Handwritten Line Detection via an EM algorithm

Francisco Cruz, Oriol Ramos Terrades
Universitat Autònoma de Barcelona
Computer Vision Center, Barcelona, Spain
{fcruz,oriolrt}@cvc.uab.es

***Abstract*—In this paper we present a handwritten line segmentation method devised to work on documents composed of several paragraphs with multiple line orientations. The method is based on a variation of the EM algorithm for the estimation of a set of regression lines between the connected components that compose the image. We evaluated our method on the ICDAR2009 handwriting segmentation contest dataset with promising results that overcome most of the presented methods. In addition, we prove the usability of the presented method by performing line segmentation on the George Washington database obtaining encouraging results.**

I. Introduction

Text line segmentation is an old problem of great interest within the Document Analysis community. The first proposed methods were focused on printed text line segmentation and big advances on printed documents has been done since then. Indeed, existing methods works reasonably well if document layout is not so much complex and consequently, many research efforts has moved on solving layout analysis problems. However, for handwritten text documents, text line segmentation has become a required step for most of handwritten text recognition engines, since the most performing methods work at line level and therefore, line segmentation methods have still to be improved.

There are different sources of difficulties particular from handwritten text. Some, we find them from the writing text style. For instance, if the interline space is small, it is common to find touching characters between lines. Line orientation can slightly change within the same text region, probably due to the writer's body position when the document was written. Moreover, character size variability is higher than in printed documents. Other difficulties are linked to text location. For instance, when text is located in the document margins, either as an extension of previous lines or as independent annotations, text line segmentation methods fail more than when text is located in the central part of a document. In this paper we present a line segmentation method devised to deal with these difficulties.

To do so, our starting working hypothesis comes from the following assumptions and final observation. Let us assume for a moment that we have already segmented the text lines of a document. Then, we can also assume that text pixels of a fixed text line are distributed according a regression line, which means in particular that their residuals are distributed according to a Normal distribution, and therefore we can represent each text line by regression lines equations. Our working hypothesis is that we will be to extract text lines if we are able to extract such regression lines.

Two main concerns directly arise from these initial assumptions. On the one hand, it is not clear that Normal assumption is generally satisfied for segmented text lines. More worrying could seem that line regressions can be computed on segmented text lines and actually it is the problem we are tackling in this paper.

We have overcame this last concern by proposing a probabilistic version of line regression equations where probabilities are updated following and Expectation-Maximization (EM) algorithm scheme [11] . The EM algorithm has largely been used for estimating pdf models with hidden variables. Two classical examples of such models are Gaussian mixture models (GMM) and Hidden Markov models (HMM) but it can be applied to a wider range of family distributions. Moreover, experiments will show that violation of Normal distribution assumption has not a relevant impact on the final results.

The rest of the paper is organized as follows: In section II we show a review of other works that have tackled the problem of line segmentation on similar text configurations. Then, section III describe the details of the different parts of our method. Section IV details the performed experiments and the obtained results. Finally, we show the conclusions that we have reached and the future research lines arising from this work in section V.

II. Related work

Many different methods have been proposed to deal with the different challenges present in a line segmentation problem [1]. Probably one of the most common methods are the projection profile based approaches [2]. Here, a set of pixels is projected on a fixed axis to estimate the number and location of the lines according to the peaks and valleys observed in the resultant histogram. This technique have proved to obtain good results on typewritten documents, however, if we deal with handwritten documents, issues as the skew of the lines or touching characters between lines can be a problem. Several variations of this technique have been proposed to deal with some of these situations. In [3], the authors present a method based on adaptive local projection profiles to deal with multi-skewed documents. Another example is the work of Marinai and Nesi in [4], where multiple projections along different axes are computed to identify symbols from musical sheets. The location of the lines within the document can be also a problem for these methods. In [5], the authors apply an image meshing to progressively determine the location of the lines, and then estimate the orientation by means of the Wigner-Ville distribution on the projection histogram profile.

Apart from this family of methods, different approaches are required to deal with other possible problems. It is common to find documents that include curved lines, overlapping words or multiple skews, and a well-known technique that helps in these cases is the Hough transform. In [6], the authors use this technique to compute the global orientation on handwritten documents as a previous step to the final line extraction. Another example is the work of G. Louloudis *et al.* in [7], where they apply a block-based Hough transform to detect the lines in documents with multiple orientations and complex writing styles with remarkable results.

Many different approaches have been also developed to tackle the line segmentation task. In [8], fuzzy runlenghts are used to detect handwritten annotations in complex layouts. In [9], a script-independent method based in active contours is applied with good results for the detection of curved and skewed handwriting lines. In addition, another works have focused on the influence of line segmentation in other posterior tasks. In [10] the authors demonstrate that a good result on the line segmentation process significantly affects to the results of a posterior word segmentation task.

III. EM FOR REGRESSION LINES

For the kind of given problem purposes, we have defined a document as a set of paragraphs, which in turn, are also composed of a mixture of text lines, all of them sharing similar orientation and similar interlinear space. According to this definition, first we find out the document paragraphs and then, for each paragraph, we apply the proposed EM algorithm described along this section. In this way, we are able to deal with multiple text line orientations in multiple paragraphs regardless their position in the original document. Therefore, and without lose of generality, we have considered in the remainder of this section that a document is only composed by one single paragraph.

The proposed EM for regression lines is based on the well-known EM algorithm but adapted for finding text lines. The rational of our approach is the following. For a given text region, we extract connected components and compute their gravity center. Then, if we knew which were the text lines of a given document. Then, for each one we could estimate regression lines of the gravity centers of connected components. Indeed, if text lines did not show warping effects, such regression lines should be a good estimation of text line positions. The main problem of this approach is however, that we do not know which are the text lines. Nevertheless, the EM for regression line algorithm introduced in this paper will help us to find out them. Once we have estimated such regression lines, we have to segmented them.

Therefore, the proposed method we can summarized by the following three steps algorithm:

1) Initial parameters model estimation: paragraphs region, number of lines and regression line parameters
2) EM iteration: fit initial model to the real regression lines of gravity centers
3) Text line segmentation: Assign to each pixel the most probable text line

Before providing more details on these three steps in the remainder of this section, we introduce in what follows the theoretical model used and recall the lineal regression equations as well.

A. Line regression model

The proposed method mix a subset of random variables (x, y) distributed on a triangulated irregular grid, capturing image content information and a subset of categorical random variables (l, t) labeling document entities according lines numbers (l), text blocks (t) and spatial relationships between l, t variables. In this paper we investigate a simple conditional random field (CRF) model where dependencies between hidden variables l, t are not considered. Therefore, the model is given by the following expression:

$$p(l, t|x, y) = \prod_n \frac{p(x_n, y_n|l_n, t_n)p(l_n, t_n)}{p(x_n, y_n)} \tag{1}$$

where probabilities $p_n(x_n, y_n|l_n, t_n)$ are modeled by Normal distributions linked to regression lines. More specifically, regression lines techniques are used when, given a cloud of image coordinates, it is assumed that they are linearly correlated but corrupted by white noise. White noise distributions follow a Normal law, centered in the origin and variance $\sigma^2_{l,t}$. We have modeled these Normal distributions by the conditional probabilities $p_n(x_n, y_n|l_n, t_n)$:

$$p(x, y|l, t) \propto \exp\left\{\frac{(a_{l,t}x + b_{l,t} - y)^2}{2\sigma^2_{l,t}}\right\} \tag{2}$$

where line parameters $a_{l,t}$ and $b_{l,t}$ and noise variance $\sigma^2_{l,t}$ are given by the usual linear regression equations:

$$\begin{aligned} a_{l,t} &= \frac{\sum_n (x_n - \overline{x})(y_n - \overline{y})}{\sum_n (x_n - \overline{x})^2} \\ b_{l,t} &= \frac{1}{N}\sum_n (y_n - a_{l,t}x_n) \\ \sigma^2_{l,t} &= \frac{1}{N-2}\sum_n (a_{l,t}x_n + b_{l,t} - y_n)^2 \end{aligned} \tag{3}$$

The main problem of the equations above is that we can applied them only on the cloud of points (x_n, y_n) belonging to a fixed line l, t while the problem statement actually is, the inverse one. Given a set of points group them according whether their belong to the same text line, or not.

We have tackled this problem using the EM algorithm for estimated a mixture of Normal distributions, where the number of components will be linked to the number of text lines and the variance of each component to the $\sigma^2_{l,t}$ values defined above. More specifically, we will estimate a Gaussian mixture model (GMM) with as much components as number of text lines.

B. Model initialization

The first stage of our method relies in the analysis of the structural features of the text. The aim of this stage is to estimate the number of paragraphs within the document, their orientation, the number of lines of each of them and a first estimation of regression line parameters for each line of each paragraph as well.

For this purpose, we have followed a sequential heuristic based on the centroids of connected components and the graph associated to the Delaunay triangulation generated from those centroids. At this point, if there were several paragraphs located in different areas of the document, some of the triangles must show longer edges than the majority of the rest. Taking this into account, we have computed the perimeter of each triangle and we have removed the ones whose perimeter is greater that a fixed threshold. The result of this operation can be seen as disconnecting the Delaunay graph. Thus, we have defined a paragraph as a connected component in the resulting graph after removing big triangles.

After this process, we have to estimate the orientation of each detected paragraphs. To do so, we have taken the set of vectors defined by the triangle edges of the Delaunay triangulation. Since each triangle edge is defined by two directional vectors, we have considered the ones with a positive direction, e.g., vectors from left to right, and we have normalize them. Thus, the orientation for a given paragraph is computed as the mean vector of the Delaunay triangulation edges.

Next, we have to estimate the number of text lines of each paragraph. We have estimated them as de quotient between paragraph height and the estimated interlinear space. To estimate the interlinear space, first we have projected triangle edges to the vector orthogonal to the paragraph orientation, as some triangle edges will mostly oriented like the paragraph orientation, we have considered only edges over a fixed angle with respect to it. the interlinear space is the mean value of the projection values.

Finally, we have computed the initial configuration of the regression lines by distributing equitably the estimated number of lines along the paragraph. Variance $\sigma_{l,t}$ is computed according the Eq. (3) but assigning each connected component to the nearest regression line. Note that initially all the lines have the same orientation and interline space, will be during the EM stage that we will be able to fit each line according to the connected components distribution.

C. EM regression lines estimation

This iterative algorithm starts with the rough initial estimation of model parameters described above. Each iteration consists of two steps, namely *Expectation* (E) and *Maximization* (M) steps. During the E-step, we compute the expectation of the hidden variables (l, t) according to the probability model parameters estimated in the previous iteration. Then and based on that, new model parameters are computed in the M-step to be used in the next iteration. It was proven that the EM algorithm converges to a local maximum of the complete likelihood function.

The EM algorithm version for regression lines estimation is reduced to compute a Gaussian mixture model (GMM). Each mixture component is therefore, the residue of a regression line (l, t). In other words, if we define the residue of a regression line as the difference between the actual value y_n and the given by the regression line: $\bar{y}_n = a_{l,t}x_n + b_{l,t}$, then $r_{l,t} = y_n - \bar{y}_n$ follows a Normal distribution centered at $b_{l,t}$ which is estimated by a GMM. Therefore, the usual updating GMM equations provide the equivalent equations to estimate regression parameters $b_{l,t}$ (mean of a Gaussian component) and variance $\sigma_{l,t}$:

$$\begin{aligned} p_{new}(t,l) = \alpha^{new}_{l,t} &= \frac{1}{N}\sum_n p_{old}(t,l|x_n,y_n) \\ b^{new}_{l,t} &= \frac{\sum_n (y_n - a^{new}_{l,t}x_n)p_{old}(l,t|x_n,y_n)}{\sum_n p_{old}(l,t|x_n,y_n)} \\ \sigma^{2,new}_{l,t} &= \frac{\sum_n (a^{new}_{l,t}x_n + b^{new}_{l,t} - y_n)^2 p_{old}(l,t|x_n,y_n)}{\sum_n p_{old}(l,t|x_n,y_n)} \end{aligned} \tag{4}$$

The equation for updating parameter $a^{new}_{l,t}$ is obtained by maximizing the log-likelihood estimator (MLE). Straightforward manipulation gives the expression sought:

$$a^{new}_{l,t} = \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})p_{old}(l,t|x_n,y_n)}{\sum_n (x_n - \bar{x})^2 p_{old}(l,t|x_n,y_n)} \tag{5}$$

It can be observed the similarity between the obtained updates and the classical regression equations in Eq. (3). Indeed, these equations generalized in the degenerate case, where $p(l, t|x, y)$ is a Dirac distribution. Finally, $p(x_n, y_n)$ is obtained by adding conditional probabilities: $p(x, y) = \sum_{t,l} p(x, y|l, t)p(l, t)$.

D. Pixel labeling

Once we have obtained an estimation of regression lines for each text line we have to segmented them. The proposed heuristic in section III-B provides an over estimation of the number of text lines. We saw as this over estimation in the number of text lines make easier the convergence of the GMM linked to the regression lines. Nevertheless, for text line segmentation we have to remove those regression lines whose actually does not correspond to any text line. A pruning criterion based on marginal probabilities $p_{new}(t, l)$ has proven to be enough.

In addition, text lines containing letters having ascenders and descents, such as for instance *p, b, l,...*, may touch other text lines and consequently, giving connected components belonging to two lines at the same time, if the interlinear space is small. Thus, we can not assign to the whole connected component the centroid label and we have to label each pixel.

On the contrary, we have followed the same idea we have used to compute $\sigma_{t,l}$ when we have initialized our model. We have labeled each pixel, in the connected component, according to the nearest regression line. Then, using the last parameters $a_{l,t}$, $b_{l,t}$, $p(t, l)$ obtained from the EM algorithm and the variance computed at pixel level, we have computed the probability $p(t, l|x, y)$ according the the model given in Eq. (2) and the Bayes formula. Finally, we have assigned to each pixel the line maximizing $p(t, l|x, y)$:

$$(\hat{l}, \hat{t}) = \arg\max_{l,t} \frac{p(x, y|l, t)p(l, t)}{p(x, y)} \tag{6}$$

IV. Experiments

To test the performance of our method we have carried out experiments on two reference datasets in line segmentation. On the one hand, we wanted to compare our method with other reference works, so that we have used the ICDAR2009 Handwriting Segmentation Contest dataset [12]. This dataset is composed of 200 handwriting documents written by a set of different authors in several languages (English, French, German, and Greek), which increase the variability of writing styles and characters used. The documents include only textual regions, so that is not expected to find any graphical elements.

On the other hand, we wanted also to prove the usability of our method in other type of documents. For this purpose we have used the documents from the George Washington database [13]. This database is composed of 20 color images from the George Washington Papers at the Library of Congress dated from the 18th century. The documents are written in English language showing a longhand script. This database adds a set of different challenges with respect to the previous one due to the old script style, overlapping lines and a more complex layout. In addition, as there is not a defined ground truth for the task of line segmentation on this database, we have manually created our own specially for this task. For this reason, it is not possible to compare our results with any other methods, so that we present the results as an indicator of the flexibility of our method to work on other type of documents.

We have performed a single experiment consisting of the line segmentation on both datasets. We have defined a set of stop criterion for the execution of the EM algorithm. In the one hand, we have empirically fixed the maximum number of iterations of the EM algorithm to 200. We have set this value to ensure a proper adjustment of the regression lines, since in the majority of the cases the convergence was reached under this value. The convergence criterion was established according to the Kullback-Leibler divergence between the probability distributions computed on two consecutive iterations. We have considered a convergence value of $\epsilon \ll 0$.

Next we show the results obtained in the experiments on both datasets. The scores have been computed using the same evaluation software developed for the ICDAR2009 Handwriting Segmentation contest, so we are able to directly compare our method with the rest of contestants. The results are shown in terms of the following metrics: Detected lines (M), one to one matches (o2o), Detection Rate (DR%), Recognition Accuracy (RA%), and F-Measure (FM%).

	M	o2o	DR (%)	RA (%)	FM (%)
CUBS	4036	4016	99.55	99.50	99.53
ILSP-LWSeg-09	4043	4000	99.16	98.94	99.05
PAIS	4031	3973	98.49	98.56	98.52
CMM	4044	3975	98.54	98.29	98.42
CASIA-MSTSeg	4049	3867	95.86	95.51	95.68
Proposed	**4061**	**3858**	**95.60**	**95.00**	**95.20**
PortoUniv	4028	3811	94.47	94.61	94.54
PPSL	4084	3792	94.00	92.85	93.42
LRDE	4423	3901	96.70	88.20	92.25
Jadavpur Univ	4075	3541	87.78	86.90	87.34
ETS	4033	3496	86.66	86.68	86.67
AegeanUniv	4054	3130	77.59	77.21	77.40
REGIM	4563	1629	40.38	35.70	37.20

TABLE I. Results compared with the ICDAR2009 Handwriting Segmentation Contest [12]

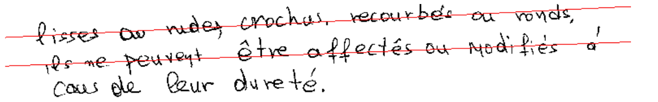
lisses ou rides, crochus, recourbés ou ronds,
ils ne peuvent être affectés ou modifiés à
caus de leur dureté.

Fig. 1. Example of excluded connected components from short lines at the end of a page.

Despite of overcoming most of the works presented in the ICDAR2009 Handwriting Segmentation contest, obtained results on ICDAR2009 dataset shows there is still place for improvement comparing to reported results, see Table I. However, we have identified some common error cases that we will correct when we add more document information by means of local features to the model proposed in section III.

We have grouped the observed error cases in three main groups. The first one comes from short text lines, usually located at the end of the document. More specifically, as the number of connected components composing short text lines is lower than the number of connected components of other text lines, final probability $p(l, t)$ is significantly smaller that the corresponding probability for the other regression lines. Consequently, these components are excluded from the line estimation process, and the final regression line is purged. An example of this first type of error is shown in Figure 1, where we can see as in the resultant regression lines we have lost the components belonging to the last line. This error correspond almost to the 40% of the reported errors, so the method performance will be significantly improve when this type of error will definitely fixed.

The second type of error is related to an over estimation of the number of lines and it represents around the 50% of the reported errors. As a result of this, we have observed two possible situations. The first one refers when an extra line is located between two regression lines correctly estimating text lines. In this situation the overall result is slightly affected, because the reminder lines have been correctly matched.

The second situation causing this type of error refers when an extra line is located crossing several regression lines (see Figure 2 for an example). In this case, such regression lines approximate isolated connected components in order to achieve a local maximum in the EM algorithm to achieve algorithm convergence. Such isolated connected component can come from *dots* in letters as *i* or *j*, but also from wrong extraction of connected components, probably due to document noise. Such lines are not removed in the pruning step, since accurate represent *noisy* connected components and therefore, the overall precision is highly penalized.

Finally, we have grouped a set of not so common mistakes in the third group of errors. This type of mistakes represent around the 10% of the reported cases. They are mainly produce by an under estimation of the number of lines, resulting in labeling two or more consecutive lines with the same lines number, or obtaining crossed lines, which penalize the one to one matches as well as the precision and recall values.

Results obtained on the George Washington database are shown in Table II. We have applied our method directly on this database without taking into account special features present in this collection of documents. However, we have

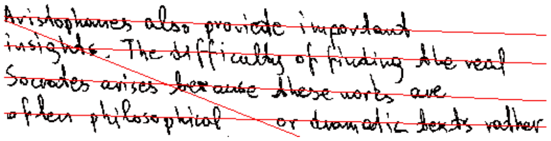
Aristophanes also provide important
insights. The difficulty of finding the real
Socrates arises because these works are
often philosophical or dramatic texts rather

Fig. 2. Error case produced by an over estimation of the number of lines.

identified some elements on the images that we know will affect to the performance of the method. On the one hand, some documents may include separator lines that have been labeled with the same label or a different one according to the distance to the closest word. Figure 3 illustrates one of these artifacts, on which the separator line has to be labeled as an individual line, but our method has included its unique connected component among the connected components of the text elements above. This situation also appears on short text elements that appear along the documents, as signatures, dates or footnotes. In these cases is also common that two regression lines finish trying to be fit between all the connected components in the region, as shown in Figure 4. On the other hand, the notarization process applied in the extraction of the connected components incorporates some noise into the image that is considered as additional connected components that disturb the estimation of the regression lines, affecting to the overall result. Nevertheless, despite of these document-related issues, we consider that the obtained results are similar to other reported results on the same document collection despite direct comparison can not been done.

	M	o2o	DR (%)	RA (%)	FM (%)
Proposed	**631**	**551**	**82,6**	**87,3**	**84,8**

TABLE II. RESULTS ON THE GEORGE WASHINGTON DATASET

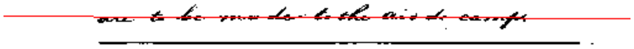

Fig. 3. Example of separator line included in a wrong regression line.

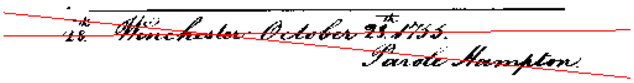

Fig. 4. Two regression lines trying to fit on connected components from different lines.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a handwritten line segmentation method based on a variation of the EM algorithm for regression line estimation. We have estimated a set of regression lines fitting the connected components on the document image. We have evaluated the proposed method on two reference datasets with promising results. In the case of the ICDAR2009 handwriting segmentation dataset, we have overcame most of the presented methods using very simple structural features such the centroids of connected components. We expect that taking into account more image information results will be improved. We have also validated our method on the George Washington database, demonstrating that our method is able to be applied on other type of documents without modifications on the main scheme.

In addition, we have some ideas to be included in future versions of our method. We plan to improve the estimation of the orientation and the number of lines by enriching the model used in this algorithm. Finally, we plan to integrate our method into a complete handwriting recognition system, so we will be able to evaluate its performance according to the results obtained from the handwriting recognition module.

REFERENCES

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, “Text line segmentation of historical documents: a survey,” *Int. J. Doc. Anal. Recognit.*, vol. 9, no. 2, pp. 123–138, Apr. 2007.

[2] R. P. d. Santos, G. S. Clemente, T. I. Ren, and G. D. C. Cavalcanti, “Text line segmentation based on morphology and histogram projection,” in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ser. ICDAR ’09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 651–655.

[3] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein, “Line segmentation for degraded handwritten historical documents,” in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ser. ICDAR ’09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1161–1165.

[4] S. Marinai and P. Nesi, “Projection based segmentation of musical sheets,” in *Document Analysis and Recognition, 1999. ICDAR ’99.*, sep 1999, pp. 515 –518.

[5] N. Ouwayed and A. Belaïd, “A general approach for multi-oriented text line extraction of handwritten document,” *International Journal on Document Analysis and Recognition*, vol. 14, no. 4, 2011.

[6] V. Shapiro, G. Gluhchev, and V. Sgurev, “Handwritten document image segmentation and analysis,” *Pattern Recognition Letters*, vol. 14, no. 1, pp. 71 – 78, 1993.

[7] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, “Text line detection in handwritten documents,” *Pattern Recogn.*, vol. 41, no. 12, pp. 3758–3772, 2008.

[8] Z. Shi and V. Govindaraju, “Line separation for complex document images using fuzzy runlength,” in *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL’04)*, ser. DIAL ’04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 306–.

[9] S. S. Bukhari, F. Shafait, and T. M. Breuel, “Script-independent handwritten textlines segmentation using active contours,” in *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, ser. ICDAR ’09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 446–450.

[10] D. Fernandez, J. Llados, A. Fornes, and R. Manmatha, “On influence of line segmentation in efficient word segmentation in old manuscripts,” in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, sept. 2012, pp. 763 –768.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” 1977.

[12] B. Gatos, N. Stamatopoulos, and G. Louloudis, “Icdar 2009 handwriting segmentation contest,” in *Document Analysis and Recognition, 2009. ICDAR ’09. 10th International Conference on*, july 2009, pp. 1393 – 1397.

[13] A. Fischer, A. Keller, V. Frinken, and H. Bunke, “Lexicon-free handwritten word spotting using character hmms,” *Pattern Recogn. Lett.*, vol. 33, no. 7, pp. 934–942, 2012.

[14] A. Asi, R. Saabni, and J. El-Sana, “Text line segmentation for gray scale historical document images,” in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, ser. HIP ’11. New York, NY, USA: ACM, 2011, pp. 120–126.