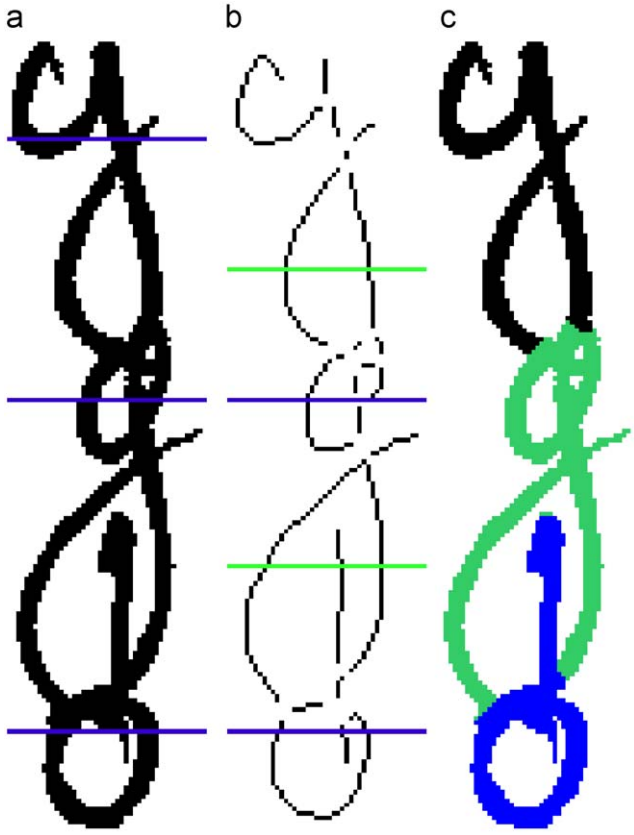

Fig. 4. Example of a connected component split into three parts. (a) lines y_i; (b) zone Z_i (starting from green 'dotted' line to blue 'solid' line); (c) final result.

(C) For every line i, $i = 1 \dots n-1$, we define zones Z_i taking into account the following constraint:

$$y_i + \frac{y_{i+1} - y_i}{2} < y < y_{i+1} \tag{6}$$

Then, we compute the skeleton of the CC, detect all junction points and remove them from the skeleton if they lie inside zone Z_i. If no junction point exists in the segmentation zone Z_i we remove all skeleton points on the center of the zone (Figs. 3(b) and 4(b)).

(D) For every zone z_i flag with id '1' the skeleton parts that intersect with line i. All other parts are flagged with id '2'. Finally, in each zone z_i separation of the initial CC into different segments is accomplished by assigning to a pixel the id of the closest skeleton pixel (Figs. 3(c) and 4(c)).

4. Word segmentation

The word segmentation procedure is divided into two steps. The first step deals with the computation of the distances of adjacent components in the text line image and the second step concerns the

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

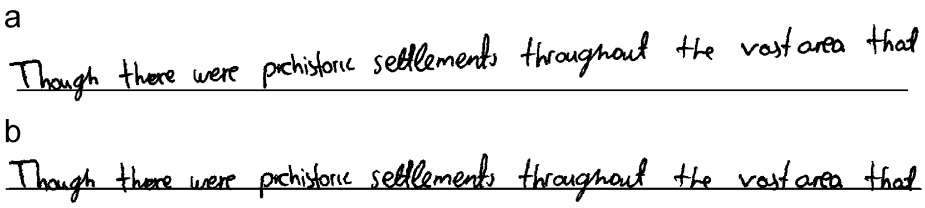

Fig. 5. (a) A text line image and (b) the same text line after skew correction.

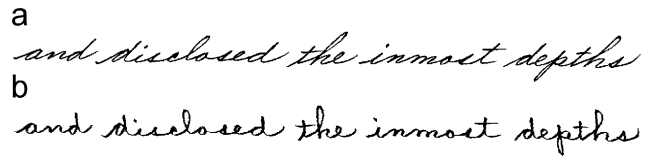

Fig. 6. (a) A text line image and (b) the same text line after slant correction.

classification of the previously computed distances as either inter-word gaps or inter-character gaps. For the first step, we propose the average of two different metrics: the Euclidean distance metric [22] and the convex hull-based metric [23]. The classification of the computed distances is performed using a well-known methodology from the area of unsupervised clustering techniques, the Gaussian mixtures.

4.1. Distance computation

In order to calculate the distance of adjacent components in the text line image, a pre-processing procedure is applied. The pre-processing procedure concerns the correction of the skew angle as well as the dominant slant angle [34] of the text line image (Figs. 5 and 6). The computation of the gap metric is considered not on the CCs but on the OCs, where an OC is defined as a set of CCs whose projection profiles overlap in the vertical direction (see Fig. 7).

We define as distance of two adjacent OCs the average value of the Euclidean distance and the convex hull-based distance. The Euclidean distance between two adjacent OCs is defined as the minimum Euclidean distance among the Euclidean distances of all pairs of points of the two adjacent OCs (Fig. 8). For the calculation of the Euclidean distance we apply a fast scheme that takes into consideration only a subset of the pixels of the left and right OCs instead of the whole number of black pixels. In order to define the subset of pixels of the left OC, we include in this subset the rightmost black pixel of every scanline. The subset of pixels for the right OC is defined by including the leftmost black pixel of every scanline. Finally, the Euclidean distance of the two OCs is defined as the minimum of the Euclidean distances of all pairs of pixels.

We calculate the convex hull-based metric as follows: Given a pair of adjacent OCs C_i and C_{i+1}, let H_i and H_{i+1} be their convex hulls, respectively. Let L be the line joining the centers of gravity (or centroid) of H_i and H_{i+1}. Let P_i and P_{i+1} be the points of intersection of L with the hulls H_i and H_{i+1}, respectively. The gap between the two convex hulls is defined as the Euclidean distance between the points P_i and P_{i+1} (see Fig. 9).

4.2. Gap classification

For the gap classification problem a novel approach is used. This approach is based on the unsupervised classification of the already computed distances into two distinct classes representing the word inter-class and the word intra-class, respectively. To this end, we adopt the use of Gaussian mixtures, a methodology which, to the best of our knowledge, was never used in previous works on word segmentation. A mixture model based clustering is based on the idea that each cluster is mathematically presented by a parametric distribution. We have a two clusters problem so every cluster is modeled with a Gaussian distribution. The algorithm that is used to calculate the parameters for the Gaussians is the EM algorithm. We use this methodology since the Gaussian mixtures is a well known unsupervised clustering technique with many advantages which comprise: (i) the mixture model covers the data well, (ii) a density estimation for each cluster can be obtained and (iii) a "soft" classification is available. For a detailed description on the Gaussian mixtures the reader should consult [35].

5. Evaluation and experimental results

5.1. Performance evaluation methodology

In the literature, the performance evaluation of a segmentation algorithm is mainly based on visual criteria in order to calculate the percentage of the correct segments. Manual observation of the segmentation result is a very tedious, time consuming and not in all cases unbiased process. To avoid user intervention, we use an automatic performance evaluation technique based on comparing the detected segmentation result with an already annotated ground truth. Similar evaluation strategies have been followed in several document segmentation competitions, such as ICDAR2003, ICDAR2005 and ICDAR2007 [36–38] page segmentation and ICDAR2007 handwriting segmentation competitions [7]. The performance evaluation is based on counting the number of matches between the areas detected by the algorithm and the areas in the ground truth. We use a table whose values are calculated according to the overlap of the labeled pixel sets as either text lines or words and the ground truth.

Let I be the set of all image points, G_i the set of all points inside the i text line (or word) ground truth region, R_j the set of all points inside the j text line (or 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:

$$MatchScore\,(i,j) = \frac{T(G_i \cap R_j \cap I)}{T((G_i \cup R_j) \cap I)} \tag{7}$$

The performance evaluator searches within the *MatchScore* Table for pairs of one-to-one, one-to-many or many-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 th. A g_one-to-many match ($go2m$) is a ground truth text line (word) that "partially" matches with two or more text lines (or words) in the detected result. A g_many-to-one match ($gm2o$) corresponds to two or more ground truth text lines (or words) that "partially" match with one detected text line (or word). A d_one-to-many match ($do2m$) is a detected text line (or word) that "partially" matches two or more text lines (or words) in the ground truth. Finally, a d_many-to-one match ($dm2o$) corresponds to two or more detected text lines (or

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

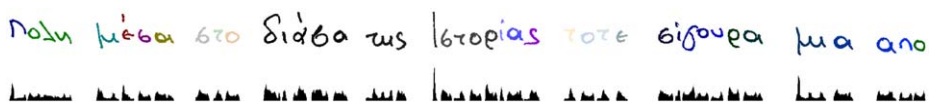

Fig. 7. The overlapped components of the text line image are defined based on the projection profile.

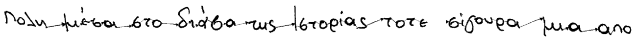

Fig. 8. The arrows define the Euclidean distances between adjacent overlapped components.

words) that "partially" match one text line (or word) in the ground truth.

If N is the count of ground truth segments (text lines or words), M is the count of result segments (text lines or words), and w_1, w_2, w_3, w_4, w_5, w_6 are pre-determined weights, we calculate the detection rate (DR) and recognition accuracy (RA) as follows:

$$DR = w_1 \frac{o2o}{N} + w_2 \frac{g\_o2m}{N} + w_3 \frac{g\_m2o}{N} \tag{8}$$

$$RA = w_4 \frac{o2o}{M} + w_5 \frac{d\_o2m}{M} + w_6 \frac{d\_m2o}{M} \tag{9}$$

where the entities $o2o$, g_o2m, g_m2o, d_o2m and d_m2o are calculated from *MatchScore* table (Eq. (7)) following the steps of [39] and correspond to the number of one-to-one, g_one-to-many, g_many-to-one, d_one-to-many and d_many-to-one, respectively.

A global performance metric can be defined if we combine the values of *DR* and *RA*. We can define the following *F-Measure* (*FM*):

$$FM = \frac{2 * DR * RA}{DR + RA} \tag{10}$$

5.2. Experimental results

The proposed methodology is tested on a modern set of handwritten images as well as on a historical set of handwritten images. The modern set that we use is the test set of the ICDAR2007 handwriting segmentation competition [7]. To this end, we conducted several experiments which are described in detail in the following sections. For the sake of comparison we implemented three state of the art techniques for text line segmentation and two for word segmentation. For text line segmentation we implemented the projection profiles [30], the fuzzy RLSA [6] and the Hough based approach described in [4]. The techniques to which we compare the proposed methodology for the word segmentation procedure are the projection profiles [30] and the RLSA [8]. Parameters n_1 and n_2 in the proposed text line segmentation methodology (Section 3.2) are experimentally defined ($n_1 = 5$, $n_2 = 9$). Also, the evaluator's acceptance threshold *th* (Section 5.1) for text line and word segmentation is defined 0.95 and 0.9, respectively, in order to comply with the corresponding values of the ICDAR2007 handwriting segmentation contest [7].

5.2.1. The modern handwritten set

The modern set of handwritten images is the test set of ICDAR2007 handwriting segmentation competition and contains 80 images. None of the documents include any non-text elements (such as lines, drawings, pictures and logos) and are all written in several languages including English, French, German and Greek. Almost all documents have two or more adjacent text lines touching in several areas. Some of them have variable skew angles among text lines. Furthermore, there are document images having text lines with different skew directions as well as document images having text lines with converse skew angles along the same text line. The appearance of accents is common in Greek and French handwritten documents. All documents are written from different writers and in the majority of the documents the distance of adjacent text lines is very small leading to a highly dense text. For all images, we have the corresponding ground truth in terms of text lines and words. The total number of text lines is 1773, while the corresponding number for words is 13311.

The detailed comparative results for text line segmentation in terms of DR, RA, FM and the number of matches are given in Table 2. In this table we also include all methodologies that participated in the handwriting segmentation competition of ICDAR2007. It is worth noting that our methodology slightly outperforms all the other approaches achieving DR 97.4% and RA 97.4%. Figs. 10 and 11 contain two text line segmentation results of the proposed methodology as well as of the three state-of-the-art methodologies.

Concerning word segmentation, we combined the distance metrics defined in Section 4.1 with two gap classification techniques. These comprise (i) the proposed Gaussian mixture classification methodology and (ii) the methodology described in [40] that uses the median of the lengths of the white pixels on the scanline which has the maximum number of black to white transitions. Table 3 contains experimental results for word segmentation in terms of DR, RA, FM and the number of matches. The average value of both distance metrics when using the proposed gap classification methodology (Gaussian mixtures) yields the best results (DR: 93.9%; RA: 90.8%). The word segmentation module takes as input the result of the proposed text line segmentation technique. Although, there is a small improvement in performance between the proposed methodology and the one presented in [40], the advantage of the proposed methodology is that it is parameter free, contrary to the methodology described in [40] where an experimentally defined factor is required for the calculation of the final threshold. Finally, Table 4 contains experimental results after combination of all text line segmentation methods with all word segmentation methods. In this table we have also included the results of all methodologies that participated in the ICDAR2007 Handwriting segmentation competition. As it is reported in Table 4, the proposed methodology outperforms all other methodologies. In order to visualize the results of the word segmentation procedure we demonstrate the word segmentation result of the proposed methodology and of the best combinations of text line and word segmentation on two handwritten documents in Figs. 12 and 13, respectively.

5.2.2. The historical handwritten set

The set of historical handwritten images contains 40 images taken from the historical archives of the University of Athens [41] and from George Washington's collection from the Library of Congress [42]. Both subsets contain images with relatively straight text lines while the text is rather sparse. The appearance of accents is common on Greek historical handwritten documents. Finally, all documents are written by different writers. For all images, we have manually created the corresponding ground truth in terms of text lines and words. The total number of text lines was 1095 while the corresponding number for words was 8576.

The detailed comparative results for text line segmentation in terms of DR, RA, FM and the number of matches are listed in Table 5.

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

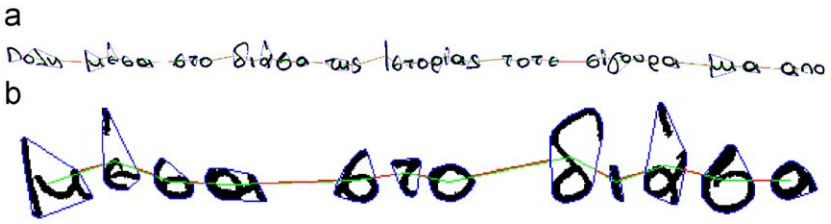

Fig. 9. (a) Handwritten text line image (b) zoomed version of handwritten text line image. The convex hulls are defined with the blue color. The green line determines the line segment that connects two gravity centers. Finally, the red line segment is the convex hull-based distance.

Table 2
Comparative experimental results for line segmentation over the test set of ICDAR2007 handwriting segmentation competition.

Line segmentation	*N*	*M*	*DR*	*RA*	*FM*	*o2o*	*go2m*	*gm2o*	*do2m*	*dm2o*
Projection [30]	1773	1610	58.3	66.5	62.1	925	222	216	102	482
Fuzzy RLSA [6]	1773	1813	77.6	75.3	76.4	1288	76	277	126	186
Hough [4]	1773	1984	88.5	78.4	83.1	1532	14	136	66	29
ICDAR2007 Handwriting Competition										
ILSP_LWSeg	1773	1773	97.3	97	97.1	1713	5	34	17	10
PARC	1773	1756	92.2	93	92.6	1604	40	76	34	85
UOA-HT	1773	1770	95.5	95.4	95.4	1674	14	54	27	28
BESUS	1773	1904	86.6	79.7	83	1494	9	151	72	21
DUTH-ARLSA	1773	1894	73.9	70.2	72	1214	149	227	107	354
RLSA	1773	1877	44.3	45.4	44.8	632	264	346	122	757
Proposed	1773	1770	97.4	97.4	97.4	1717	6	34	17	13

Fig. 10. Text line segmentation result on a Greek handwritten document produced by (a) proposed method; (b) Hough method [4]; (c) fuzzy RLSA method [6] and (d) projection profile method [30].

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

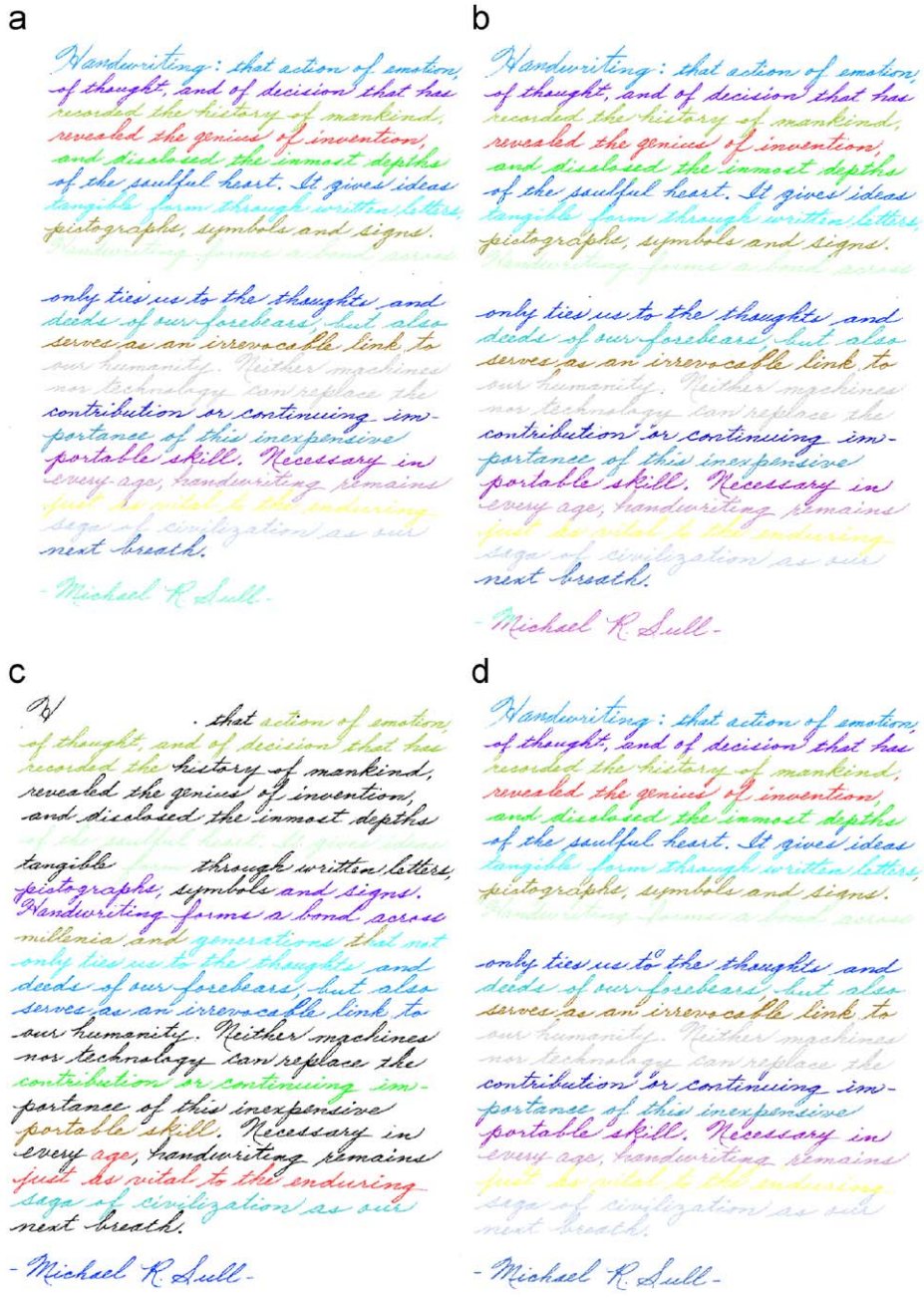

Fig. 11. Text line segmentation result on a cursive English handwritten document produced by (a) proposed method; (b) Hough method [4]; (c) fuzzy RLSA method [6] and (d) projection profile method [30].

Table 3
Experimental results for word segmentation using combinations of distance metrics and gap classification methodologies over the test set of ICDAR2007 handwriting segmentation competition.

Distance metric		Classification methods		*N*	*M*	*DR*	*RA*	*FM*	*o2o*	*go2m*	*gm2o*	*do2m*	*dm2o*
Euclidean	Convex	LT	GM										
✓		✓		13311	13322	91.8	91.7	91.7	11933	326	869	410	732
	✓	✓		13311	13249	91.8	92.3	92.0	11953	334	779	367	764
✓	✓	✓		13311	13334	92.4	92.1	92.2	12018	302	828	390	673
✓			✓	13311	13666	93.4	90.3	91.8	12106	202	1137	535	427
	✓		✓	13311	13622	93.2	90.5	91.8	12093	206	1082	514	454
✓	✓		✓	13311	13655	93.9	90.8	92.3	12190	175	1062	503	371

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

Table 4
Comparative experimental results for line and word segmentation over the test set of ICDAR2007 handwriting segmentation competition.

Line segmentation	Word segmentation	*N*	*M*	*DR*	*RA*	*FM*	o2o	go2m	gm2o	do2m	dm2o
Proposed	Proposed	13311	13655	93.9	90.8	92.3	12190	175	1062	503	371
Proposed	Projections [30]	13311	17762	85.8	59	69.9	9638	287	6876	2689	742
Proposed	RLSA [8]	13311	14931	83	71.2	76.6	9957	395	4022	1699	1045
Hough [4]	Proposed	13311	13873	89.8	85.3	87.5	11536	227	1489	670	554
Hough [4]	Projections [30]	13311	18734	83.8	54.3	65.9	9309	275	7128	2773	696
Hough [4]	RLSA [8]	13311	15306	82	68.3	74.5	9815	329	4118	1723	892
Fuzzy [6]	Proposed	13311	12782	81.2	84.5	82.8	10349	450	1391	649	1173
Fuzzy [6]	Projections [30]	13311	16581	74.3	55.1	63.3	8165	486	6444	2517	1388
Fuzzy [6]	RLSA [8]	13311	15024	81	69	74.5	9644	444	4133	1746	1181
Projections [30]	Proposed	13311	11645	64.1	74.8	69.0	7956	721	1594	724	2336
Projections [30]	Projections [30]	13311	15803	60.1	47	52.7	6308	655	6140	2330	2176
Projections [30]	RLSA [8]	13311	13826	75.3	70.8	73.0	8971	741	3513	1484	1811
ICDAR2007 Handwriting Competition											
ILSP-LWSeg	ILSP-LWSeg	13311	13027	90.3	92.4	91.3	11732	303	834	378	819
PARC	PARC	13311	14965	84.3	72.8	78.1	10246	422	3482	1524	1088
UOA-HT	UOA-HT	13311	13824	91.7	87.6	89.6	11794	263	1418	668	602
BESUS	BESUS	13311	19091	80.7	52	63.3	9114	327	6172	2449	823
DUTH-ARLSA	DUTH-ARLSA	13311	16220	80.2	61.3	69.5	9100	394	5896	2440	954
RLSA	RLSA	13311	13792	76.9	74	75.4	9566	639	2064	920	1633

Fig. 12. Word segmentation result on Greek handwritten document when using (a) proposed methodology, (b) Hough [4] for line segmentation and RLSA [8] for word segmentation, (c) fuzzy RLSA [6] for line segmentation and RLSA [8] for word segmentation and (d) projection profile [30] for line segmentation and RLSA [8] for word segmentation.

It is worth pointing out that our methodology outperforms the other three approaches achieving DR 98.9% and RA 99.1%. Fig. 14 contains a document image after text line segmentation using the proposed methodology and the three state of the art text line segmentation techniques.

The same combination of experiments that is applied for the modern set is also applied on the historical set on the word segmentation problem. Table 6 contains the experimental results in terms of DR, RA, FM and the number of matches. It is clear that the average of the two distance metrics using Gaussian mixtures as the gap classification methodology yields the best results (DR: 86.8%/RA: 84.2%). The input to the word segmentation module is the result of the proposed methodology for text line segmentation (as is the case for the modern set). Table 7 contains experimental results after combination

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

Fig. 13. Word segmentation result on cursive English handwritten document when using (a) proposed methodology, (b) Hough [4] for line segmentation and RLSA [8] for word segmentation, (c) fuzzy RLSA [6] for line segmentation and RLSA [8] for word segmentation and (d) projection profile [30] for line segmentation and RLSA [8] for word segmentation.

Table 5
Comparative experimental results for line segmentation over the historical set.

Line segmentation	*N*	*M*	*DR*	*RA*	*FM*	*o2o*	*go2m*	*gm2o*	*do2m*	*dm2o*
Projection [30]	1095	1098	77.4	77.2	77.3	800	63	128	57	134
Fuzzy [6]	1095	1093	83.9	83	83.4	881	19	132	65	41
Hough [4]	1095	1511	76.8	51.6	61.7	718	3	493	242	7
Proposed	1095	1093	98.9	99.1	99.0	1082	2	4	2	4

of all text line segmentation methods with all word segmentation methods. In order to visualize the results of the word segmentation procedures we demonstrate the word segmentation result of the proposed methodology and of the best combination of text line and word segmentation in Fig. 15.

Fig. 16 summarizes some problems that are encountered at the word segmentation procedure. Due to non uniform spacing between adjacent words there are cases that parts of adjacent words are merged (undersegmentation) and cases where parts of the same word are split into two or more words (oversegmentation). The

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

Fig. 14. Text line segmentation result on a historical document produced by (a) proposed method; (b) Hough method [4]; (c) fuzzy RLSA method [6] and (d) projection profile method [30].

Table 6
Experimental results for word segmentation using different distance metrics and different classification methods over the historical set.

Distance metric		Classification methods		*N*	*M*	*DR*	*RA*	*FM*	*o2o*	*go2m*	*gm2o*	*do2m*	*dm2o*
Euclidean	Convex	LT	GM										
✓		✓		8576	8080	81.1	87.8	84.3	6655	697	510	235	1546
	✓	✓		8576	8282	81.8	85.9	83.8	6674	643	733	342	1419
✓	✓	✓		8576	8305	83.4	87.2	85.3	6854	586	630	292	1280
✓			✓	8576	8685	85.9	84.8	85.3	7022	444	953	440	959
	✓		✓	8576	8810	85.1	82.5	83.8	6895	445	1190	554	956
✓	✓		✓	8576	8809	86.8	84.2	85.5	7087	397	1050	485	843

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

Table 7
Comparative experimental results for line and word segmentation over the historical set.

Line segmentation	Word segmentation	*N*	*M*	*DR*	*RA*	*FM*	*o2o*	*go2m*	*gm2o*	*do2m*	*dm2o*
Proposed	Proposed	8576	8809	86.8	84.2	85.5	7087	397	1050	485	843
Proposed	Projections [30]	8576	8869	67.7	66.4	67.0	5053	914	2118	907	2448
Proposed	RLSA [8]	8576	10056	64.7	53.1	58.3	4611	684	3096	1264	1665
Hough [4]	Proposed	8576	11698	72	47.9	57.5	5035	371	4217	1454	863
Hough [4]	Projections [30]	8576	11453	63	45.4	52.8	4215	823	3955	1705	2241
Hough [4]	RLSA [8]	8576	9946	60.6	50.1	54.9	4275	669	3022	1236	1631
Fuzzy [6]	Proposed	8576	8818	80.3	77.6	78.9	6454	456	1278	597	981
Fuzzy [6]	Projections [30]	8576	8887	62.2	60.6	61.4	4502	942	2394	1036	2529
Fuzzy [6]	RLSA [8]	8576	10084	63.7	52	57.3	4512	691	3113	1268	1679
Projections [30]	Proposed	8576	8324	72.9	75.9	74.4	5789	598	1256	558	1561
Projections [30]	Projections [30]	8576	8930	57.9	56.4	57.1	4161	926	2294	959	2573
Projections [30]	RLSA [8]	8576	9865	62.8	52.8	57.4	4420	770	3010	1229	1871

Fig. 15. Word segmentation result on a historical document when using (a) proposed methodology, (b) Hough [4] for line segmentation and RLSA [8] for word segmentation, (c) fuzzy RLSA [6] for line segmentation and RLSA [8] for word segmentation and (d) projection profile [30] for line segmentation and RLSA [8] for word segmentation.

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

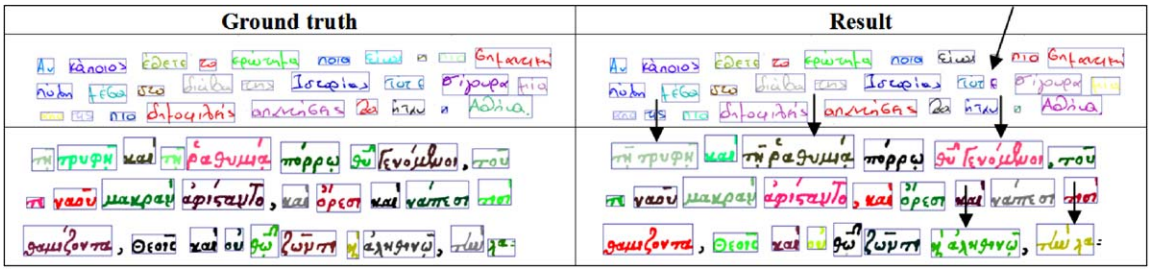

Fig. 16. Indicative errors of the proposed word segmentation methodology. Arrows indicate components on the result that are either undersegmented or oversegmented.

Table 8
Computational costs for text line segmentation.

Text line segmentation methodologies	Time (s)
Fuzzy run-length	25.0
Projection profiles [26]	0.1
Hough [4]	0.2
Proposed	0.4

Table 9
Detailed computational costs of the proposed word segmentation method.

Computational cost for proposed segmentation methodology	Time (s)
Skew detection and correction	2.2
Slant detection and correction	12.2
Gaussian mixture classification	0.9
Total	18.2

problem is even more obvious on the historical documents.

Table 8 presents the computational cost of all text line segmentation methodologies that are involved in our experimentation procedure. The environment used for development was Borland C++ 6 while the PC used was an Intel Core 2 Quad at 2.8 Ghz with 4 Gb of Ram. The typical size of the binary image is 2500×2000. Finally, Table 9 contains the computational cost of the proposed word segmentation method. We give detailed times for skew and slant detection and correction as well as Gaussian mixture classification. From this table it is obvious that most of the time is consumed for the detection and correction of the dominant slant angle.

6. Conclusions and future work

In this paper, we present a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words. The main novelties of the proposed approach consist of (i) the extension of a previously published work for text line segmentation [1] taking into account an improved methodology for the separation of vertically connected text lines and (ii) a new word segmentation technique based on an efficient distinction of inter and intra-word gaps using the combination of two different distance metrics. The distance metrics that we use comprise the Euclidean distance metric and the convex hull-based metric. The distinction of the two classes is considered as an unsupervised clustering problem for which we make use of the Gaussian mixture theory in order to model the two classes. From our experimental results it is shown that the proposed methodology outperforms existing state-of-the-art methods for text line and word segmentation in handwritten documents. Future work mainly concerns the improvement of the proposed word segmentation methodology using a punctuation detection method as well as a feedback from the character segmentation and recognition modules.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Program under Grant agreement no. 215604 (project IMPACT) and by the Greek Secretariat for Research and Development under the PENED 2001 framework.

References

[1] G. Louloudis, K. Halatsis, B. Gatos, I. Pratikakis, A block-based Hough transform mapping for text line detection in handwritten documents, in: The 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2006), La Baule, France, October 2006, pp. 515–520.
[2] E. Bruzzone, M.C. Coffetti, An algorithm for extracting cursive text lines, in: Proceedings of the Fifth International Conference on Document Analysis and Recognition, Bangalore, India, 1999, p. 749.
[3] M. Arivazhagan, H. Srinivasan, S.N. Srihari, A statistical approach to handwritten line segmentation, in: Document Recognition and Retrieval XIV, Proceedings of SPIE, San Jose, CA, USA, February 2007, p. 6500T-1-11.
[4] L. Likforman-Sulem, A. Hanimyan, C. Faure, A Hough based algorithm for extracting text lines in handwritten documents, in: Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, Canada, 1995, pp. 774–777.
[5] Y. Pu, Z. Shi, A natural learning algorithm based on hough transform for text lines extraction in handwritten documents, in: Proceedings of the 6th International Workshop on Frontiers in Handwriting Recognition, Taejon, Korea, 1998, pp. 637–646.
[6] Z. Shi, V. Govindaraju, Line separation for complex document images using fuzzy runlength, in: First International Workshop on Document Image Analysis for Libraries, 2004, p. 306.
[7] B. Gatos, A. Antonacopoulos, N. Stamatopoulos, ICDAR2007 Handwriting Segmentation Contest, in: The Ninth International Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, September 2007, pp. 1284–1288.
[8] F.M. Wahl, K.Y. Wong, R.G. Casey, Block segmentation and text extraction in mixed text/image documents, Computer Graphics and Image Processing 20 (1982) 375–390.
[9] Z. Shi, S. Setlur, V. Govindaraju, Text extraction from gray scale historical document images using adaptive local connectivity map, in: The Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 794–798.
[10] D.J. Kennard, W.A. Barrett, Separating lines of text in free-form handwritten historical documents, in: The Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp. 12–23.
[11] A. Lemaitre, J. Camillerapp, Text line extraction in handwritten document with Kalman filter applied on low resolution image, in: The Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp. 38–45.
[12] L. Likforman-Sulem, A. Zahour, B. Taconet, Text line segmentation of historical documents: a survey, International Journal on Document Analysis and Recognition 9 (2) (2007) 123–138.
[13] S. Nicolas, T. Paquet, L. Heutte, Text line segmentation in handwritten document using a production system, in: Proceedings of the Ninth IWFHR, Tokyo, Japan, 2004, pp. 245–250.
[14] Y. Li, Y. Zheng, D. Doermann, Detecting text lines in handwritten documents, in: The 18th International Conference on Pattern Recognition (ICPR'06), pp. 1030–1033.
[15] C. Weliwitage, A.L. Harvey, A.B. Jennings, Handwritten document offline text line segmentation, in: Proceedings of Digital Imaging Computing: Techniques and Applications, 2005, pp. 184–187.
[16] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D.K. Basu, Text line extraction from multi-skewed handwritten documents, Pattern Recognition 40 (6) (2007) 1825–1839.
[17] F. Yin, C. Liu, Handwritten text line extraction based on minimum spanning tree clustering, in: International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, November 2007, pp. 1123–1128.

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016

[18] F. Yin, C. Liu, Handwritten text line segmentation by clustering with distance metric learning, in: International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 229–234.
[19] N. Stamatopoulos, B. Gatos, S.J. Perantonis, A method for combining complementary techniques for document image segmentation, in: International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 235–240.
[20] P.P. Roy, U. Pal, J. Llados, Morphology based handwritten line segmentation using foreground and background information, in: International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 241–246.
[21] X. Du, W. Pan, T.D. Bui, Text line segmentation in handwritten documents using Mumford–Shah model, in: International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 253–258.
[22] G. Seni, E. Cohen, External word segmentation of off-line handwritten text lines, Pattern Recognition 27 (1) (1994) 41–52.
[23] U. Mahadevan, R.C. Nagabushnam, Gap metrics for word separation in handwritten lines, in: The Third International Conference on Document Analysis and Recognition, Montreal, Canada, 1995, pp. 124–127.
[24] S.H. Kim, S. Jeong, G.-S. Lee, C.Y. Suen, Word segmentation in handwritten Korean text lines based on gap clustering techniques, in: The Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 189–193.
[25] T. Varga, H. Bunke, Tree structure for word extraction from handwritten text lines, in: The Eight International Conference on Document Analysis and Recognition, Seoul, Korea, 2005, pp. 352–356.
[26] G. Kim, V. Govindaraju, Handwritten phrase recognition as applied to street name images, Pattern Recognition 31 (1) (1998) 41–51.
[27] C. Huang, S. Srihari, Word segmentation of off-line handwritten documents, in: Proceedings of the Document Recognition and Retrieval (DRR) XV, IST/SPIE Annual Symposium, San Jose, CA, USA, January 2008.
[28] F. Luthy, T. Varga, H. Bunke, Using hidden Markov models as a tool for handwritten text line segmentation, in: The Ninth International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 8–12.
[29] R. Manmatha, J.L. Rothfeder, A scale space approach for automatically segmenting words from historical handwritten documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1212–1225.
[30] E. Ataer, P. Duygulu, Retrieval of Ottoman documents, in: Proceedings of the Eighth ACM SIGMM International Workshop on Multimedia Information Retrieval, 26–27 October 2006, Santa Barbara, CA, USA.
[31] U.V. Marti, H. Bunke, Text line segmentation and word recognition in a system for general writer independent handwriting recognition, in: The Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 159–163.
[32] G. Kim, V. Govindaraju, S. Srihari, An architecture for handwritten text recognition systems, International Journal on Document Analysis and Recognition 2 (1999) 37–44.
[33] T. Stafylakis, V. Papavassiliou, V. Katsouros, G. Carayiannis, Robust text-line and word segmentation for handwritten documents images, in: International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008, pp. 3393–3396.
[34] A. Vinciarelli, J. Luettin, A new normalization technique for cursive handwritten words, Pattern Recognition Letters 2 (9) (2001) 1043–1050.
[35] J.M. Marin, K. Mengersen, C.P. Robert, Bayesian Modelling and Inference on Mixtures of Distributions, Handbook of Statistics, vol. 25, Elsevier-Sciences, Amsterdam, 2005.
[36] A. Antonacopoulos, B. Gatos, D. Karatzas, ICDAR 2003 page segmentation competition, in: The Seventh International Conference on Document Analysis and Recognition (ICDAR'03), Edinburgh, Scotland, August 2003, pp. 688–692.
[37] A. Antonacopoulos, B. Gatos, D. Bridson, ICDAR2005 page segmentation competition, in: The Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea, August 2005, pp. 75–79.
[38] A. Antonacopoulos, B. Gatos, D. Bridson, ICDAR2007 page segmentation competition, in: The Ninth International Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, September 2007.
[39] I. Phillips, A. Chhabra, Empirical performance evaluation of graphics recognition systems, IEEE Transaction of Pattern Analysis and Machine Intelligence 21 (9) (1999) 849–870.
[40] G. Louloudis, B. Gatos, I. Pratikakis, Line and word segmentation of handwritten documents, in: International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 247–252.
[41] ⟨http://pergamos.lib.uoa.gr/dl/object/col:histarch⟩.
[42] ⟨http://memory.loc.gov/ammem/gwhtml/gwhome.html⟩.

About the Author—GEORGIOS E. LOULOUDIS was born in 1979 in Athens, Greece. He received his Computer Science diploma in 2000 and his Master degree in 2002, both from the Department of Informatics and Telecommunications of University of Athens, Athens, Greece. He is currently a Ph.D. student in the Department of Informatics and Telecommunications, University of Athens, Greece, in the field of Handwritten Character Recognition. His main research interests are in image processing and document image analysis, OCR and pattern recognition.

About the Author—BASILIS G. GATOS was born in 1967, in Athens, Greece. He received his Electrical Engineering Diploma in 1992 and his Ph.D. degree in 1998, both from the Electrical and Computer Engineering Department of Democritus University of Thrace, Xanthi, Greece. His Ph.D. thesis is on Optical Character Recognition Techniques. In 1993, he was awarded a scholarship from the Institute of Informatics and Telecommunications, NCSR "Demokritos", where he worked till 1996. From 1997 to 1998 he worked as a Software Engineer at Computer Logic S.A. From 1998 to 2001 he worked at Lambrakis Press Archives as a Director of the Research Division in the field of digital preservation of old newspapers. From 2001 to 2003 he worked at BSI S.A. as Managing Director of R&D Division in the field of document management and recognition. He is currently working as a Researcher at the Institute of Informatics and Telecommunications of the National Center for Scientific Research "Demokritos", Athens, Greece. His main research interests are in image processing and document image analysis, OCR and pattern recognition. He has more than 80 publications in journals and international conference proceedings and has participated in several research programs funded by the European community. He is a Member of the Technical Chamber of Greece and Program Committee Member of several international Conferences (e.g. ICFHR 2008, ICDAR 2009).

About the Author—IOANNIS PRATIKAKIS received the Diploma degree in Electrical Engineering from the Demokritus University of Thrace, Xanthi, Greece in 1992, and the Ph.D. degree in 3D Image processing and analysis from Vrije Universiteit Brussel, Brussels, Belgium, in January 1999. From March 1999 to March 2000, he was at IRISA/ViSTA group, Rennes, France as an INRIA postdoctoral fellow. He is currently working as a Research Scientist at the Institute of Informatics and Telecommunications of the National Center for Scientific Research "Demokritos", Athens, Greece. His research interests include document image analysis, 2D and 3D image analysis, image and volume sequence analysis and content-based multimedia search and retrieval with a particular focus on 3D objects. He has published more than 90 papers in journals, book chapters and conference proceedings in the above areas. He has participated in more than 15 national and international R&D projects. Finally, he has served as a Co-Chair of the Eurographics Workshop on 3D object retrieval in 2008 and 2009 as well as a Co-Editor for the special issue on 3D object retrieval at the International Journal of Computer Vision.

About the Author—CONSTANTIN HALATSIS was born in Athens, Greece, in 1941. He received his B.Sc. degree in Physics in 1964 and his M.Sc. degree in Electronics in 1966, both from the University of Athens. In 1971 he received his Ph.D. degree in Computer Science from the University of Manchester, UK. From 1971 to 1981 he was a researcher with the Computer Centre of NRC Demokritos in Athens. In 1981 he became Full Professor of Computer Science at the University of Thessaloniki, Greece, and in 1988 he was invited as Full Professor of CS at the Department of Informatics and Telecommunications of the University of Athens, where he is now since then. He has a wide area of research interests including logic design, computer architectures, fault-tolerant computing, artificial intelligence, and theory of computation. He is the Author/Co-Author of more than 100 technical papers published in refereed international scientific journals and conference proceedings.

Please cite this article as: G. Louloudis, et al., Text line and word segmentation of handwritten documents, Pattern Recognition (2009), doi: 10.1016/j.patcog.2008.12.016