On the evaluation of handwritten text line detection algorithms

Bastien Moysset and Christopher Kermorvant
A2iA
39 rue de la Bienfaisance, Paris, France
{*bm,ck*}*@a2ia.com*

***Abstract*—Even if numerous text line detection algorithms have been proposed, the algorithms are usually compared on a single database and according to a single metric. In this paper, we study the performance of four different text line detection algorithms, on four databases containing very different documents, and according to three metrics (ZoneMap, ICDAR and recognition error rate). Our goal is to provide a more comprehensive empirical evaluation of handwritten text line detection methods and to identify what are the key points in the evaluation. We show that the different algorithms yield very different results depending on the type of documents and that two of them are constantly better than the others. We also show that the ZoneMap and the ICDAR metric are strongly correlated, but the ZoneMap metric provides greater detail on the error types. Finally we show that the geometric metrics are correlated to the recognition error rate on easy to segment databases, but this has to be confirmed on difficult documents.**

***Keywords*-Document Layout Analysis, Text line detection, Handwriting recognition, Evaluation metrics**

I. Introduction

Handwritten text line detection is a critical step in the processing of handwritten documents. The detection and extraction of text lines is a pre-requisite for the recognition of the text, and a text line badly detected or segmented cannot be correctly recognised. Numerous algorithms have been developed in order to tackle this problem [1].

These algorithms can be classified in five different categories. First, in the projection based methods, lines are separated by finding the minima in the projection histograms, either globally or locally [2]. Some techniques [3], [4] use blurring or morphological operations to get a smearing-like result where the connected components correspond the different lines. Hough transform methods can also be adapted to handwritten texts [5]. Other techniques include finding minimum paths to join left and right borders of the page without crossing text lines [6], [7]. Finally stochastic methods [8] can be used to separate the image in line and inter-line classes.

Despite the numerous papers published and several competition organized [9] on the subject, the existing algorithms are usually evaluated on a single database for which they have been tuned and have seldom been evaluated in the context of the complete application, which is handwritten text recognition. Indeed, what really matters is to evaluate how much a good text line segmentation improves the recognition results.

In this paper, the text line detection algorithms are evaluated both with geometrical metrics based on the ground-truth values of text line boxes and with a textual metric which evaluates the impact of the text line detection on the recognition error rate.

Moreover, we evaluate the different algorithms on four different document image databases, from well written synthetic handwritten pages (IAM) to real historical documents (Numen-RA). Our goal is to show the behaviour of the algorithms on databases on which they have not been tuned.

The structure of this paper is the following: we first describe the text line detection algorithms, the databases and the metrics and finally present the results of the evaluation.

II. Handwritten Text Line Detection algorithms

We first describe the different algorithms chosen for the evaluation. One algorithm was chosen in each of the different categories of handwritten text line detection algorithms. We did not tuned them for a particular database in order not to bias the results.

A. Projection algorithm

This technique is based on the horizontal projection histogram of pixel intensity. The histogram is smoothed and the local minima are extracted. Consecutive minima define the lines boundaries. Finally, thin boxes due to noise are removed. This technique is a very simple baseline.

B. Rectangle-based filtering

This smearing-like method is inspired by [3]. First, a median filtering using a rectangular mask with the same orientation as the text is applied to the image. Then, a binarization is applied using the Otsu algorithm. The binary image obtained shows horizontal components representing text lines. The bounding box of these components is extracted and defines the text line position. Finally the small rectangles corresponding to noise are removed.

C. Shredding method

This technique is detailed in [7]. The height of a text line is estimated by taking the median value in the histogram of the height of the connected components. Then, a rectangular

median filtering is applied to the image. Starting from the leftmost pixels, a path toward the right is created following the valleys between the lines. At each pixel, the path is extended to the direction of the whiter pixels in a scope defined by the median height of connected components. The same process is performed from right to left. The detected text lines are composed of the pixels than have not been crossed by these two paths. Finally, small boxes are removed.

D. Hough-based method

This technique, inspired by [5], is based on the Hough transform. The connected components of the document are extracted and the median height of these components is computed. Based on this value, connected components are classified in three categories corresponding to normal components, small components (usually dots or diacritics) and big components that may belong to several text lines. Normal components are horizontally split in sub-components whose gravity centres are used as voting points in the Hough array. Maxima in the Hough array correspond to main lines. The components of the normal subset are joined to a line if at least half of its sub-components are voting for the Hough array point corresponding to this line. Finally, small components are joined to the closer line and big components are split between the neighbouring lines.

III. Image databases

For this study, we have selected four image databases showing a diversity in type of writing, language, epoch of writing and for which ground-truth values were available for line position or text. Samples of these image databases are shown in Figure 1. We did not select the database used for the ICDAR2009 text line detection competition [9]: we considered it too easy since most of the line detection methods achieved more the 90% of correct detection rate.

A. IAM database

The IAM database V3.0 [10] is composed of 1539 English handwritten pages. The subset used for the evaluation in this paper is the official *Validation2* subset, which is disjunct from the training set of the recognizer (official *Train* subset). The documents are clean, carefully written and the lines are, in general, well separated and easy enough to segment. The ground-truth is provided for textual content and location at word and line levels. The *Validation2* subset contains 115 images and 940 text lines.

B. Rimes database

The RIMES database [11] was developed for the evaluation of automated systems for handwritten document processing. Since the documents were carefully produced, this database is similar to the IAM database, but in French. It is composed of 5599 handwritten pages, with human made ground-truth transcription. The ground-truth values for text lines positions are not available. The subset used for the evaluation was the ICDAR2011 evaluation set[1] composed of 200 pages. The recognizer was trained on the ICDAR2011 train set which is disjunct from the evaluation set.

C. OpenHaRT 2010 database

The OpenHart database was provided by the NIST for the OpenHaRT 2010 contest[2]. In this paper, we used for evaluation the MADCAT_Phase1_DevTest set which is composed 845 images with 14.160 annotated line positions. This database is composed of binary images of Arabic handwritten texts. The high number of diacritics in Arabic writing associated to text lines frequently overlapping or touching make the text lines segmentation relatively difficult.

D. Numen-RA database

The Numen-RA database, provided by Numen Digital, is a set of 13.649 historical documents in old creole French containing land surveying reports. The evaluation set is composed of 665 images with 14.862 lines. The text line segmentation of this database encounters several challenges that commonly occurs when dealing with historical document: luminosity variation, background noise, appearance of verso writings, black bands around the image or annotations in the margins.

IV. Evaluation methods

In our experiments, we have used three different metrics for the evaluation of the text line detection methods. Two of them, the ZoneMap metric and the ICDAR metric are based on the image only, whereas the third one is based on the recognition error rate.

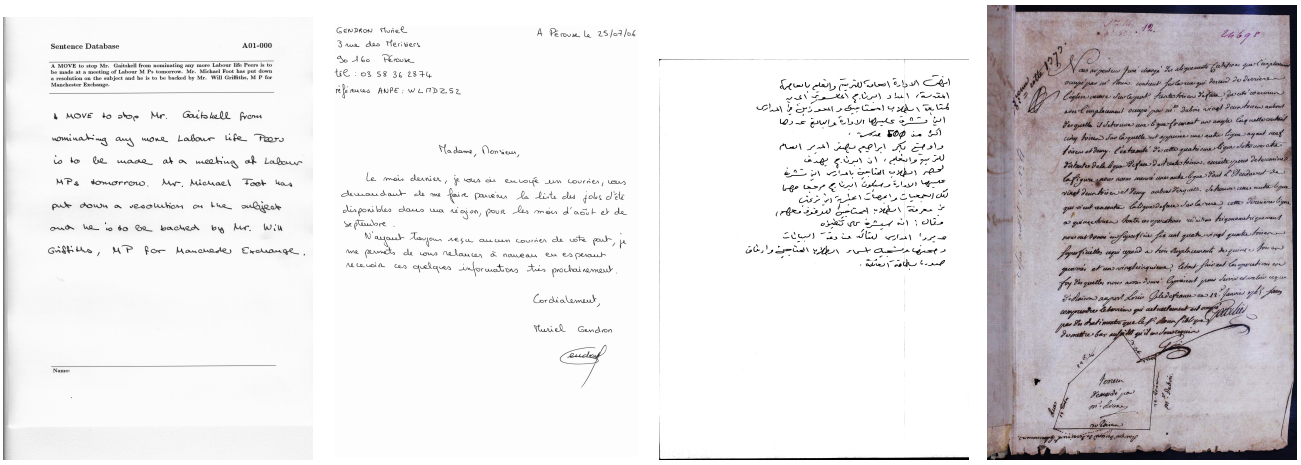
A. ZoneMap metric

The ZoneMap metric was developed by the Laboratoire National de métrologie et d'Essais (LNE) for the Maurdor campaign[3]. This metric was designed to evaluate the accuracy of document layout systems, but it can be applied to text line detection. The algorithm of this metric first computes the strength of the link between each hypothesis zone and each reference zone. This strength is defined in Equation 1 with R as the reference box, H as the hypothesis box and where $\texttt{Surface}(Box)$ corresponds to the number of black pixels in the box.

$$f(R,H) = \left(\frac{\texttt{Surface}(H \cap R)}{\texttt{Surface}(H)}\right)^2 + \left(\frac{\texttt{Surface}(H \cap R)}{\texttt{Surface}(R)}\right)^2 \quad (1)$$

Then, the zones are grouped by decreasing strength values if it doesn't lead to a situation with at the same time

[1] http://www.rimes-database.fr
[2] http://www.nist.gov/itl/iad/mig/hart2010.cfm
[3] http://www.maurdor-campaign.org/



(a) IAM (b) Rimes (c) OpenHaRT (d) Numen-RA

Figure 1: An example of image from the four databases used for the evaluation.

several hypothesis boxes and several reference boxes in the same group. If it does, no grouping is done. Each group is considered as a match, a merge, a split, a miss or a false acceptance according to the number of references and hypothesis boxes it contains. The error associated to each group corresponds to the number of black pixels that are not well classified:

- for a match configuration, it is the number of black pixels that are in the union of the boxes but not in their intersection;
- for a miss or a false acceptance configuration, it is the number of black pixels in the box;
- for a merge or a split configuration it is the number of black pixels that are not in the biggest intersection between an hypothesis and a reference box.

Total error for the document image is the sum of the group errors normalized by the total number of black pixels in the reference boxes.

B. ICDAR metric

The second metric used in our experiments is the metric used in the ICDAR 2009 text line detection competition [9], initially proposed by [12] for graphics recognition systems. This metric is based on threshold which defines the proportion of black pixels that the hypothesis and reference zones must have in common to be considered as matching. A match score between each reference zone and each hypothesis zone is computed according to Equation 2.

$$\texttt{matching\_score} = \frac{\texttt{Surface}(H \cap R)}{\texttt{Surface}(H \cup R)} \qquad (2)$$

If the match score is greater than the threshold, the count of one to one matches is incremented. Then, the detection rate (DR) is defined as the number of one to one matches divided by the number of reference line boxes. Similarly, the recognition accuracy (RA) is defined as the number of one to one matches divided by the number of hypothesis line boxes. Finally, the error metric is defined by Equation 3.

$$\texttt{error} = 1 - \frac{2 * DR * RA}{DR + RA} \qquad (3)$$

C. Recognition metric

The goal of a text line detection algorithm is to provide well located text lines to the text recognizer. The quality of the text line extraction has a strong impact on the recognition results. In order to evaluate this impact, we have used the recognition result of a handwriting text recognizer as a metric. We have trained a Multi-dimensional Long-Short Term Memory (MDLSTM) recurrent neural network as described in [13] . The recognizer was trained on isolated words so that the training does not rely on a specific text line detection method. The decoding was done with a vocabulary of size 2318 for IAM and of size 2318 for Rimes (0% of out-of-vocabulary words in both cases). No language model was used. The text lines images provided by the four text line detection methods were processed by the recognizer and the recognition error rate was computed using sclite form the NIST SCTK scoring package [4].

V. Results

We have systematically evaluated the four text line detection methods, on the four databases and using the three metrics (when it was possible). All the evaluation results are presented on Table I.

A. Comparison of the ZoneMap and ICDAR error metric

The figure 2 shows the relation between the values of the ZoneMap metric and the ICDAR metric for the four line

[4] http://www.itl.nist.gov/iad/mig/tools/

	Projection			Rectangle			Shredding			Hough		
	ZoneMap	ICDAR	RecoErr	ZoneMap	ICDAR	RecoErr	ZoneMap	ICDAR	RecoErr	ZoneMap	ICDAR	RecoErr
IAM	48.4	62.3	72.4	4.2	3.3	23.9	3.6	7.4	23.7	75.5	64.0	91.1
OpenHaRT	46.8	50.2	*	29.3	33.1	*	78.9	55.6	*	158.1	61.8	*
Numen-RA	96.2	98.1	*	61.6	73.7	*	31.2	60.6	*	166.2	89.5	*
Rimes	*	*	63.9	*	*	13.2	*	*	17.5	*	*	87.5

Table I: Evaluation of four line detection algorithms (Projection, Rectangle, Shredding and Hough), on four databases (IAM, OpenHaRT, Numen-RA and Rimes) according to three error metrics (ZoneMape, ICDAR and recognition error rate). For all the metrics, a lower value is better, and the score can be greater than 100. A star (*) indicates that the evaluation could not be computed either because the ground-truth is missing or the recognizer was not available.

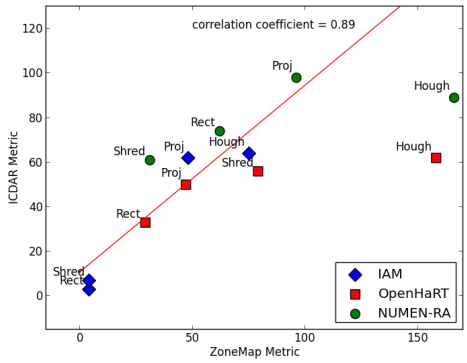


Figure 2: Correlation of the ZoneMap and ICDAR metrics on three different databases (IAM, OpenHaRT, NUMEN-RA) and for four different line detection algorithms.

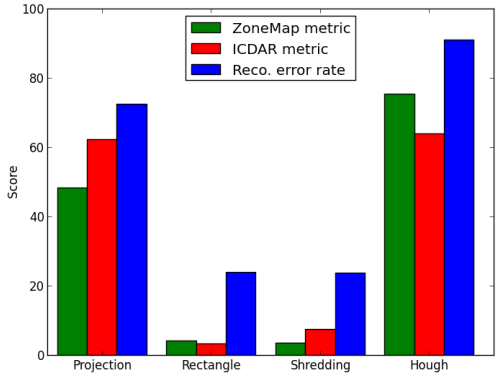


Figure 4: Comparison of line detectors with the three metric (ZoneMap, ICDAR, recognition error rate) on the IAM database. For all the metrics, a lower value is better

detection methods on the three databases with line location ground-truth values (IAM, OpenHaRT, NUMEN-RA). The two metrics are generally well correlated: if we exclude two outliers, the correlation coefficient is 0.89.

However, the ZoneMap metric offers a more detailed analysis of the errors, as shown on Figure 3. Looking at the detailed results, one can see that the same algorithm can suffer from very different types of error, depending on the database.

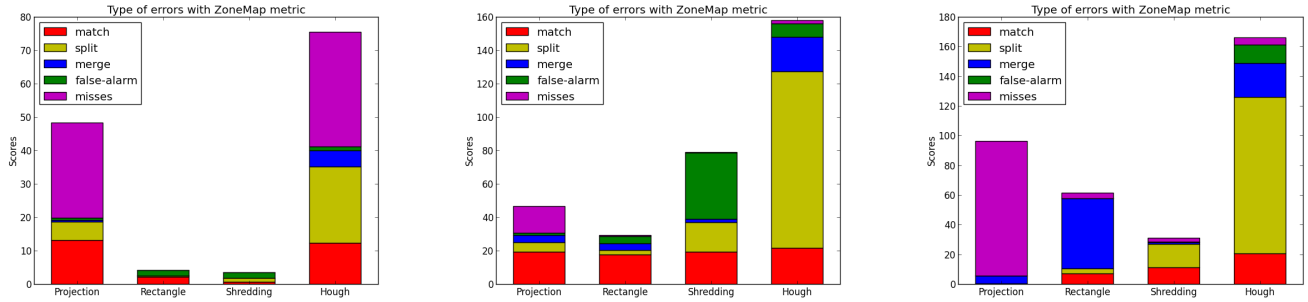
B. Comparison of the different algorithms

The comparison of line detection algorithms results with the different databases confirm the fact that some algorithms are more adapted than the other to certain image types. In particular, we observe that the shredding-based method and the rectangle median filtering method show good results on a easy database such as IAM. The OpenHaRT2010 database is more difficult due to closer lines with a high number of overlapping components and a lot of diacritics between lines. On this database, the shredding and rectangle manifest a very different behaviour: the rectangle filtering method still shows correct results but the shredding method shows poorer results because it is harder to find a proper path between lines. On the contrary on the historical documents of the Numen-RA database, the rectangle filtering method is disrupted by the annotations or by the black bands in the margin while the shredding method works better. In conclusion, none of the four methods outperform the other on the four databases. Only the Hough method under-perform the other methods on all the databases.

C. Relation between the layout metrics and the recognition metric

The IAM database provides the ground-truth for both the line location and the textual content. The layout metrics (ZoneMap and ICDAR) were compared to the recognition rate for the four text line detection algorithms, as shown on Figure 4. The ZoneMap metric seems to be more correlated to the recognition error rate than the ICDAR metric. For example, the projection and Hough methods have about the same score with the ICDAR metric (62-64 %) but different recognition error rate, which is reflected by ZoneMap metric. More data must be collected, for example on the OpenHaRT2010 and the Numen-RA database, to corroborate this analysis.

Finally, if we consider two similar databases such as IAM and Rimes (Table I), we can see a very similar behaviour of the different line detection algorithms: on both databases, the two best algorithms are Rectangle and Shredding whereas



(a) IAM (b) OpenHaRT2010 (c) Numen-RA

Figure 3: Detailed error analysis of the different line detection algorithms with the ZoneMap error rate.

Hough and Projection yield poor results. This illustrates a certain stability of the algorithms on similar databases.

VI. Conclusion

In this paper, we have evaluated four different handwritten text line detection algorithms, on four different databases and using three different metrics. Our goal is to provide a more comprehensive empirical evaluation of handwritten text line detection methods and this paper is a first step into this direction. To go further, the geometrical metrics can still be improved to better reflect the impact of the different type of errors on the recognition. For example, the ICDAR 2009 and ZoneMap metrics consider equally the split of a line and the merge of two lines or the false alarm and the misses whereas merging two lines has a stronger impact on text recognition than splitting a line (vertically). We have shown that the algorithms behave very differently depending on the type of documents: for a more comprehensive evaluation, we need to consider an even wider scope of document diversity.

Acknowledgement

This work was funded by the French Grand Emprunt-Investissements d'Avenir program through the PACTE project.

References

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 123–138, Sep. 2006.

[2] A. Zahour, L. Likforman-Sulem, W. Boussellaa, and B. Taconet, "Text Line Segmentation of Historical Arabic Documents," in *International Conference on Document Analysis and Recognition*, 2007.

[3] Z. Shi, S. Setlur, and V. Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines," in *International Conference on Document Analysis and Recognition*, 2009.

[4] V. Papavassiliou, V. Katsouros, and G. Carayannis, "A Morphological Approach for Text-Line Segmentation in Handwritten Documents," in *International Conference on Frontiers in Handwriting Recognition*, 2010.

[5] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," *Pattern Recognition*, vol. 42, no. 12, pp. 3169–3183, Dec. 2009.

[6] R. Saabni and E.-S. Jihad, "Language-Independent Text Lines Extraction Using Seam Carving," in *International Conference of Document Analysis and Recognition*, 2011.

[7] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines," *International Conference on Document Analysis and Recognition*, 2009.

[8] T. Stafylakis, V. Papavassiliou, V. Katsouros, and G. Carayannis, "Robust Text-line and Word Segmentation for Handwritten Documents images," in *International Conference on Acoustics, Speech and Signal Processing*, 2008.

[9] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICDAR2009 handwriting segmentation contest," *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, pp. 25–33, Jun. 2010.

[10] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, Nov. 2002.

[11] E. Augustin, M. Carré, E. Grosicki, J.-M. Brodin, E. Geoffrois, and F. Preteux, "RIMES evaluation campaign for handwritten mail processing," in *Proceedings of the Workshop on Frontiers in Handwriting Recognition*, 2006.

[12] I. Phillips and A. Chhabra, "Empirical performance evaluation of graphics recognition systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 849–870, 1999.

[13] F. Menasri, J. Louradour, A.-L. Bianne-Bernard, and C. Kermorvant, "The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition," in *Document Recognition and Retrieval Conference*, 2012.