Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-169476
IOS Press

Unsupervised multi-language handwritten text line segmentation

Miguel Ángel García-Calderón*, René Arnulfo García-Hernández and Yulia Ledeneva
Autonomous University of the State of Mexico, Instituto Literario #100, Col. Centro, Toluca, State of Mexico

Abstract. Text Lines Segmentation (TLS) affects the performance of Manuscript Text Recognition (MTR) systems from document images. At the same time, the TLS task consists of two tasks: the first is Text Lines Localization (TLL) and the second is the Search of the Path that Divides neighboring Lines (SPDL) of handwritten text. The TLS task depends on the type of language, author's writing style, pen type and document quality. In this paper, Projected Energy Map with Alpha blending (PEM-Alpha) is presented as an unsupervised method for the TLL task, which can work with lines that are touching or overlapping. In addition, SPDL-GA is proposed as a method for SPDL task which finds the line that best splits the text. The experimentation is carried out with a standard collection of historical multilingual documents. Through experimentation it is demostrated that the proposed methods outperform other state-of-the-art methods, even in documents with mixed languages. In addition, few parameters required by PEM-Alpha and SPDL-GA are automatically calculated.

Keywords: Handwritten text line segmentation, text line segmentation, document image processing, projection profile, segmentation, historical documents

1. Introduction

It is estimated that the writing was created in the year 3200 B.C. which allowed the human to transmit their knowledge to later generations. Today, there are large digital collections of historical documents in libraries and national archives for free access. There are projects to facilitate access to the handwritten information stored in libraries and national archives [1, 2]. However, all that knowledge has been exploited little because access to such information requires experienced paleographers in writing styles and variations of languages. There are projects and platforms that facilitate the manual transcription of manuscript documents [2]. An example is the Transcribed Bentham project in which 1,009 manuscripts were transcribed in a period of 6 months employing 1,207 people [2].

In the case that a document has more than one language, it would be necessary for the human to know the languages involved in order to transcribe the document. An example of a document with more than one language is the rosette stone that allowed the translation of Egyptian hieroglyphs in the seventeenth century [3].

That is why there is a need to generate systems that allow analyzing images of documents with the aim of recognizing manuscript texts. Manuscript Text Recognition (MTR) systems need as input the image of the line to be transcribed. Therefore, the systems of Text Lines Segmentation (TLS) have to locate and then to extract the text lines from the image of a page. The first stage of the TLS task is Text Lines Localization (TLL), where the starting and ending point of a line must be determined. From the points found should be Search the Path that Divides Lines (SPDL) of handwritten text that best separates two

*Corresponding author. Miguel Ángel García-Calderón, Autonomous University of the State of Mexico, Instituto Literario #100, Col. Centro, Toluca 50000, State of Mexico. E-mail: mgcalderon@outlook.es.

neighboring lines of text from the start point to the end point [4].

The problem of MTR and TLS is that the writing contained in the documents is consistent with the language, time, region and style of the author [5]. In the writing style of the author, there are variables such as character shape, character size, space between characters, space between lines, touching lines, overlapping lines (see Fig. 1), ornamentation and scratched text.

However, before locating and extracting lines of text, it is necessary to preprocess the images in order to eliminate inherent variations in document quality and digitization such as surface type, noise, resolution, inclination, etc. [6].

All the above variables increase the complexity of the TLS, thus some developed methods are for specific languages and writing styles [7, 8].

There are standard collections with several documents in different languages for the TLS of handwritten documents, such as the one presented in ICDAR 2009. However, it is not publicly available. In [9] a public collection is created using part of the documents of ICDAR 2009 collection and other ones. The collection in [9] has different languages such as Spanish, English, Arabic, Chinese and Arabic-Spanish. For such reason, it is necessary to develop TLS methods that can work even in documents with mixed languages.

Figures 1, 2 and 3 show an example of handwritten document in Spanish, Arabic and Arabic-Spanish, respectively.

There are several methods proposed for TLL: projection-based [10–12] grouping [13, 14] and learning methods [8, 15], etc. In these works, to be able to compare them to other methods they solve the problem of SPDL by means of a line between the previously found points.

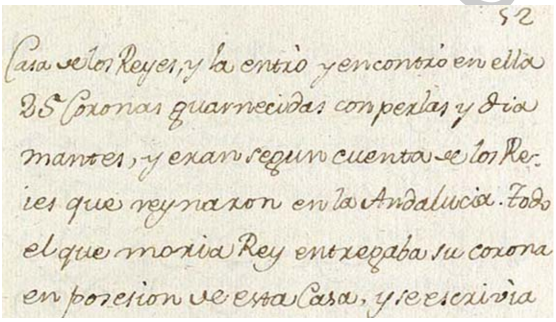

Fig. 1. Example of a handwritten document in Spanish language.

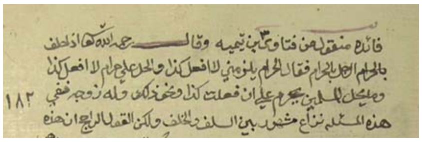

Fig. 2. Example of a handwritten document in Arabic language.

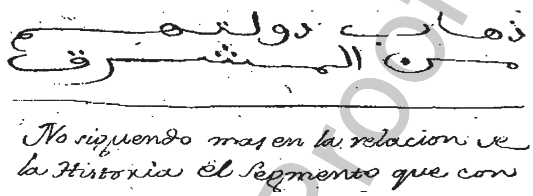

Fig. 3. Example of a handwritten document with combined languages. The first section contains handwritten text in Arabic language and the second one shows handwritten text in Spanish language.

On the one hand, related works of the TLL that are based on the Horizontal Projection Profile (HPP) use the projection of a histogram to determine the number of lines of text by peaks, see Fig. 5. The works mentioned above are more focused on finding the lines of text that in finding the points between the lines of text that could better separate those lines. This type of work first finds the local maximum values (peaks) in the horizontal projection profile and later determines an average interval between these peaks in order to find the cut-off points. One of the problems of these works is that there are several local maximum values in a single line of text which solves some works by smoothing the projection profile [10, 16] or by determining an average among the local maximum values [10]. Recently in [16], it is proposed to obtain an energy map of the document image to enhance the difference between the maximum and minimum points. In this paper we propose a method for TLL that is focused on local minimum values (gaps between lines of text) from Projected Energy Map with Alpha blending (PEM-Alpha).

On the other hand, for the SPDL problem some related works separate the text looking for a path from left to right that better separates the lines, that is to say these methods locally seek to minimize the number of crossing points in the letters. Most of the works related to the SPDL [11, 16, 17] develop an overall optimization of the path that separates text, in which the difference lies in the proposed function that should minimize such algorithm. In this paper we propose a method for SPDL using a Genetic Algorithm (GA) that optimizes the minimum number of crossing points (SPDL-GA) in handwritten letters.

This paper is organized as follows. In Section 2, we present the related work. In Section 3, we detail the different steps followed by the proposed system. The experimental results on multilingual collection are reported in Section 4 and the conclusions are drawn in Section 5.

2. Related work

The first and second part of this section briefly describes the TLL related works that work from the bottom up and from the top down, respectively. The third section describes the work related to the SPDL, along with its cost functions.

2.1. Preprocessing

In the state-of-the-art, it is emphasized that the binarization stage [18, 19], skew correction [20–22], and noise reduction [11, 23, 24], are fundamental steps for the task of *image document analysis* and, therefore, the task of TLS.

The methods proposed in [6, 11, 25, 26] performing a preprocessing step before the TLL or SPDL stage. On the other hand, in the following works [8, 13, 14, 27] it is assumed that the input is a binarized image of a document with already corrected inclination and single column text information.

The works that does not have a preprocessing stage can increase its performance when these methods are applied [16].

2.2. Bottom-up TLL approaches

The methods in this category group basic elements of the image as pixels, characters or connected components [7, 13, 24] to form line patterns [9].

These methods perform well on documents containing groups of lines of text with different lengths and inclinations in each paragraph. Figure 4 shows an example of the documents in which these approaches perform best.

These clustering-based methods cannot segment document images with lines of text that intersect vertically.

2.3. Top-down LLT approaches

These types of methods are based on learning, horizontal projection and energy map.

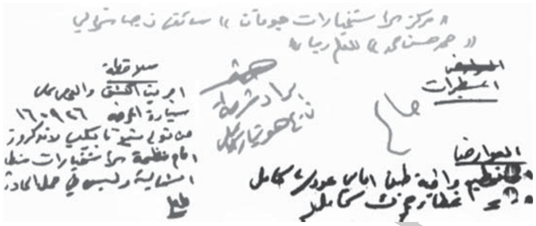

Fig. 4. Example document where the bottom-up methods have better performance [24]. This image shows each group of text with different tones.

2.3.1. Methods based on learning

Learning-based methods require a training sample for TLL and a training sample with paths for the TLS [27], in this sense learning-based methods are language-dependent.

2.3.2. Methods based on HPP

The methods based on Horizontal Projected Profile (HPP) are most commonly used to locate lines of text in images of documents printed on machines [28]. Some of these methods cannot be applied directly to handwritten text documents because they need a clear separation between neighboring text lines. Text line skew variability and touching line components also influence the performance of these methods.

Usually, these types of methods are focused on locating the peaks in order to identify the separation between each line of text. However, when applying this technique to a document with handwritten text (Fig. 5), it is impossible to find peaks with the same height and width. Thus, the HPP-based methods present a set of thresholds that have to be defined empirically for each collection of documents [10, 11, 16, 28]. In addition, the main problem of the methods in this category is that these methods are based on locating the peaks in the HPP. For example, the document in Fig. 5 only contains four text lines, but the HPP of the document detects five lines (peaks).

The works in [10, 11, 16] present a HPP-based method of the histogram to estimate the position of each handwriting line (local maximum values). However, they have problems estimating the whitespace

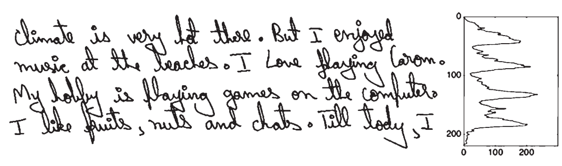

Fig. 5. Example of extraction of the horizontal projection profile of a historical handwritten document.

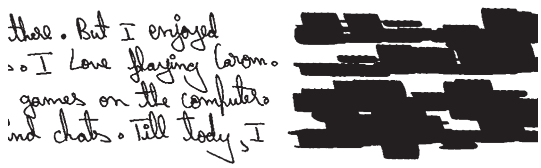

Fig. 6. Example of filling in blanks between characters using the method proposed in [29].

between neighboring lines (local minimum values) because there are overlapping and touching lines. In the work of [10] only the stage of the TLL is realized. To apply a HPP-based method it is necessary that the text is represented in horizontal lines, so it is not possible to apply directly to documents like the one in Fig. 4.

2.3.3. Methods based on energy map

An Energy Map (EM) is the process by which a document is scanned in order to eliminate the spaces between characters and words [9, 17, 29]; making larger the differences between the local maximums and minimums.

The works presented in [9, 11, 14, 17, 29] reduce the white space between each character and each word by applying energy maps based on the gradient operator (EM-gradient) [9] or a specific function (EM-F) [29], etc. To do this, the image of the document is smoothed or translated to the original image in order to generate an energy map, as shown in Fig. 6. After applying this process in some methods it is proposed to group the regions with the most information [29], in other works it is proposed to generate HPP of the energy map [9, 11].

Note that in Fig. 6 the gaps between characters and words disappear, but also the white space between neighboring text lines. It is important to keep the space between neighboring text lines in order to facilitate the search of the path that allows segmenting. Therefore, these methods have problems when separating documents with handwriting lines that intersect vertically with neighboring lines.

2.4. Methods for SPDL

Some works search for the path with the most amount of white space but perform a local search, thus they do not guarantee an optimal path [8, 13, 14]. In [11], a local search of the path is made considering as a cost function the least amount of black pixels within the path.

In the method presented in [16], an adaptation of the seam carving method to find the best path is used. Seam carving is a path of pixels connected from top to bottom in an image with one pixel in each row. Eventually, the path with the smallest overall penalty, or cost, is the desired solution. To avoid that a method deviated to a local minimum, it is necessary to use a global optimization technique as discussed in [9, 17].

3. Proposed methods

On the one hand, the hypothesis of this paper is that the accuracy of the TLL can be improved if the search is focused more on the local minimum values (white gaps between lines) of horizontal projection profiles of alpha energy maps of manuscript texts, which would reduce the many local minima that may occur even when the lines are touching or overlapping.

On the other hand, our hypothesis is that the accuracy of the SPDL can be improved if a non-linear path that crosses the smallest number of letters is minimized globally from the initial point to the final point. In this way, the TLS task can be improved.

In the first section we describe the proposed method in general. In the second section we give the details of the proposed energy map based on the alpha blending. In the third section we describe how the parameters are automatically determined for the proposed energy map. Finally, our proposed method for SPDL is described.

3.1. General steps for TLS

The input of our method is a grayscale image where the angle of inclination has been corrected, the noise has been eliminated and the background variations have been filtered. The steps to follow are:

1. Automatically determine the best parameters for the EM-Alpha of a subset of the collection (see Section 3.2.1).
2. For each image in the collection, the EM-Alpha with the parameters of step 1 (Section 3.2) is generated and, then, the HPP from the histogram of the EM-Alpha image is extracted. The result of this step is a Projected Energy Map with Alpha Blending (PEM-Alpha).
3. Remove from the PEM-Alpha the local minimum values which are lower than a threshold (in percentage) based on the average of the local maximum values.

4. Locate the local minimum values in the PEM-Alpha to determine the start and end points to draw the path.
5. Find a path between the start and end point that best split the neighboring text lines (see Section 3.3).

3.2. Proposed energy map

Unlike previous work, in this paper we propose a new energy map based on the alpha blending [30]. The goal of generating an EM with the alpha blending is to generate a HPP in which all local minimum values go down to zero. The result of this stage is a Projected Energy Map based on Alpha Blending (PEM-Alpha) where it is possible to locate each line of text.

First, the alpha blending is applied to the image to obtain an EM-Alpha. After that, this image is binarized to reduce local minimum values.

Alpha blending is a simple method for transparently overlaying two images [30], I_{BG} and I_{FG}, within a window size (w). The original image I is covered by the slice image, whose transparency is controlled by the value $\propto$ in the form:

$$\text{Alpha}\,(I, w, \alpha)^r = I_{BG}\,(u+w) + (1-\propto)\cdot I_{FG}\,(u+w)$$

where $0 \leq \propto \leq 1$, $\propto = 0.5$, u is the position on the x-axis and r is the number of times that the Alpha blending is applied.

Unlike the binary image of the document, the EM-Alpha image is a grayscale image.

High-energy regions (most black regions) correspond to center of text lines and low-energy regions correspond to the top and down text lines. Figure 7 shows an example of regions with high-energy and low-energy.

Binarization of alpha energy map (EM-Alpha) allows to remove pixels with low energy leads to minimal information loss in comparison to directly extraction of projection profile (see Fig. 8). From the extraction of the HPP, we can perform better the TLL as can be seen in Fig. 8.

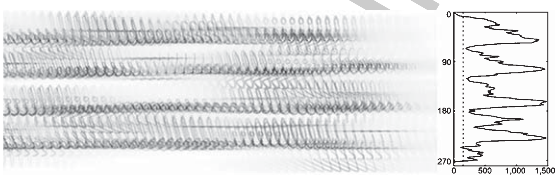

Fig. 7. Example of proposed EM-Alpha in grayscale and its HPP.

In this algorithm, it is possible to generate a HPP where the distance between the peaks and valleys increases (see Fig. 9) unlike the HPP shown in Fig. 7, leaving a larger gap between the local minimums which reach 0.

To determine the points of origin to trace the path is necessary to locate the valleys with a length greater than 1 pixel in the HPP. Figure 9 shows the beginning and the end of each valley in the HPP. In Fig. 10, the initial ($P0$) and final valley (Pf) found by PEM-Alpha method are drawn in the corresponding text.

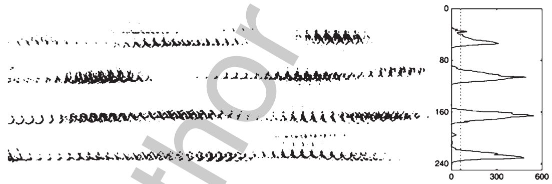

Fig. 8. Example of the HPP of a binarized EM-Alpha.

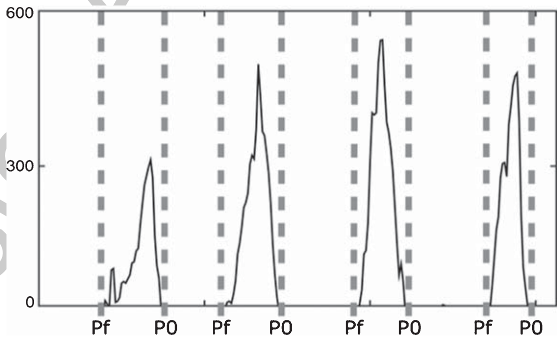

Fig. 9. Projection profile extraction from EM-alpha of English E18 document of the [9] collection.

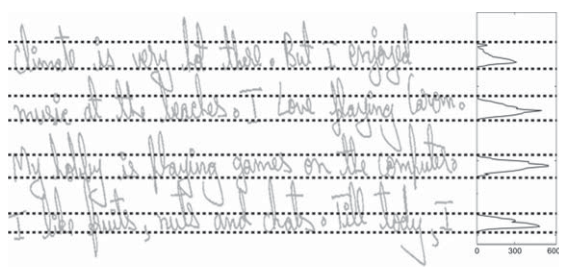

Fig. 10. Founded medial seams by finding the valleys of the HPP showed on Fig. 9. Image document English E18 of the [9] collection.

3.2.1. Finding the best parameters for PEM-Alpha

For applying the alpha blending, it is necessary to define the number of slices (r) for the medial seam computation and the window size (w) for translate the image in every slice.

The problem with these parameters is that they depend on the author's language and style of writing. For example, a very small window will not achieve the purpose of filling spaces between characters. Likewise, an unsuitable slice size will not preserve the proper information.

Given a minimum and maximum window range $[w_{min}, w_{max}]$; and a minimum and maximum slice range $[r_{min}, r_{max}]$ in a grayscale image I, where the bits 1 s correspond to the black pixels and the 0 s bits to the background, the appropriate size of r and w is the one that produces the most bits of 1 s applying the alpha blending with $\propto = 0.5$ to all combinations of r and w in $w_{min} \leq w \leq w_{max}$ and $r_{min} \leq r \leq r_{max}$. It means:

$$Find\,(I, r, w) = \max\left(\sum_{\substack{w_{min}\leq w\leq w_{max}\\ r_{min}\leq r\leq r_{max}}} Bits1\,(Bin\,(Alpha\,(I, w, \propto)^{r}))\right)$$

3.3. Search for the path that divides lines of text based on a genetic algorithm (SPDL-GA)

In this stage we propose to carry out a global search for a set of intermediate points between the initial and final points (found in the previous stage) to generate the optimal path to segment a pair of neighboring lines with handwritten text. In particular, we propose to use the GA because they have proven to be one of the best techniques to solve optimization problems.

In the first stage of a GA, a set of random solutions is generated (*population generation step*) and evaluated with a metric that quantifies the quality of the solutions (*fitness function step*) in order to minimize or maximize the aptitude of each individual [31].

The solution to a problem is not absolute, on the contrary, there is a possible set of solutions where some of them are better than others. In a next step, a selection of the best solutions (*parent selection step*) is extracted, so that the GA proposes a new population by mixing (*crossing step*) some fragments of the genes of the set of best solutions in order to generate better solutions (principle of evolution).

After several iterations mixing the genes of the best individuals, repeated solutions are generated. To solve this problem the GA applies a small variation (*mutation step*) to the genes of each individual of the new population to explore new solutions [31].

At the end of the mutation stage the new population is evaluated and the process is repeated until a satisfactory solution is found or until an arbitrary criterion is achieved to stop the search (*stop condition*).

3.3.1. Preprocessing step

Before encoding the individual it is necessary to know the relative position of each initial and final point of the path.

3.3.2. Chromosome encoding

A GA needs to encode each solution (chromosome) using a canonical representation. For TLS it is proposed to represent the genes of chromosome (C) with a vector of size n (number of intermediate points between the initial and final point of the path) with values in base k. The base of representation of the genes (k-base) is determined as $k = \left|P_f - P_0\right| + 1$.

3.3.3. Initial population step

After knowing the coding of individuals, the first generation is created randomly where each gene can take a value between 0 and k, *i.e.* ($C_{i=1\ldots n} = Random\,[0, k]$).

3.3.4. Fitness function step

One of the most important elements of genetic algorithms is the fitness function. For this problem the fitness function is considered as a minimization function in which it is necessary to look for a path that crosses the smallest number of black pixels in a grayscale image but which is the shortest path so that it gets as close as possible to a straight line between the initial and final point as the human normally does. Thus, if there is a clear division between two text lines it is possible to find the straight line from the start point to the end point. Therefore, the fitness function counts the number all pixels with $Bits1 < MostCommon\,Intensity$ found in the generated path plus the sum of the value of the genes C_i.

$$FA\,(C_n) = |C_1| + \sum_{i=2}^{n}\left(Bits1\left(P_{C_{i-1}}, P_{C_i}\right) + |C_i|\right)$$

3.3.5. Parent selection step

In this stage each chromosome has associated a value of aptitude that allows to select the best chromosomes. The principle of evolution states that it is natural to improve the fitness of individuals when two good solutions are crossed. In this stage of the GA it is foreseen to apply the selection operator by tournament.

3.3.6. Crossover step

A new chromosome is created from the pair of parents selected by the classic operator of uniform cross at n-points.

3.3.7. Mutation step

According to the theory of evolution, mutation occurs with a very low probability (about 0.1%), however, it is important that it happens to ensure the evolution. For this step, we mutate randomly (with a very low probability per gene) the genes according to the k-base of the chromosome coding, i.e.:

$$Mutation\,(C_i) = Random\,[0, k]\,.$$

4. Experimental evaluation

The evaluation section is divided into three stages, the first stage describes the collection of documents used, and the second stage evaluates the performance of methods to find the number of lines of each document. The third stage evaluates the performance of methods to generate the best path for handwritten text line segmentation.

4.1. Datasets

After performing an exhaustive search we find few standard sets of public document data to segment lines of handwriting, most of works use private data to perform the experimentation stage [8, 11, 16, 27]. Another problem is that some related research only uses an unspecified subset of documents from a public data [7, 16, 17]. Due to this problem, a collection of images of historical documents was created in four languages (Arabic, Chinese, English, and Spanish) and the combination of Arabic-Spanish, it is available on [9]. The experimentation in this paper is performed using the whole collection of [9], which consists of 315 historical manuscripts with 216 Arabic pages (3972 lines), 20 English pages (187 lines), 15 Spanish pages (197 lines) and 10 Arabic-Spanish pages (124 lines).

4.1.1. Preprocessing

In order to make the results of our proposed methods comparable to [10, 16], the same preprocessing is applied to the entire collection. In this case, the skew correction of all the documents in the collection is achieved using the Radon transformation [21].

4.2. Evaluation methodology

In order to better appreciate the relevance of our proposed TLS methods, the evaluation of the TLL and SPDL stages has been divided.

For the TLL stage an evaluation metric based on the presented in [10] is proposed, but the same metric is not used because it only evaluates the number of bad separator identifications and the number of false positives is limited to one.

Similar to the metric of [10], we evaluate the performance of TLL methods according to number of handwritten text lines correctly identified (true positive) minus the number of incorrect separators (false positives). A separator is correct if it is located between two adjacent or neighboring lines. With the proposed evaluation metric two types of error are considered. The first type of error occurs when two neighboring lines of text are identified as a single line. The second type of error occurs when there is more than one separator that segments a single line of text two or more. It is important to consider both aspects because this affects the MTR task.

There are many evaluation schemes for the SPDL However, many recent evaluation methods are based on MathScore. MathScore was introduced by Yanikoglu [32] and it is defined as the percentage of the foreground pixels of G_u covered by R_u minus the percentage of the foreground pixels of R_u outside of G_v.

Let G_u be set of all points of the i ground truth region, R_v - set of all points of the j result region, $T\,(s)$ is a function that counts the elements of set s. $MatchScore\,(u, v)$ represents the matching results of i ground truth region and j result region as follows:

$$MatchScore\,(u, v) = \frac{T\,(G_u \cap R_v)}{T\,(G_u \cup R_v)}$$

It is worth mentioning that this score is used to assess the performance of proposed methods in ICDAR 2007, ICDAR 2014 and ICFHR 2010 Handwriting Segmentation Contest.

There are few multilanguage systems for the TLS, so the comparison of the results is carried out with the original systems of Ptak [10] and Arvanitopoulos [16].

The Arvanitopoulos system [16] performs the tasks of TLL and SPDL for historical documents, being better than the system proposed in [16]. However, the Ptak system [10] is only for the TLL task, so it is not included in the complete task. Since previous systems require a set of parameters a priori, exhaustive tests were performed for the dataset used with all possible combinations. It is important to address that the following experimentation is the result of processing all the documents in the collection of [9].

Tables 1 and 2 show, according to the language, best parameters obtained in the experimentation of the systems of Ptak [10] and Arvanitopoulos [16], respectively.

For estimating the parameters for the TLL, in the first step, 5 documents are randomly selected from the whole collection. In the second step, for each document, the value of r and w is chosen (between [1,100] and [5, 500], respectively) that produces greater entropy in the energy map with alpha blending. The above step produces five estimations of the minimum and maximum parameters for r and w. In this case, for this collection we obtain: $w_{min} = 7,\ w_{max} = 30,\ r_{min} = 10$ and $r_{max} = 50$ s.

Using the above range for w between [7, 30] and for r between [10, 50] the maximum entropy for the Arabic, English, Spanish and Arabic-Spanish collection are obtained: $r = 30$ and $w = 7$, only for the Chinese the window change to $w = 15$. In this case, the window size for Chinese is higher because the white space between the words also is larger (see Fig. 14).

After performing some experiments, the percent threshold to eliminate the non-relevant local minimums is 10% for all collections.

Table 3 shows the results achieved with our method for the five sub-collections and the comparison with the systems of Arvanitopoulos [16] and Ptak [10]. Also, in Table 3, it is included an average per system.

Table 1
Best configuration found for the Arvanitopoulos system [16]

Language	smooth	slides	sigma	offset
Arabic	0.0001	6	2	1
Chinese	0.00007	7	2	4
English	0.001	3	2	1
Spanish	0.001	3	2	1
Arabic-Spanish	0.00001	3	10	2

Table 2
Best configuration found for the Ptak system [10]

Language	Threshold	slides
Arabic	0.98	12
Chinese	0.997	12
English	0.99	11
Spanish	0.99	8
Arabic-Spanish	0.99	10

Table 3
Results for the identification of lines in the whole collection

Method	Arvanitopoulos's System [16]	Ptak's System [10]	Proposed PEM-Alpha
Arabic	94.45%	96.33%	98.98%
Chinese	97.18%	95.77%	98.59%
English	94.13%	94.84%	99.20%
Spanish	98.30%	92.38%	98.72%
Combined	75.35%	87.09%	97.18%
Average	91.88%	93.28%	98.53%

Table 4
Comparison of accuracy to our method to [16]

Language	Arvanitopoulos's system [16]	Proposed PEM-Alpha+BRL-GA
Arabic	93.68%	97.83%
Chinese	99.35%	99.68%
English	95.15%	98.34%
Spanish	97.30%	98.72%
Combined	82.90%	95.45%
Average	93.67%	98.00%

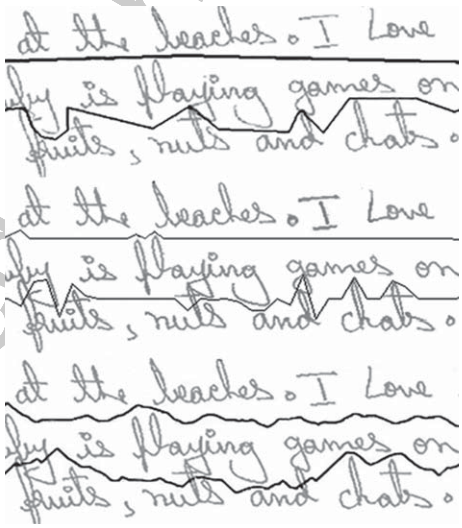

Fig. 11. Visual comparison of evaluated methods on English language.

Analyzing the results of Table 3, it is concluded that PEM-Alpha has the best accuracy compared to other systems. Note that other methods are more affected with the mixed collection.

4.3. Experimentation for the full task of TLS

In this section we compare our proposed PEM-Alpha+SPDL-GA methods to the Arvanitopoulos system [16]. In order to adjust the parameters of the genetic algorithm, the following parameters are

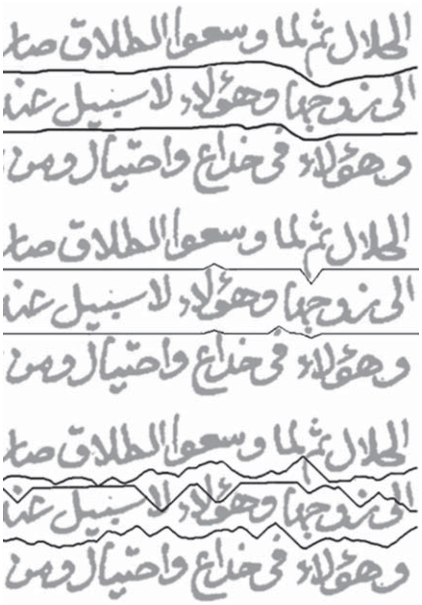

Fig. 12. Visual comparison of evaluated methods on Arabic language.

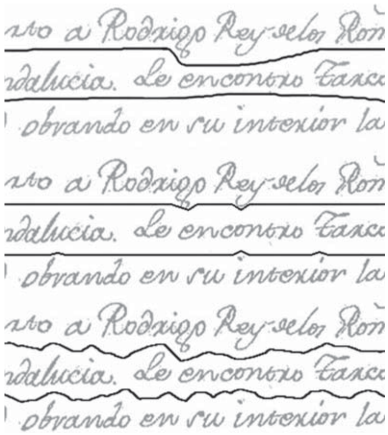

Fig. 13. Visual comparison of evaluated methods on Spanish language.

obtained for the whole experiment: number of chromosomes of 10, number of genes per chromosome of $n = 100$, size of tournament of 3, crossing points 2, probability of mutation of 0.2% and stop criterion of 70 generations.

Table 4 shows the results achieved by our proposed methods. As it can see, our method surpasses in all the collections to the system of Arvanitopoulos [16]. Also, in Table 4 is included an average per system.

From Figs. 11–15 are shown in the upper part an example of the separation made by the human, in the middle part by our method and in the lower part by the Arvanitopoulos system; for English, Arabic, Spanish, Chinese and Arabic-Spanish collections, respectively.

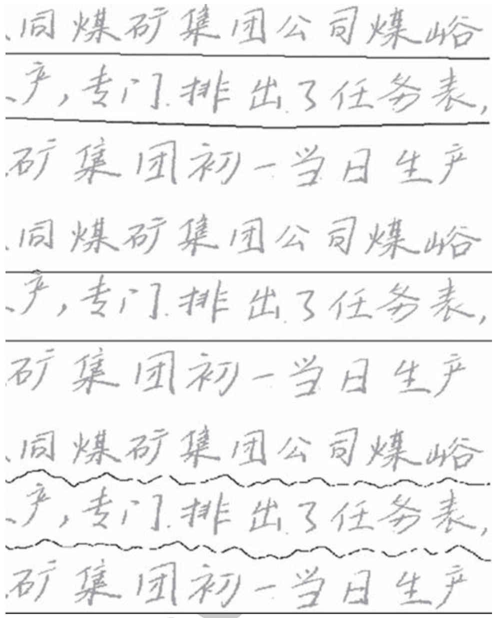

Fig. 14. Visual comparison of evaluated methods on Chinese language.

Fig. 15. Visual comparison of evaluated methods on combined languages, Arabic at the top and Spanish at the bottom.

5. Conclusions and future work

The analysis of images of handwritten documents is important to access the information they contain.

It is worth remembering that the most valuable treasure of humanity is the knowledge that have generated diverse cultures, which is already digitized in historical documents. This paper proposes the PEM-Alpha and SPDL-GA methods for the multi-language text line segmentation, which is a subtask of the recognition of handwritten texts. In particular, for the task of locating text lines, the PEM-Alpha method is proposed which can automatically calculate few parameters required according to the type of language. Using the same standard data set proposed in [9] for four languages and mixed languages, our PEM-Alpha method outperforms other systems in all sub-collections. In this sense, it is concluded that for TLL it is better to search for local minimum from the horizontal projection profile of an energy map based on the alpha blending.

For the problem of finding the path that best divides the neighboring text lines, the SPDL-GA method is proposed, which allows finding a non-linear path that minimizes the cut-off points between the letters and the distance between the initial and final points. According to the experimentation, our method surpasses the method recently proposed in [16]. As can be seen, at first glance, in Fig. 11, our method has more similarity to the path made by the human.

It is necessary to emphasize that in both proposed methods it is not necessary to adjust the parameters of the proposed method for each sub-collection of documents, therefore this is an advance in comparison to the current works. The methods proposed in this work surpassed other systems in documents with Spanish, Arabic, English, Chinese and documents containing two languages (Arabic and Spanish), therefore the methods presented here are unsupervised multi-language methods for the handwritten text line segmentation task.

In the future it would be interesting to test documents with greater complexity where text lines with more complex splices are available.

Acknowledgments

This work was partially support of Mexican Government CONACYT Thematic Networks program (Language Technologies Thematic Network project 281795). We also thank UAEMex for their assistance.

References

[1] J.A. Sánchez et al., TranScriptorium: A european project on handwritten text recognition, in *Proceedings of the 2013 ACM symposium on Document engineering*, Florence, Italy, 2013, pp. 227–228.

[2] T. Causer and V. Wallace, Building a Volunteer Community: Results and Findings from Transcribe Bentham, *Digit Humanit Q* **6**(2) 2012.

[3] S. Medina Morán, ¿Un error en la piedra de rosetta?, *REB. Revista de educación bioquímica* **30**(1) (2011), 21–27.

[4] T. Steinherz, E. Rivlin and N. Intrator, Offline cursive script word recognition – a survey, *Int J Doc Anal Recognit* **2**(2) (1999), 90–110.

[5] A. Khandelwal, P. Choudhury, R. Sarkar, S. Basu, M. Nasipuri and N. Das, Text Line Segmentation for Unconstrained Handwritten Document Images Using Neighborhood Connected Component Analysis, in *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, New Delhi, India, 2009, pp. 369–374.

[6] U.V. Marti and H. Bunke, Text line segmentation and word recognition in a system for general writer independent handwriting recognition, in *Proceedings of Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 159–163.

[7] D. Valy, M. Verleysen and K. Sok, Line Segmentation Approach for Ancient Palm Leaf Manuscripts Using Competitive Learning Algorithm, *2016 15th Int Conf Front Handwrit Recognit ICFHR*, 2016, pp. 108–113.

[8] G. Peng, P. Yu, H. Li and L. He, Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai, *2016 Int Conf Audio Lang Image Process ICALIP*, 2016, pp. 336–340.

[9] R. Saabni, A. Asi and J. El-Sana, Text line extraction for historical document images, *Pattern Recognit Lett* **35** (2014), 23–33.

[10] R. Ptak, B. Zygadlo and O. Unold, Projection Based Text Line Segmentation with a Variable Threshold, *Int J Appl Math Comput Sci* **27**(1) (2017), 195–206.

[11] M.W.A. Kesiman, J.-C. Burie and J.-M. Ogier, A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript, *2016 15th Int Conf Front Handwrit Recognit ICFHR*, 2016, pp. 325–330.

[12] P. Jindal and B. Jindal, Line and Word Segmentation of handwritten text documents written in Gurmukhi Script using mid point detection technique, in *2015 2nd International Conference on Recent Advances in Engineering Computational Sciences (RAECS)*, 2015, pp. 1-6.

[13] V.K. Koppula and A. Negi, Segmentation of closely set and touching lines in handwritten document images using fringe maps, *Int Conf Converg Technol-2014*, 2014, pp. 1–6.

[14] A. Nicolaou and B. Gatos, Handwritten Text Line Segmentation by Shredding Text into its Lines, in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 626–630.

[15] Y.-H. Tseng and H.-J. Lee, Recognition-based Handwritten Chinese Character Segmentation Using a Probabilistic Viterbi Algorithm, *Pattern Recogn Lett* **20**(8) (1999), 791–806.

[16] N. Arvanitopoulos and S. Süsstrunk, Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts, in *2014 14th International Conference on Frontiers in Handwriting Recognition*, 2014, pp. 726–731.

[17] M. Liwicki, E. Indermuhle and H. Bunke, On-Line Handwritten Text Line Detection Using Dynamic Programming, *Ninth Int Conf Doc Anal Recognit ICDAR 2007*, 2007, pp. 447–451.

[18] I. Pratikakis, K. Zagoris, G. Barlas and B. Gatos, ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016), in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2016, pp. 619–623.

[19] Y. Wu, P. Natarajan, S. Rawls and W. AbdAlmageed, Learning document image binarization from data, in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3763–3767.

[20] A. Papandreou, B. Gatos, G. Louloudis and N. Stamatopoulos, ICDAR 2013 Document Image Skew Estimation Contest (DISEC 2013), in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1444–1448.

[21] R. Kapoor, D. Bagai and T.S. Kamal, A new algorithm for skew detection and correction, *Pattern Recognit Lett* **25**(11) (2004), 1215–1229.

[22] S.C. Hinds, J.L. Fisher and D.P. D'Amato, A document skew detection method using run-length encoding and the Hough transform, in *[1990] Proceedings. 10th International Conference on Pattern Recognition* **1** (1990), 464–468.

[23] A. Prachanucroa and S. Phongsuphap, Marginal noise removal for scanned document images by projection profile based method, in *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2013, pp. 17–20.

[24] Y. Li, Y. Zheng, D. Doermann and S. Jaeger, Script-Independent Text Line Segmentation in Freestyle Handwritten Documents, *IEEE Trans. Pattern Anal Mach Intell* **30**(8) (2008), 1313–1329.

[25] Z. Shi and V. Govindaraju, Line separation for complex document images using fuzzy runlength, in *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings*, 2004, pp. 306–312.

[26] L. Likforman-Sulem, A. Zahour and B. Taconet, Text line segmentation of historical documents: A survey, *Int J Doc Anal Recognit IJDAR* **9**(2–4) (2006), 123–138.

[27] Q.N. Vo and G. Lee, Dense prediction for text line segmentation in handwritten document images, in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3264–3268.

[28] J. Ha, R.M. Haralick and I.T. Phillips, Document page decomposition by the bounding-box project, in *Proceedings of 3rd International Conference on Document Analysis and Recognition* **2** (1995), 1119–1122.

[29] X. Du, W. Pan and T.D. Bui, Text line segmentation in handwritten documents using Mumford-Shah model, *Pattern Recognit* **42**(12) (2009), 3136–3145.

[30] W. Burger and M.J. Burge, Digital Image Processing: An Algorithmic Introduction Using Java. Springer-Verlag London, 2008.

[31] K.L. Du and M.N.S. Swamy, Search and Optimization by Metaheuristics: Techniques and Algorithms Inspired by Nature, Springer International Publishing (2016).

[32] B.A. Yanikoglu and L. Vincent, Pink Panther: A Complete Environment For Ground-Truthing And Benchmarking Document Page Segmentation, *Pattern Recognit* **31**(9) (1998), 1191–1204.