Baseline Estimation in Arabic Handwritten Text-Line
Evaluation on AHTID/MW Database

Anis Mezghani[1], Slim Kanoun[1,2], Souhir Bouaziz[2], Maher Khemakhem[1] and Haikal El Abed[3]
[1]*MIRACL Lab, ISIMS, University of Sfax, Sfax, Tunisia*
[2]*University of Sfax, National School of Engineers (ENIS), Sfax, Tunisia*
[3]*Braunschweig Technical University, Institute for Communications Technology (IfN), Braunschweig, Germany*

Keywords: Baseline Estimation, Handwritten Text-Line, AHTID/MW Database, Ground Truth.

Abstract: Baseline extraction is one of the most important phases for handwriting recognition. Due to the complexity of the Arabic scripts, baseline detection of Arabic handwritten text-lines is a difficult task compared to other languages. In this work, a method which combines some baseline extraction techniques used in literature was presented to provide a fine estimation of baseline in Arabic handwritten text-lines. For evaluation purpose, the AHTID/MW database was extended by a baseline ground truth annotation. The database is freely available for researchers worldwide which enable other researchers to test their baseline detection systems.

1 INTRODUCTION

In pattern recognition field, handwritten text recognition is considered as one of the most complicated problem. Moreover, the complexity is increased in Arabic language because the text is written cursively in addition to the complexity of the text characteristics (Al-Badr and Mahmoud, 1995). The existing research on recognizing Arab text is still limited compared with Latin or China's languages. In the Arabic OCR (Optical Character Recognition) system, preprocessing stage is the most important because it directly affects the reliability and efficiency in the segmentation and feature extraction process (Farooq et al., 2005). After enhancing the quality of the image, one of the prior parts of the text preprocessing is the estimation of the writing line called Baseline.

The baseline is a vertical reference position for the characters and subwords in a handwritten text-line image. The baseline has been used by most of the Arabic OCR systems; it can be used in skew normalization (Pechwitz and Märgner, 2003), to segment the Arabic text into words or characters (Amin, 1998) and to make the text ready for the feature extraction stage (El-Hajj et al., 2005). In Arabic handwritten text, classic methods of baseline estimation such as the horizontal projection are not suitable because of wide variety of writing styles and specific characteristics such as cursive writing and large number of dots.

Considering these issues, we propose to develop a method that provides a fine estimation of baseline in Arabic handwritten text-line. To the best of our knowledge, there is no Arabic handwritten text-line dataset with ground truth information. Therefore, we extended the AHTID/MW database (Mezghani et al., 2012) by a baseline ground truth annotation of text-line images. This database is freely available for the scientific community and may be used as a benchmark database where researchers can evaluate and compare their algorithms and results with other published works.

In Section 2, we will outline the published works related to baseline detecting methods. Section 3 describes the developed method of baseline estimation in Arabic handwritten text-line. In Section 4, the AHTID/MW database structure along with baseline ground truth data is presented. Experimental results are reported in Section 5. Finally, concluding remarks are given in Section 6.

2 RELATED WORK

Different baseline extraction methods are presented in literature. Al-Shatnawi and Omar (2008)

Mezghani A., Kanoun S., Bouaziz S., Khemakhem M. and El Abed H. (2013).
Baseline Estimation in Arabic Handwritten Text-Line - Evaluation on AHTID/MW Database.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 430-434
DOI: 10.5220/0004218704300434

classified the Arabic baseline extraction methods into four different groups based on the techniques used. The simplest one is based on horizontal projection. Elgammal and Ismail (2001) detected the baseline by finding peak value of horizontal projection profile in a printed text-line. This method has the defect to be very sensitive to the skew (Pechwitz and Märgner, 2002). A modified projection technique based on rotating word image through different angular inclinations is presented by Al-Rashaideh (2006). The baseline is identified by finding the maximum value and corresponding angle among all the peak values is obtained. Pechwitz and Märgner (2002) proposed the only one work to detect Arabic handwriting baseline according to the word skeleton. The main idea of this approach is to calculate robust features from the skeleton and use these features for classifying the connected components into baseline relevant and baseline irrelevant areas. In a subsequent step, a regression analysis of points of the relevant objects is done to estimate the final baseline position. Some researchers extract baselines after correcting the slant of the word by a linear regression of the critic points of the contour having nearly the same horizontal positions (Farooq et al., 2005). Burrow (2004) presented a method based on angle detection by principle components analysis.

Other methods such as the minimization of the entropy and Hough transform based methods, which are used for Latin script, are developed and applied on Arabic script (Côté et al., 1996; Likforman-Sulem et al., 1995). These methods have the defect to be expensive in term of calculation time. Lemaitre et al. (2009) proposed a script independent method for baseline detection. This method is based on the principle of the perceptive vision, which combines several points of view of the same word (from low to high resolution). Boubaker et al. (2009) described a baseline detection method which considers geometric and topologic features. It is tested on online and offline short Arabic handwritten writing. Recently, a two-stage Persian/Arabic baseline detection and correction algorithm is presented by Ziaratban and Faez (2008). The first stage estimates the writing path of a text-line by a fitted curve based on candidate baseline pixels, which are detected using template matching algorithm. Then the slant and position of the components in the line is adjusted. In the second stage, the baseline for each subword is corrected. Other method of tracing the baseline in handwritten Persian/Arabic text-line is proposed by Nagabhushan and Alaei (2010). This method is based on preparing patches of black and white blocks all along the text-line, identifying some candidate points and regressing a curve through these candidate points to trace the baseline.

The majority of methods presented in literature failed in estimating the correct baseline for handwritten text having greater number of ascenders and descenders. Menasri et al. (2008) described a baseline extraction method of words overcoming some difficulties in Arabic script such as the presence of loops and various shapes for a group of two or three dots. Inspired by this work, we developed a baseline estimation method adapted to Arabic handwritten text-lines.

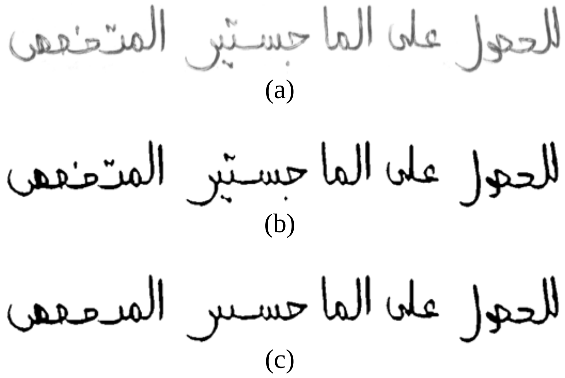
3 BASELINE ESTIMATION

After scanning a document, some basic preprocessing tasks like image binarization and noise reduction have to be performed to increase the readability of the input by the baseline detection system. Using the binary image, we perform a noise reduction filtering. Small holes, produced by writing and binarization process, are closed and the unwanted information is deleted by using the opening and closing morphology operation respectively (Figure 1(b)).

The developed baseline detection process consists of three stages: the first one is a basic stage leading to the detection and removing of diacritical marks. The second stage extracts the upper baseline and the lower baseline based on the horizontal projection histogram. In the final stage, we estimate more precisely the baseline using support points.

3.1 Diacritical Marks Elimination

More than half of Arabic letters include in their shape dots which can be one, two or three dots. The presence of these dots, called diacritical marks, in their positions allows us to differentiate between letters that belong to the same family shape. These diacritical marks lie in either above or under the baseline depending on the character. In order to circumvent the bad influences in the process of baseline detection when using horizontal projection, we start by removing the diacritical marks based on the size of the connected components as described in (Menasri et al., 2008). A sample of a text-line image after removing the diacritical marks is shown in Figure 1(c).



(a)

(b)

(c)

Figure 1: Arabic handwritten text-line image from AHTID/MW database: (a) before binarization; (b) after binarization and noise reduction; (c) after diacritical marks elimination.

3.2 Primary Baseline Estimation

For primary baseline estimation, we used the horizontal projection method. In general, this method is used to detect two baselines in each input image, upper and lower baseline (El-Hajj et al., 2005). Horizontal projection histogram will be disturbed by many kinds of noises. Among them is the succession of descenders, or long tails under baseline that could lead to a high peak in the histogram under the baseline. In Arabic script, loops are often located in the baseline. So, to overcome this problem, we start by detecting the loops to pre-locate a horizontal band. This horizontal band is three times the maximum length of loops centered on loops.

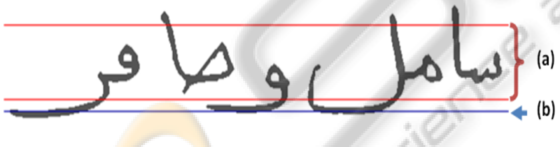


Figure 2: Pre-localization of horizontal band by the use of loop position: (a) pre-localization of horizontal band; (b) high peak due to the succession of descenders.

After pre-localization based on loops, we compute the projection histogram and calculate the global maximum. Baselines correspond to local minima above and below the global maximum which are lower than 1/3 of the global maximum. The coefficient 1/3 is obtained from tests. Figure 3 illustrates improved baselines of a text-line image.

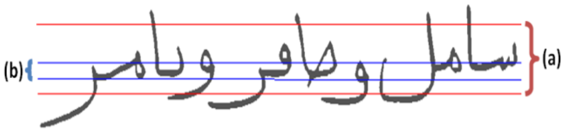


Figure 3: Improvement of horizontal band: (a) search area of the global maximum; (b) primary baseline estimation.

3.3 Fine Baseline Estimation

In this step, we evaluate more precisely the lower baseline using support points. Those support points are singular points of the skeleton and local minimums located in the baseline.

- A skeletonization of the text-line image is performed using the Toumazet algorithm (Toumazet, 1990). The Thomé algorithm (Thomé, 1978) was utilized to bring the thickness to a single pixel taking into account saving the geometry, location and connections. Singular points of the skeleton are defined as points for which one stroke starts inside the baseline and finishes under the baseline (Figure 4).

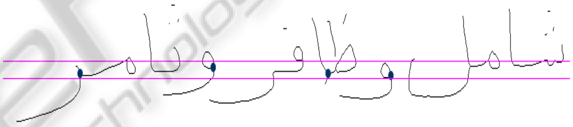


Figure 4: Detected singular points from skeleton text-line.

- Local minima points are deduced from the contour of sub-words. We retain only local minimums located in the baseline.

Based on all these support points, a linear interpolation is applied to detect the approximate baseline. This fine detection of the baseline adapts well to small changes in the inclination of the writing in the same text-line. Figure 5 gives an example of the resulting baseline estimation for an Arabic handwritten text-line.

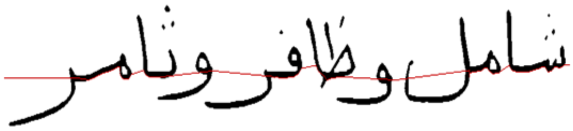
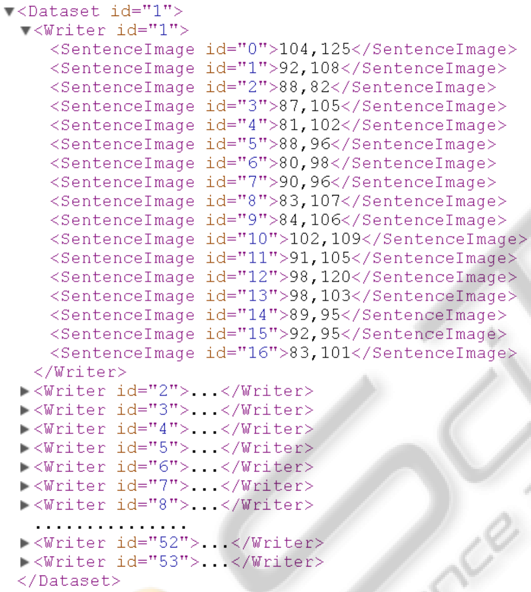


Figure 5: Baseline estimation of an Arabic handwritten text-line.

4 BASELINE GROUND TRUTH DESCRIPTION

To evaluate the present work, we extended the AHTID/MW database developed by Mezghani et al. (2012) by a baseline ground truth annotation. The AHTID/MW database contains 3710 text-line images written by 53 Arabic native writers. These images are divided into 4 sets. For each of these sets, we provide a baseline ground truth of the data using an XML file. An example of such XML file is given in Figure 6.

A baseline is drawn on each text-line image manually. This straight line should give a good estimation of the writing line. The baseline is parameterized by two values, *Y1* and *Y2*, which represent endpoints of the baseline.



```
▼<Dataset id="1">
  ▼<Writer id="1">
     <SentenceImage id="0">104,125</SentenceImage>
     <SentenceImage id="1">92,108</SentenceImage>
     <SentenceImage id="2">88,82</SentenceImage>
     <SentenceImage id="3">87,105</SentenceImage>
     <SentenceImage id="4">81,102</SentenceImage>
     <SentenceImage id="5">88,96</SentenceImage>
     <SentenceImage id="6">80,98</SentenceImage>
     <SentenceImage id="7">90,96</SentenceImage>
     <SentenceImage id="8">83,107</SentenceImage>
     <SentenceImage id="9">84,106</SentenceImage>
     <SentenceImage id="10">102,109</SentenceImage>
     <SentenceImage id="11">91,105</SentenceImage>
     <SentenceImage id="12">98,120</SentenceImage>
     <SentenceImage id="13">98,103</SentenceImage>
     <SentenceImage id="14">89,95</SentenceImage>
     <SentenceImage id="15">92,95</SentenceImage>
     <SentenceImage id="16">83,101</SentenceImage>
   </Writer>
  ▶<Writer id="2">...</Writer>
  ▶<Writer id="3">...</Writer>
  ▶<Writer id="4">...</Writer>
  ▶<Writer id="5">...</Writer>
  ▶<Writer id="6">...</Writer>
  ▶<Writer id="7">...</Writer>
  ▶<Writer id="8">...</Writer>
   ...............
  ▶<Writer id="52">...</Writer>
  ▶<Writer id="53">...</Writer>
 </Dataset>
```

Figure 6: An example of a baseline ground truth data files.

5 RESULTS AND DISCUSSION

Experiments have been carried out on AHTID/MW database. It consists of 3710 Arabic handwritten text-line images containing 22896 words (Mezghani et al., 2012) with ground truth information. Due to the fact that the baseline ground truth is provided, so it is possible to evaluate the baseline. The error of a baseline is calculated as the average distance between the ground truth and the proposed result (figure 7). Two thresholds are proposed with this metric: with an average pixel error less than 5 pixels, the baseline is considered as *good*, whereas with an error up to 7 pixels, the baseline is *acceptable*. In order to obtain the average pixel error, we discretized the baselines into equal intervals fixed from tests at 20 pixels. The average pixel error is defined as the average of distances between the ground truth and the proposed result at different positions.

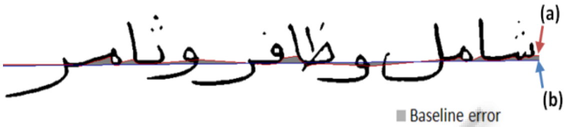


Figure 7: Visualization of the baseline error: (a) estimated baseline; (b) baseline ground truth.

The results of the proposed baseline estimation method are reported in Table 1. We must point that our text-line dataset is complex due to the presence of different handwriting styles. So, the dot removing algorithm proposed in section 3.1 is not able to detect all diacritical marks. The absence of other Arabic handwritten text-line datasets with baseline ground truth information disables as to obtain a precise comparison of our method with existing work. Therefore, we invited interested researchers to use the AHTID/MW database.

Table 1: Performance of the proposed baseline estimation method.

Baseline quality (average pixel error)	Percentage (%)
Good (<5)	84.3
Acceptable (<7)	88.7

6 CONCLUSIONS

In this work, a baseline estimation method of Arabic handwritten text-lines is presented. In the first stage, the diacritical marks are eliminated. The upper and the lower baseline were extracted thanks to the horizontal projection method. Finally, we estimate more precisely the baseline using support points.

To evaluate the present work, we extended the AHTID/MW database developed by Mezghani et al. (2012) by baseline ground truth information of text-line images. This database contains 3710 text-line images written by 53 Arabic native writers. We would like to note that the AHTID/MW database, including baseline ground truth annotation, is freely available to interested researchers worldwide.

REFERENCES

Al-Badr, B., Mahmoud, S., 1995. Survey and bibliography of Arabic optical text recognition. *Signal Processing*. Vol.41(1): 49–77.

Al-Rashaideh, H., 2006. Preprocessing phase for Arabic Word Handwritten Recognition. *Electronic Scientific Journal*. Vol.6 (1): 11–19.

Al-Shatnawi, A., Omar, K., 2008. Methods of Arabic Baseline Detection -The State of Art. *International Journal of Computer Science and Network Security*. Vol.8 (10):137–142.

Amin, A., 1998. Off-line Arabic character recognition: the state of the art. *Pattern Recognition*. Vol.31(5): 517–530.

Boubaker, H., Kherallah, M., Alimi, M. A.,2009. New Algorithm of Straight or Curved Baseline Detection for Short Arabic Handwritten Writing. *International Conference on Document Analysis and Recognition*. 778–782.

Burrow, P., 2004. *Arabic handwriting recognition*, Thesis. University of Edinburgh. England.

Côté, M., Chériet, M., Suen, C., Lecolinet, E., 1996. Détection des Lignes de Base de Mots Cursifs à l'aide de l'Entropie. *Colloque sur l'Intelligence Artificielle dans les Technologies de l'Information*.

Elgammal, A. M., Ismail, M. A., 2001. A Graph-Based Segmentation and Feature Extraction Framework for Arabic Text Recognition. *International Conference on Document Analysis and Recognition*. 622–626.

El-Hajj, R., Likforman-Sulem, L., A., Mokbe, C., 2005. Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling. *International Conference on Document Analysis and Recognition*. 893–897.

Farooq, F., Govindaraju, V., Perrone, M., 2005. Preprocessing Methods for Handwritten Arabic Documents. *International Conference on Document Analysis and Recognition*. 267–271.

Lemaitre, A., Camillerapp, J., Coüasnon, B., 2009. Multi-script Baseline Detection Using Perceptive Vision. *Biennial Conference of the International Graphonomics Society*.

Likforman-Sulem, L., Hanimyan, A., Faure, C., 1995. A Hough based algorithm for extracting text lines in handwritten documents. *International Conference on Document Analysis and Recognition*. 774–777.

Menasri, F., Vincent, N., Augustin, E., Cheriet, M., 2008. Un système de reconnaissance de mots arabes manuscrits hors-ligne sans signes diacritiques. *Conférence Internationale francophone sur l'écrit et le document*.

Mezghani, A., Kanoun, S., Khemakhem, M., El Abed, H., 2012. A Database for Arabic Handwritten Text Image Recognition and Writer Identification. *International Conference on Frontiers in Handwriting Recognition*. 397–400.

Nagabhushan, P., Alaei, A., 2010. Tracing and Straightening the Baseline in Handwritten Persian/Arabic Text-line: A New Approach Based on Painting-technique. *International Journal on Computer Science and Engineering*. Vol.2 (4): 907–916.

Pechwitz, M., Märgner, V., 2003. HMM Based approach for handwritten Arabic Word Recognition Using the IFN/ENIT DataBase. *International Conference on Document Analysis and Recognition*. 890–894.

Pechwitz, M., Märgner, V., 2002. Baseline Estimation For Arabic Handwritten Words. *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition*. 479–484.

Thomé, S., 1978. Prétraitement du chiffre manuscrit. *Congrès AFCET, France*. 568–576.

Toumazet, J. J., 1990. *Traitement de l'image par l'exemple*, Sybex.

Ziaratban, M., Faez, K., 2008. A Novel Two-Stage Algorithm for Baseline Estimation and Correction in Farsi and Arabic Handwritten Text line. *International Conference on Pattern Recognition*. 1–5.