# Text Normalization Method for Arabic Handwritten Script

**Tarik Abu-Ain[1], Siti Norul Huda Sheikh Abdullah[1], Khairuddin Omar[1], Ashraf Abu-Ein[2], Bilal Bataineh[1] & Waleed Abu-Ain[1]**

[1]Pattern Recognition Research Group, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia.
[2]Computer Engineering Department, Al-Balqa' Applied University, Faculty of Engineering Technology, 15008 Amman, 11134 Jordan.
Email: tabuain@siswa.ukm.edu.my

**Abstract.** Text normalization is an important technique in document image analysis and recognition. It consists of many preprocessing stages, which include slope correction, text padding, skew correction, and straight the writing line. In this side, text normalization has an important role in many procedures such as text segmentation, feature extraction and characters recognition. In the present article, a new method for text baseline detection, straightening, and slant correction for Arabic handwritten texts is proposed. The method comprises a set of sequential steps: first components segmentation is done followed by components text thinning; then, the direction features of the skeletons are extracted, and the candidate baseline regions are determined. After that, selection of the correct baseline region is done, and finally, the baselines of all components are aligned with the writing line. The experiments are conducted on IFN/ENIT benchmark Arabic dataset. The results show that the proposed method has a promising and encouraging performance.

**Keywords:** *Arabic handwriting script; baseline detection; handwritten text normalization; preprocessing; slant correction; sub-word extraction.*

## 1 Introduction

For languages written horizontally such as the Arabic language (see Table 1), the text line virtually splits into three regions: upper, middle and lower regions [1]-[2]. The upper region contains upper dots, upper diacritic, and ascenders. The middle or the baseline region is the main part of the characters, where the loops and connection points between characters lay. Finally, the lower region contains descenders, lower dots and lower diacritic [3]-[4].

In Arabic-text, writing baseline is defined as an unreal horizontal straight line, where all characters lay and join in a specific part of each character [5]. In Arabic printed texts, the baseline is detected ideally by using the horizontal projection histogram. This is done by finding the row that contains the highest

Received September 23rd, 2013, Revision October 9th, 2013, Accepted for publication November 28th, 2013.
 ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2013.7.2.5

number of foreground pixels as shown in Figure 1(a). Nevertheless, there are problems in using this method on the handwritten text such as the existence of a wide variety of free writing styles and irregularity in sub-words alignment. This problem appears when one of these six letters "د،ذ،ر،ز،و،أ" is located in the beginning or middle of the word, leading to an inaccurate detection of a straight baseline for the text, as shown in Figure 1(b).This contradicts the definition of baseline as stated earlier in this paragraph. Slant problem occurs if the vertical primitives of the text are standing in slant form on the text baseline. The text slant existence affects directly in many processes in document image analysis especially in character segmentation process that use the vertical projection histogram.

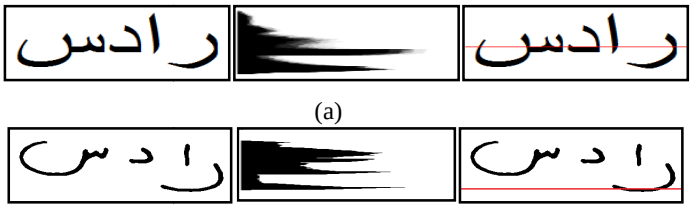
**Table 1** Arabic Letters and Shapes in the Position of a Text.

| Beginning | Middle | End | Isolated |
|---|---|---|---|
| - | - | ـأ | أ |
| بـ | ـبـ | ـب | ب |
| تـ | ـتـ | ـت | ت |
| ثـ | ـثـ | ـث | ث |
| جـ | ـجـ | ـج | ج |
| حـ | ـحـ | ـح | ح |
| خـ | ـخـ | ـخ | خ |
| - | - | ـد | د |
| - | - | ـذ | ذ |
| - | - | ـر | ر |
| - | - | ـز | ز |
| سـ | ـسـ | ـس | س |
| شـ | ـشـ | ـش | ش |
| صـ | ـصـ | ـص | ص |

| Beginning | Middle | End | Isolated |
|---|---|---|---|
| ضـ | ـضـ | ـض | ض |
| طـ | ـطـ | ـط | ط |
| ظـ | ـظـ | ـظ | ظ |
| عـ | ـعـ | ـع | ع |
| غـ | ـغـ | ـغ | غ |
| فـ | ـفـ | ـف | ف |
| قـ | ـقـ | ـق | ق |
| كـ | ـكـ | ـك | ك |
| لـ | ـلـ | ـل | ل |
| مـ | ـمـ | ـم | م |
| نـ | ـنـ | ـن | ن |
| هـ | ـهـ | ـه | ه |
| - | - | ـو | و |
| يـ | ـيـ | ـي | ي |

In this work, a text normalization method for baseline detection and slant correction in handwritten Arabic-text for character segmentation purposes is proposed. The method consists of several steps, which are: components labeling and segmentation, components text thinning, skeletons direction features extraction, candidate baselines regions determination, correct baseline region selection, components baselines alignment and slant correction. Set of experiments are conducted on a benchmark database for Arabic handwritten text called IFN/ENIT. The result is compared with a set of well-known methods such as Pechwitz [6], Farooq [7], Boukerma [8] and Horizontal Projection Histogram [9]-[11]. The results of the proposed method show a better performance than the previous methods. This paper consists of five sections:

introduction to the study, the relevant literature, the proposed framework, findings of the study, and finally the conclusion and recommendations for future studies.



(a)

(b)

**Figure 1** The Result of Using the Horizontal Projection Histogram Method [9]-[11] in Baseline Detection for Arabic Text: (a) Success on a printed text, (b) Failure on a handwritten text.

## 2 Related Studies

Pechwitz proposed a study on baseline detection method based on polygonally approximated skeleton processing [6]. A study conducted by Farooq proposed another method that uses a two-step linear regression, which is applied on the local minimal points of word contour [7]. This latter method proved to be a better one compared to the previous study done by Pechwitz. However, the results of both of these methods contradict the baseline definition. The problem with these methods is that the linear regression algorithm does not work well on unaligned text.

In another study, Boukerma proposed an algorithm method that uses a piece-wise painting scheme to estimate the baseline by identifying a set of points used in the estimation process [8]. However, this method is defective in cases where large diacritics and small characters exist. On the other hand, Ziaratban *et. al*. proposed a baseline estimation algorithm using a template matching and a polynomial fitting algorithm [12]. However, this method is less effective in texts containing short words, dots and diacritics.

A study conducted by Boubaker proposed a baseline detection method for thinning text [13]. This is done by finding the relation between the text's point of alignment and their trajectory neighbor directions. However, the algorithm is not effective on short words that consist of only isolated characters. Nagabhushan, *et al*. proposed an algorithm method that uses a piece-wise painting scheme to identify points that are used to estimate the baseline [14]. However, this algorithm method is less effective on large diacritics and small characters.

From the above studies, it is obvious that most of these methods are influenced by many factors such as diacritics, isolated characters, long words as well as binding points between the characters.

## 3 The Proposed Method

The detection process of baseline location is very useful in extracting accurate information such as writing directions, ascenders, descenders, dots and diacritics. Irregularities in Arabic script handwriting style lead to irregularity in the writing of straight text components. A straight baseline detection and text slant correction are crucial steps in the pre-processing stage as a text normalization process.

As discussed above, most methods did not detect the correct baseline in texts consisting of short characters as well as of large diacritics. To overcome this problem, a framework is proposed here to estimate a baseline for each sub-word separately. Then, the estimated baselines are used to estimate a straight writing line for the whole text. Figure 2 depicts the proposed framework.

The process steps of the proposed text normalization are as follows:

**Step 1:** *Binary the text image.* This is done using Eq. (1) from the method proposed in [15]-[16] (Figure 3(a)) as follows:

$$T_w = m_w - \frac{{m_w}^2 * \sigma_w}{(m_g + \sigma_w)(\sigma_{Adaptive} + \sigma_w)} \quad (1)$$

where $T_W$ is the thresholding value of the binarization window, $m_w$ is the mean value of the pixels in the window; $m_g$ is the mean value of the global image pixels; $\sigma_{Adaptive}$ is the adaptive standard deviation for the window; $\sigma_w$ is the standard deviation of the window.

**Step 2:** *Extract the connected components (sub-words) from the text image.* This is carried out using the method proposed in [17] (Figure 3(b)).

**Step 3:** *Keep the main sub-words body only.* This is important since the presence of noise and dots affect the process of the baseline detection. This step works as the following; for components with the size of less than a threshold value ($T_1$) when calculated by the Eq. (2), they will be removed.

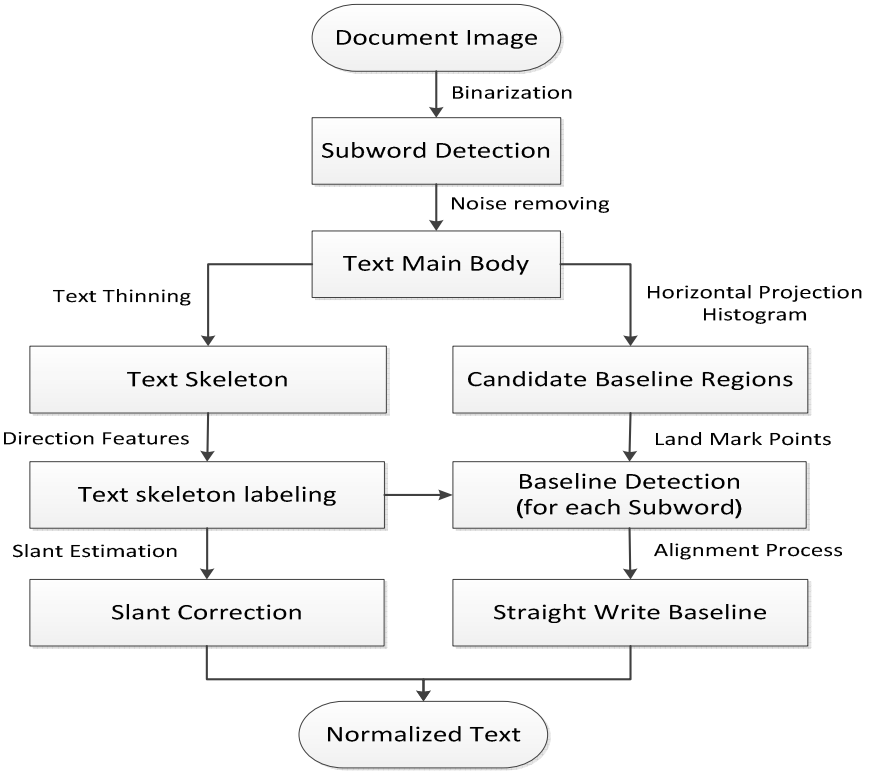



**Figure 2** The Proposed Framework for the Normalization of Handwritten Arabic Text.

$$T_1 = \left(\frac{\sum black\ pixels}{number\ of\ connected\ components}\right) / V_e \ . \quad (2)$$

where $3<V_e<4$. $V_e$ is a non-constant value that depends on the size of noise, dots and diacritics.

**Step 4:** *Text skeleton:* A robust text-thinning algorithm [18] is applied to ensure that the size of the width of the text skeleton is exactly one-pixel, as shown in Figure 3(c).

**Step 5:** *Text skeleton pixels analysis:* Perform a set of direction features [19] on the text skeleton to detect the horizontal and vertical adjacent pixels to draw a closed shape, where each one of them takes a unique label $L_H$, $L_V$ and $L_C$, respectively as shown in Figure 3(d) as follows:

- Closed shapes are detected using connected component technique [17] and are labelled by $L_C$.
- The rest of pixels are labelled based on their relationship with the 8-niebour adjacent pixels.
  - All pixels that are adjacent horizontally are labelled with $L_H$.
  - All pixels that are adjacent vertically are labelled with $L_V$.
  - Remaining pixels are labelled with $L_H$ or $L_V$ that are labelled based on their neighbour labelled pixels.

**Step 6:** *Baseline detection and straightening*

- The steps of baseline detection and straightening is as given below: for each sub-word,
- The horizontal projection histogram [9]-[11] is applied on each sub-word (see Figure 3(e)).
- Calculate the threshold value "T3", which is equal to the mean value of all the black pixels (see the vertical line in Figure 3(e)).
- Detect the candidates' baseline regions where every region is a set of continuous rows of foreground pixels that exceeds the threshold value T3 (see Figure 3(f)).
- Assign the landmark points to the pixels that gather between two different labels (see Figure 3(d)).
- The region that has the most number of landmarks will contain the baseline, which is the row that has the highest number of foreground pixels in the original text image (see Figure 3(g)).
- Because of the existence of many sub-words in the same text line, there will be a baseline for each of them, which are called local baselines. All these local baselines are aligned into one straight line called a global baseline. The exception will occur in cases of overlapping sub-words which are treated as a special case in the following process; measure the vertical distance between each of the overlapped sub-words' baseline and the nearest un-overlapped sub-words' baseline as shown in Figure 3 (h); the sub-word that has the shorter distance is aligned onto the global baseline (see Figure 3(i)).

**Step 7:** *Slant correction*

Here, we designate how the text skeleton-based representation of an Arabic text provides a considerable role in slant correction proposed algorithm. Figure 4 illuminates the proposed method for slant correction of an Arabic-text.

For each sub-word (see Figure 4(a)), do the following:

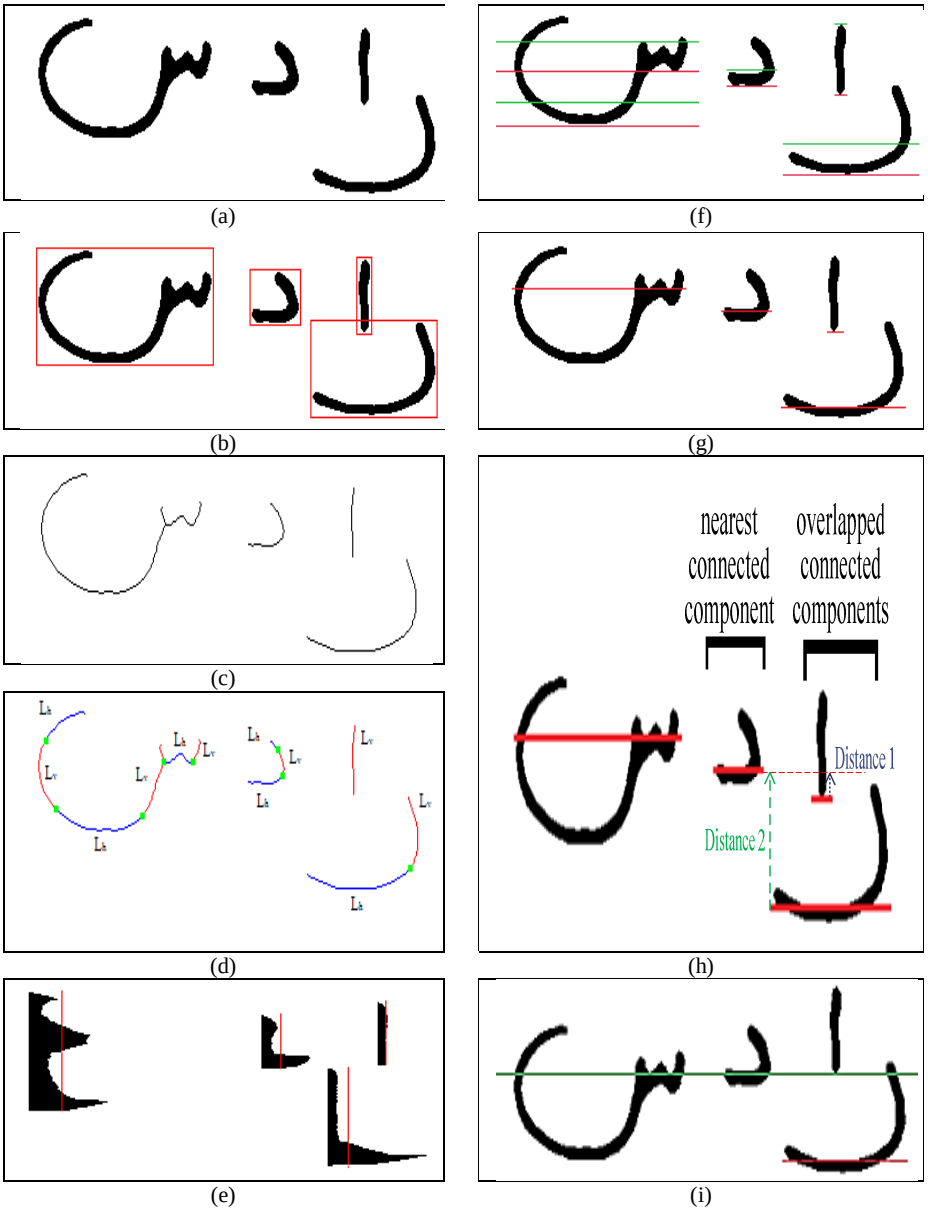
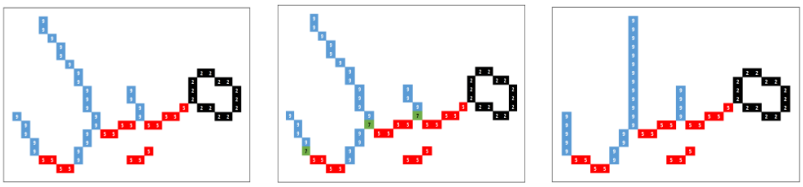



**Figure 3** An Example of a Baseline Detection and Straightening Results of a Handwritten Text Image (رادس) After Completing the First Part of the Proposed Method.

- Look for the pixels labelled as "$L_V$" that have an only one 8-neighbour foreground pixel.
- For each pixel from the previous step:
  - Trace its 8-neighbour's pixels until it reaches to none "$L_V$" labelled pixel or it reaches to the end, re-mark as base-pixel (see Figure 4(b)).
  - Shift all traced pixels labelled as "$L_V$" straight over the base-pixel (see Figure 4(c)).



(a) (b) (c)

**Figure 4** An Example of a Slant Correction Result of a Handwritten Arabic Text Image (ميل) After Completing the Second Part of the Proposed Method.

## 4 Experiments and Results

The initial results are very promising in solving many important problems that arise in baseline detection for both handwritten and machine printed cursive scripts such as in the case of diacritics, isolated characters and short words as well as the binding points between characters that lay on the baseline.

Due to the unconstrained nature of Arabic handwriting style from person to person, there is no ideal position of the handwritten text baseline (see Figure 5 (f-j)). In contrast with machine printed texts, the baseline is detected ideally using the horizontal projection profile method (see Figure 5 (a-e)).

Many experiments are conducted on the IFN/ENIT dataset [20] to validate the abilities of the proposed framework. To evaluate the proposed method performance, the results of the proposed method are compared visually with the results of Pechwitz [6], Farooq [7], Boukerma [8] (see Figure 6). Based on Figure 6, the proposed algorithm method shows better results. The proposed method is also capable to process machine printed texts as well as handwritten texts without any modification.

This study hints about various crucial problems that arise in baseline detection in handwritten Arabic text and their solutions. The problems of baseline detection for handwritten text have been largely solved in the proposed method such as in the case of diacritics, isolated characters and short words in binding the points between the characters that lay over the baseline accurately.

Moreover, all the sub-words are also intersecting on the straight baseline on the right points.

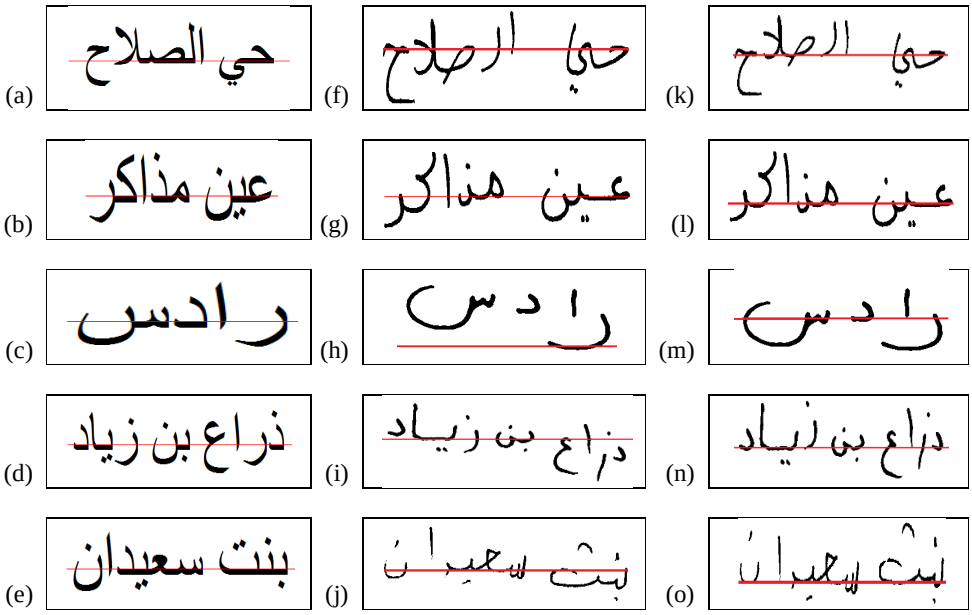



**Figure 5** Results of Baseline Detection Using: (a-e) the Ideal Baseline Obtained from the Machine Printed Text, (f-j) the Horizontal Projection Histogram Method [9]-[11] and (k-o) the Proposed Method on Handwritten Text.

## 5 Conclusion

The objective of this paper is a proposed text normalization method for handwritten Arabic script. The paper discusses seven main stages involved in this method; connected component allocation, candidate baseline regions detection, text skeleton analysis, baseline allocating for each word/sub-word, baseline straightening procedure and lastly text slant correction for segmentation purposes. The visual results prove the remarkable performance of the proposed method on the textual binary images. The results on the IFN/ENIT dataset used in the experiments with the proposed method show a significantly better performance compared to the Pechwitz [6], Farooq [7], Boukerma [8] methods. The proposed method will enhance the performances of several techniques such as, text segmentation, feature extraction and character recognition.

**Figure 6** Results of: (a-e) the Horizontal Projection Histogram Method [16], (f-j) Pechwitz Method [6], (k-l) Farooq Method [7], (m-n) Boukerma Method [8] and (o-p) the Proposed Method.

## Acknowledgements

The authors would like to thank the Faculty of Information Science and Technology and Center for Research and Instrumentation Management of the Universiti Kebangsaan Malaysia for providing facilities and financial support under Exploration Research Grant Scheme Project No. ERGS/1/2011/STG/ UKM/01/18 entitled "Calligraphy Recognition in Jawi Manuscripts using Palaeography Concepts Based on Perception Based Model" and Fundamental Research Grant Scheme No. FRGS/1/2012/SG05/UKM/02/8 entitled "Generic Object Localization Algorithm for Image Segmentation".

## References

[1] Grimes, B.F., *Ethnologue: Languages of the world*, fourteenth ed., SIL International, 2000.

[2] Lorigo, M. & Govindaraju, V., *Offline Arabic Handwriting Recognition: A Survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **28**(5), pp. 712-724, 2006.

[3] Abu-Ain, T.A.H., Abu-Ain, W.A.H., Abdullah, S.N.H.S. & Omar, K., *Off-line Arabic Character-Based Writer Identification – a Survey*, in Proceeding of the International Conference on Advanced Science, Engineering and Information Technology (ICASEIT 2011),Indonesian Students Association-Universiti Kebangsaan Malaysia, Bangi, Malaysia, pp. 161-166, 2011.

[4] Abu-Ain, T., Abdullah, S.N.H.S., Bataineh, B., Abu-Ain, W. & Omar, K., *Text Normalization Framework for Handwritten Cursive Languages by Detection and Straightness the Writing Baseline*, in International Conference on Electrical Engineering and Informatics (ICEEI 2013), UKM, Bangi, Selangor,Malaysia. pp. 654-658, 2013.

[5] Gacek, A., *ArabicManuscripts: A Vademecum for Readers*, BRILL 2009.

[6] Pechwitz, M. & Margner,V., *Baseline Estimation for Arabic Handwritten Words*, in Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR 2002), Niagara-on-the-Lake, Ontario, Canada, pp. 479-484, 2002.

[7] Farooq, F., Govindaraju, V. & Perrone, M., *Pre-Processing Methods for Handwritten Arabic Documents*, in Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR 2005), Seoul, Korea, pp. 267-271, 2005.

[8] Boukerma, H. & Farah, N., *A Novel Arabic Baseline Estimation Algorithm Based on Sub-Words Treatment*, in Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR 2010), Kolkata, India, pp. 335-338, 2010.

[9] Parhami, B. & Taraghi, M., *Automatic Recognition of Printed Farsi Texts*, Pattern Recognition, **14(**1-6), pp. 395-403, 1981.

[10] Saady, Y.E., Rachidi, A., El Yassa, M. & Driss, M., *Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character*, International Journal of Advanced Science and Technology, **33**(17), pp. 33-50, 2011.

[11] Touj, S., Amara, N.B. & Amiri, H., *Arabic Handwritten Words Recognition Based on a Plannar Hidden Markov Model*, The International Arab Journal of Information Technology, **2**(4), pp. 318-325, 2005.

[12] Ziaratban, M. & Faez, K., *A Novel Two-Stage Algorithm for Baseline Estimation and Correction in Farsi and Arabic Handwritten Text Line*, in 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, Florida, USA, pp. 1-5, 2008.

[13] Boubaker, H., Kherallah, M. & Alimi, A.M., *New Algorithm of Straight or Curved Baseline Detection for Short Arabic Handwritten Writing,* in Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, pp. 778-782, 2009.

[14] Nagabhushan, P. & Alaei, A., *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, **2**(4), pp. 907-916, July 2010.

[15] Bataineh, B., Abdullah, S.N.H.S. & Omar, K., *An Adaptive Local Binarization Method for Document Images Based on A Novel Thresholding Method and Dynamic Windows*, Pattern Recognition Letters, **32**(14), pp. 1805-1813, 2011.

[16] Bataineh, B., Abdullah, S.N.H.S., Omar, K. & Faidzul, M., *Adaptive Thresholding Methods for Documents Image Binarization*, in Pattern Recognition, ed: Springer Berlin Heidelberg, pp. 230-239, 2011.

[17] Linda, G. & Shapiro, G.C.S., *Computer Vision*: Prentice Hall, p. 608, 2002.

[18] Abu-Ain, W., Abdullah, S.N.H.S., Bataineh, B., Abu-Ain, T. & Omar, K., *Skeletonization Algorithm for Binary Images*, in International Conference on Electrical Engineering and Informatics (ICEEI 2013), UKM, Bangi, Selangor, Malaysia. pp. 690-694, 2013.

[19] Al Hamad, H. & Abu Zitar, R., *Development of an Efficient Neural-Based Segmentation Technique for Arabic Handwriting Recognition.* Pattern Recognition, **43**(8), pp. 2773–2798, 2010.

[20] Pechwitz, M., Maddouri, S.S., Märgner, V., Ellouze, N. & Amiri, H., IFN/ENIT - Database of Arabic Handwritten words, in Proceedings of Colloque international francophone surl'écrit et le document (CIFED 2002), Hammamet, Tunisie, pp. 129-136, 2002.