# A New Baseline Estimation Method Applied to Arabic Word Recognition

Fouad Slimane[1,2] - Slim Kanoun[3] - Jean Hennebert[1,4] - Rolf Ingold[1] - Adel M. Alimi[2]
[1]*DIVA Group - University of Fribourg (unifr), Bd de Pérolles 90, CH-1700 Fribourg, Switzerland*
[2]*REGIM Lab. - University of Sfax, National School of Engineers (ENIS), BP 1173, Sfax, 3038, Tunisia*
[3]*University of Sfax, National School of Engineers (ENIS), BP 1173, Sfax, 3038, Tunisia*
[4]*Computer Science Department, EIAFR, HES-SO // Fribourg, Switzerland*
*Fouad.Slimane@unifr.ch, slim.kanoun@ieee.org, Jean.Hennebert@hefr.ch, Rolf.Ingold@unifr.ch, Adel.Alimi@ieee.org*

***Abstract*—We analyse in this paper the impact of different baseline identification approaches in the case of single word recognition. We show that classical baseline identification approaches using horizontal projection histograms may fail in detecting accurately the baseline of short words, impacting the overall processing chain and inducing errors. From this observation, we propose a novel approach based on stochastic models able to propose probable baseline regions from characters features. Once the most probable baseline region is detected, we fine tune the position of the baseline with an horizontal projection histogram. We ran our experiments in the case of a printed word recognition task using the APTI database and observed a significant increase of performance.**

***Keywords*-HMM; GMM; arabic recognition; baseline;**

## I. Introduction

Our work is in the field of single word OCR recognition present in low-resolution images. Such images are usually generated by screen-based OCR applications and are also frequently found in web pages [1]. Our work is also focusing on the Arabic language that present specific difficulties. In some previous works, we have shown that state-of-the-art HMM systems can still be used successfully in this context, of course showing lower range of performances that what can be observed for traditional OCR. We proposed several adaptation of the HMM system to increase their performance, such as inclusion of duration models [2], optimization of the set of character shape models [3] and a priori Arabic font recognition for multi-font inputs [4].

Pursuing our analysis of factors impacting the performance of Arabic screen-based OCR, we realized that our system was failing more frequently to recognize short words. A tracking of the errors shown that the errors were due to false detections of the baseline. In our case, we used a classical baseline detection algorithm based on horizontal projections where the baseline is estimated at the peak of pixel distribution projected horizontally on the y axis. Arabic word can indeed be composed of as low as two characters, and, in such cases, it is easy to falsely detect the baseline position with a simple horizontal pixel distribution analysis.

This paper is organized as follows: Section 2 provides a description of the preprocessing and normalization procedures. Section 3 and 4 give a description of the feature extraction and the word recognition system. Section 5 reports on the experimental results followed by some conclusions.

## II. Preprocessing

The preprocessing phase aims at the reduction of the variability between character shapes due to mis-alignment on the y axis as explained earlier. Classically, this preprocessing phase normalizes all inputs by shifting the images so that the characters of a word or sequence of words are aligned vertically according to a common baseline.

### A. Baseline detection with horizontal projection

The horizontal projection approach is widely used for baseline identification [5] [6] [7] [8]. This method is easy to implement but needs relatively wide inputs (i.e. long sequences of characters) to estimate correctly the baseline.

### B. Data-driven baseline detection system

Our idea for baseline detection method is to detect, firstly, the probable baseline region using data-driven methods trained on local character features ans secondly, the position of the baseline is fine-tuned using the classical horizontal projection histogram, but limited to this region. The data-driven system is actually similar to the system presented in [4] for Arabic font recognition. Gaussian Mixture Models (GMMs) are used to estimate the likelihoods of three baseline positions (called here *below*, *middle* and *above* positions). Each position is represented by a single GMM (a 1-state HMM). Assuming the independence of the feature vectors, the GMMs are able to compute a global likelihood [9] of a baseline position simply by multiplying the local likelihoods of each feature vectors computed separately.

## III. Feature extraction

The GMM system for the baseline detection, as well as the HMM system for word recognition share the same feature extraction module. It is presented in more details in our previous work [4]. A feature vector is extracted from each analysis window. Using a simple right-left sliding procedure of the analysis window, no segmentation into letters is made and the word image is transformed into a sequence of feature vectors. Each feature vector has $M$ components including $M/2$ basis features concatenated with $M/2$ so-called delta coefficients computed as in [4].

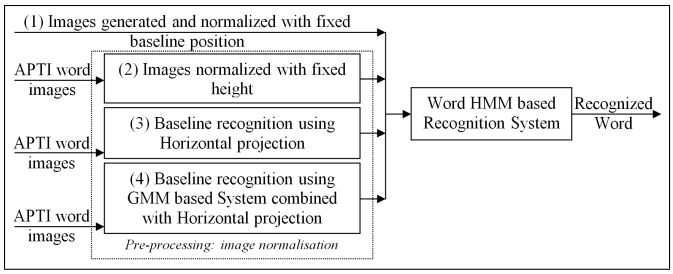


Figure 1. Overview of the 4 image normalisation and baseline detection settings.

## IV. Word recognition system description

Our word recognition system is based on HMMs. The system used in this paper has a similar architecture to the one presented in [2]. We can actually note that the baseline detection system presented in Section II-B shares a similar training-testing architecture, the only difference being in the fact that HMMs are here used instead of GMMs.

## V. Experimental results

In order to investigate the effectiveness of our baseline detection algorithm, series of tests were performed using four experimental settings (see Figure 1): (1) "Oracle" system where we know a priori the baseline position from the ground truth, (2) Blind fixed height normalization without baseline identification, (3) Standard horizontal projection baseline detection and (4) GMMs computing probable baseline regions followed by horizontal projection. To evaluate the different systems, we used the APTI database [10].

**Evaluation of the baseline identification system**. The GMMs for the baseline region identification system was trained using a subset of 2076 word images for each size in APTI. A test was performed using 1000 word images for each size. The objective of this test was to evaluate the baseline estimation performance. A correct baseline region recognition rate of 99.4 % has been measured, showing the good capacity of the system to identify baseline regions.

**Evaluation of the word recognition system**. The word recognition system based on HMMs was trained using 75'750 and tested with 18'868 different images with a size of 24 and using the "Arabic Transparent" font. All results are presented in Table I.The best word recognition rate (WRR) is 99.3 % with the "Oracle" system (1), trained and tested using generated images with a priori controlled height and baseline positions. This good result is the upper bound value assuming a perfect baseline identification condition. The "blind" system (2) is based on fixed height normalization (30 pixels) without baseline identification. The obtained WRR is rather low with 89.2 %. The standard system (3) is using a classical horizontal projection baseline detection procedure. The WRR is raising up to 96.7 % showing the importance of normalizing the images according to the estimated baseline.

Table I
Word and character recognition results

| System | WRR | CRR | System | WRR | CRR |
|---|---|---|---|---|---|
| Oracle system (1) | 99.3 | 99.9 | Standard system (3) | 96.7 | 99.0 |
| Blind system (2) | 89.2 | 99.1 | GMM based system (4) | 98.6 | 99.6 |

The last system (4) is using our new approach consisting of recognizing probable baseline regions and refining the final baseline position with horizontal projection. The obtained results show the best performance with 98.6 % WRR.

## VI. Conclusions

We proposed in this paper a novel baseline identification approach based on GMM models able to identify probable baseline regions from local features. Experiments carried on the large APTI database reported a significant gain of performance in comparison to blind normalization and to classical baseline identification using simple horizontal projection. The recognition results are actually getting close to an *Oracle* system where the baseline is a priori known.

## References

[1] S. Rashid, F. Shafait, and T. Breuel, "An evaluation of hmm-based techniques for the recognition of screen rendered text," in *ICDAR*, sept. 2011, pp. 1260 –1264.

[2] F. Slimane, R. Ingold, A. M. Alimi, and J. Hennebert, "Duration models for arabic text recognition using hidden markov models," *CIMCA*, pp. 838–843, 2008.

[3] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, "Impact of character models choice on arabic text recognition performance," *ICFHR*, vol. 0, pp. 670–675, 2010.

[4] F. Slimane, S. Kanoun, A. M. Alimi, R. Ingold, and J. Hennebert, "Gaussian mixture models for arabic font recognition," *ICPR*, pp. 2174–2177, 2010.

[5] A. M. AL-Shatnawi, S. AL-Salaimeh, F. H. AL-Zawaideh, and K. Omar, "Offline arabic text recognition - an overview," *WCSIT*, vol. 1, pp. 184–192, 2011.

[6] H. Boukerma and N. Farah, "A novel arabic baseline estimation algorithm based on sub-words treatment," in *ICFHR*, nov. 2010, pp. 335 –338.

[7] H. E. Abed and V. Margner, "Comparison of different preprocessing and feature extraction methods for offline recognition of handwritten arabicwords," *ICDAR*, pp. 974–978, 2007.

[8] J. H. AlKhateeb, J. Jiang, J. Ren, and S. Ipson, "Interactive knowledge discovery for baseline estimation and word segmentation in handwritten arabic text," *Recent advances in technologies*, 2009.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[10] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, "A new arabic printed text image database and evaluation protocols," *ICDAR*, pp. 946–950, 2009.