An Arabic Baseline Estimation Method Based on Feature Points Extraction

Arwa AL-Khatatneh, Sakinah Ali Pitchay and Musab Al-qudah

***Abstract*—Baseline estimation is an important pre-processing step for many methods in optical character recognition system such as character segmentation, detection and correction of skew and feature extraction methods. In this paper, a new baseline estimation method for Arabic handwritten text is proposed based on analysis and extraction the feature points from the subword of the thinned text. A comprehensive set of experimental results using IFN/ENIT database which is specifically for Arabic optical character recognition handwritten demonstrate the efficiency of the proposed method in overcome the failure and weaknesses in the existing methods.**

***Index Terms*— Baseline estimation, preprocessing, Arabic, feature points**

I. Introduction

THE goal of the optical character recognition (OCR) systems is to recognize the input data in form of handwritten or printed text images into electronic form with high accuracy. The Arabic is the most difficult scripts which their systems present variants of Latin OCR. However, the nature of Arabic written present specific difficulties which make the recognition is a challenging task.

In Arabic script the baseline is the virtual imaginary line which joins all the characters from a specific part [1, 2] and contains valuable information about the points that connect the characters, the position of ascender/descender and the location of diacritics in the text.

Manuscript received July 19, 2016; revised Aug 10, 2016. This work was supported in part by the Ministry of Higher Education (MOHE) Malaysia under Grant [USIM/RAGS/FST/36/50914]. The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM), Institute Science Islam (ISI) and Ministry of Higher Education Malaysia for the support and facilities provided.

Sakinah Ali Pitchay is with the Universiti Sains Islam Malaysia (USIM), Malaysia. Currently she is a senior lecturer in Faculty of Science and Technology and also Associate Fellow with Institute Science Islam. (corresponding author: sakinah.ali@usim.edu.my) *Member, IAENG*.

Arwa al-Khatatneh is a PhD student in Faculty of Science and Technology in Universiti Sains Islam Malaysia (USIM). (email: arwa_khatatneh@yahoo.com)

Musab Kasim Alqudah is a PhD student in Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia (UKM).

Therefore, the estimation of baseline is an important preprocessing step for many methods in Arabic optical character recognition (AOCR) system such as Arabic character segmentation of cursive or semi cursive text, detection and correction of skew, writing lines straightness, slant and slop corrections and feature extraction methods. The correct estimation of baseline strongly contributes the efficiency and reliability of the of the AOCR system accuracy [3-6]. The points that connect the characters with the baseline are called the feature points. The feature points that are extracted from text as seen in Fig. 1 are:

A. Branch point feature: is the point that connects two lines as in letter "T ".

B. Cross point feature: is the point that connects two lines as in "+ ".

There are two different conventions typically used for distinguish objects which are: 4-connected [22] or 8-connected [18] neighborhoods. However, in this work, these points identified by examining the eight neighbors of every skeleton pixel. A branch point has three black neighbors, and a cross point has four.

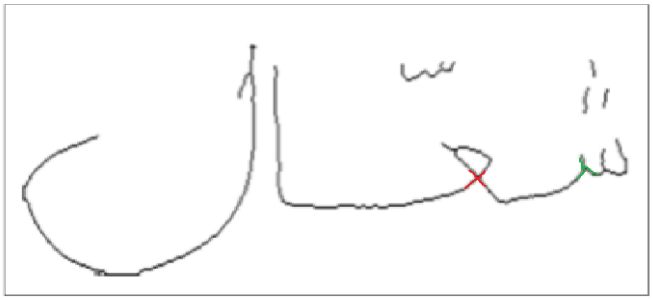


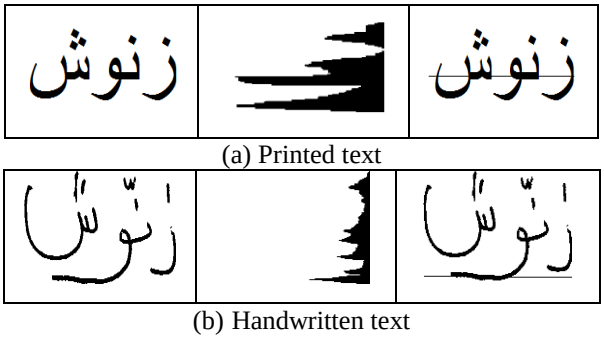
Fig. 1. Example of feature points.

The baseline in printed Arabic text can be estimated reliably using horizontal projection histogram [7] by finding the row that contains the maximum number of foreground pixels as shown in Figure 2 (a). While in handwritten Arabic scripts there are problems in using this method; the diagonally nature of handwritten text may have resulted with false baseline, the variety of font style, diacritics and dots which may lie in the range of baseline, and different of the text size and irregularity in subwords alignment as shown in Figure 2 (b).

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

These problems contradict with the definition of baseline.

These entire reasons render the baseline estimation process inaccurate and contradictory to the definition of baseline. Generally, the handwritten baseline estimation methods efficiency is dependent on a number of factors namely, the performance of the baseline search for each subword individually, the performance of the baseline search with the existence of isolated characters, and the performance of detecting the baseline for different writing style.



(a) Printed text

(b) Handwritten text

Fig. 2. Arabic text baseline estimation using the horizontal projection histogram technique.

The remainder of the paper is organised as follow. In Section II, we discuss the related works on baseline estimation with the proposed method. Section III presents the baseline estimation steps. Example results on proposed baseline estimation algorithm with comparisons to existing related works [10] and [14] are presents in Section IV. It describes the experimental setting. Finally, conclusions and future work are discussed in Section V

II. RELATED WORKS

The methods of Arabic baseline estimation have received more concern researchers in the last three decades, where the standard method used for baseline estimation is horizontal projection in 1981 by Parhami & Taraghi [8] which later improved by Timsari & Fahimi [9]. Work in [10] depends on a skeleton-based technique to estimate the baseline, their method used the skeleton lines of the polygonal word skeleton and calculate set of the feature represent the baseline relevant features. Another technique used to estimate the baseline for Arabic script is the word contour representation which proposed by [2], the method uses a two-step linear regression applied on the local minimal points of word contour. This method is better than the previous work implemented in [10].

Method in [11] used template matching and polynomial fitting algorithm to estimate the baseline. [12] proposed algorithm for thinned text by finding the relation between the text point of alignment and their channel neighbor directions. While, [13] proposed a baseline estimation algorithm using a piece-wise painting scheme to identify points that will be used to estimate the baseline. A study conducted by [14] proposed a baseline estimation method where they first remove the diacritics from text, secondly they found the branch points and cross points of word skeleton in the middle horizontal band of image, then the local baseline for each subword are detected as a line between the local minima points of the lowest outer contour of subword. Finally, they used linear regression to extract straight baseline.

Work in [15] combined enhanced horizontal projection profiles and reinforced vertical with the use of the minimum bounding box area criterion, whereas in [16] found the text skew by minimizing the area of the axis-parallel bounding box where this algorithm is script and content independent.

The proposed algorithm in [17] concentrates on the skew correction and baseline detection of Arabic documents where the skew angle is determined using a randomized Hough transform, and the baselines are extracted using y-intercept histogram.

Generally, the handwritten baseline estimation methods efficiency is dependent on a number of factors namely, the performance of the baseline search for each subword individually, the performance of the baseline search with the existence of isolated characters, and the performance of detecting the baseline for different writing style.

III. THE PROPOSED BASELINE ESTIMATION METHOD

The estimation process of correct baseline for handwritten text is difficult when the text contains subwords which are unaligned in a straight line (baseline), isolated characters or subwords of different size.

A. Pre-Baseline Estimation

Pre-Baseline Estimation of the document image is prerequisite for preparing the text for the succeeding steps. There are three steps in this stage:

- *Binarization*

The document binarization aims to increase the visibility of the useful information in the document image to extract the text while removing the noise and reducing the size of store the images in memory by removing non-useful information. This is carried out using our method in [18].

- *Connected Component*

Detection of the connected component from the text image is posed as a critical step in proposed baseline estimation for two reasons; the first one to extract the main component (sub-word) which is used to establish the baseline, and the second reason to determine the diacritics to be eliminated from the image because it reduces the performance of baseline estimation. In this research, the Connected Component Labeling (CCL) technique is employed using the method proposed in [19] as shown in Figure 3(b).

- *Diacritics Extraction*

The diacritics impose an unnecessary load on the performance of baseline estimation process. Therefore, it must be eliminated to increase the efficiency of the OCR system as in Figure 3(b). In comparison to any other isolated

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

character in Arabic scripts, the shape intensity of diacritics appears rather diminished typically [20]. It is found to be less than the threshold value TD upon the calculation of the number of foreground black pixels or text pixels of each component, and then the shape will be considered as diacritics and removed by converting its value to background pixels' value of "1".

$$T_D = (\sum_{(x,y)} I) \times 0.2 \quad (1)$$

where I (x,y) is the text pixels.

B. Baseline Estimation Process

In order to find the baseline, it is important to estimate the region that contain the baseline then detect the baseline as straight line which confirms the above definition. Based on this, the process steps of the proposed Arabic baseline estimation are as follows:

1. Detect the baseline region: in this step, the candidate region that may contain the baseline will be estimated using horizontal projection histogram (HPH) as in Fig. 3(c), where the average of black pixels intensity is calculated for each subword using (2).

$$T_S = \frac{\sum_y B(x,y)}{\sum R} \quad (2)$$

where *B(x,y)* represents the text or black pixels, and *R* is the row in image.
Any region in the horizontal histogram having a set of continuous rows of text pixels more than a threshold value T_S is considered as local candidate baseline region. Where, the local baseline for each subword detected by determines the region that has a highest number of black pixels.
2. Find textual thinning of text: A text-thinning algorithm [21] is applied to reduce text into one-pixel width skeleton for whole the text images.
3. Extract branch point features and cross point features in baseline region for each subword, while the other candidate regions are eliminated as shown in Figure 3(e).
4. Apply linear regression using these feature points, the resulted line is considered as estimated local baseline as in Figure 3(f).
5. If the subword does not contain any feature points, the baseline of this subword is inherited from horizontal nearest subwords as illustrated in Figure 3(f, g).
6. The local baseline with highest number of text pixels is considered as estimated baseline.
7. Align all subword local baselines vertically into estimated baseline.
8. In case of overlapped subwords: in some cases, the overlapping between the subwords leads to overlapping between the local baseline for subwords. Henceforth, a measurement is made to find the vertical distance between each of the overlapped subwords and the nearest un-overlapped subword local baseline. The subword having shorter distance is aligned onto the word baseline.

The example of the whole steps of the proposed Arabic baseline estimation method is presented in Fig. 3.

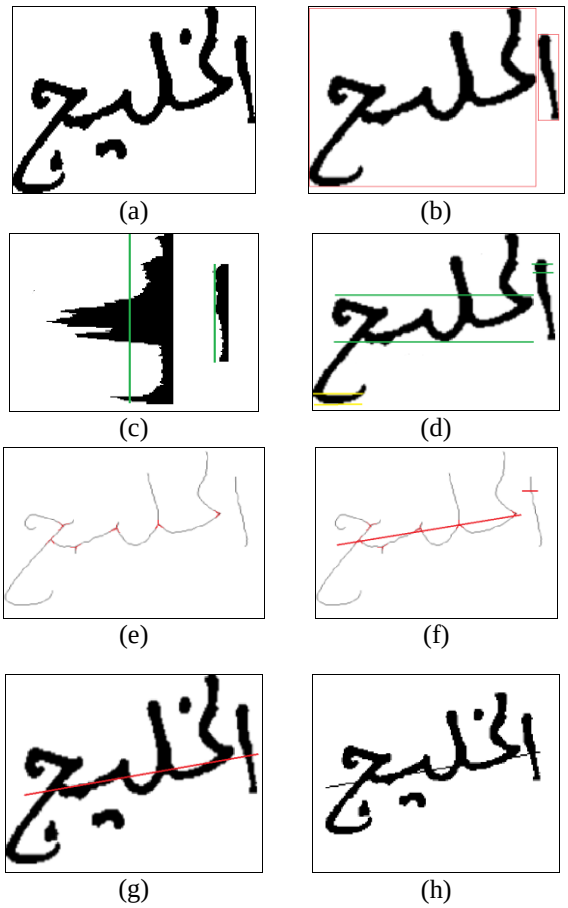


Fig. 3. An example from IFN/ENIT image number ae19_016, (a-g) steps of the proposed baseline estimation method (h) ground truth image.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed baseline estimation algorithm was tested on 115 images selected randomly from set-a of IFN/ENIT database. This database contains 26,459 handwritten Arabic images of Tunisian towns' names, with the maximum of three words, and each word consisting of one or more subwords. IFN/ENIT database provides the ground truth images for each image in database. Presently many researchers have published results on IFN/ENIT database as it is the most reliable Arabic handwritten text database. The IFN/ENIT database baseline ground-truth image is used to evaluate our baseline estimation method.

For evaluation purpose, we compare our method based on two types of experiments the visual and analytical experiments. Fig. 4 shows qualitative results of the Pechwitz [10], Boukerma [14] and proposed methods on images with different cases, where Fig. 4 represents long words with different slant angles. Fig. 5 represents case of short words with isolated characters only. The results attest to the high accuracy of the proposed method to extract baseline in case of long words with different slant angles, in case of short words with isolated characters only and the in case with long word. However, Pechwitz method is defective in the case of very short words with isolated characters only and in the case of long words with only a small region containing baseline relevant objects. Later on, Boukerma method solves

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Pechwitz problems but fail in cases when the subword contains large diacritics and small characters.

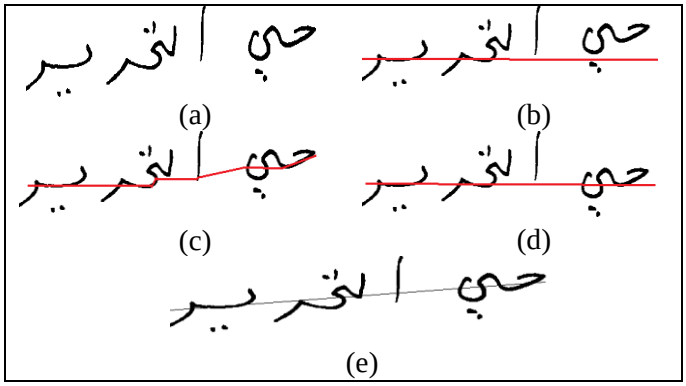


Fig. 4. Arabic handwritten baseline estimation methods result using: (a) the input image (b) Pechwitz, (c) Boukerma, and (d) the proposed method (e) ground truth images ae07_010.

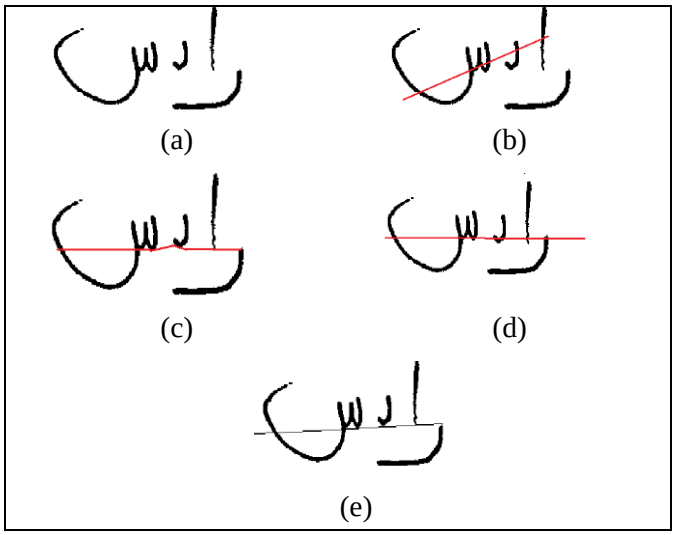


Fig. 5. Arabic handwritten baseline estimation methods result using: (a) the input image (b) Pechwitz, (c) Boukerma, and (d) the proposed method (e) ground truth images f27_13

The analytical results of the proposed method were performed to investigate the performance of our proposed method. The distance between the points in estimated baseline and the baseline ground truth image from IFN/ENIT dataset, are calculated by utilizing the rightmost and the leftmost points for the baseline. For the quality measurement the error in baseline pixels are divided into six bands (0,5,10,15,20,25). Table 1 illustrates the result of our proposed method in comparison with the existing works of Pechwitz [10] and Boukerma [14] methods.

TABLE I
RESULTS OF PECHWITZ, BOUKERMA, AND PROPOSED BASELINE ESTIMATION METHODS WITH DIFFERENT BASELINE ERROR

Baseline error in pixel	Pechwitz method	Boukerma method	Proposed method
0	0.86	1.73	2.6
5	22.6	26.95	30.43
10	48.69	60.86	62.6
15	65.21	82.61	84.34
20	79.13	89.56	90.43
25	86.08	91.3	92.17

As shown in Table 1, in the proposed method 84.34% of all cases the error is less than 15 pixels, implying considerable improvement from the previous methods, and the exact extraction according to baseline in ground truth image was 2.6 %. However, more than 15 pixels the rate will be more than 92% as the proposed method separate each of subwords using connected component labeling technique, upon the estimation process for local baseline and eliminate the supportive objects. This comparison seems to be semi-optimal since the nature of the proposed baseline and the baseline in ground truth is marginally different.

Pechwitz method based on the established linear regression of the features in polygonal approximated word skeleton. This method fails when the image contains a very short words with isolated characters only or when the image contains long words with only a small region containing baseline relevant objects. Generally, the error is less than 15 pixels in 65.21% of all cases and only achieved exact extraction was 0.86% according to baseline in ground truth image.

Boukerma method based on the established set of points on subword to estimation process. However, this method fails when the subword contains large diacritics and small characters. Generally, this method can extract 1.73% according to baseline in ground truth image and 82.61 % of all cases the error is less than 15 pixels. The case of vertical ligatures in the word as well as in the case of subwords poses a significant hindrance to this algorithm owing to the lack of individual horizontal band and thereby inherit inappropriate band from their subwords neighbours. Moreover, the failure of this method is also contributed by the wrong selection of local minima point located at bottom curve of small descenders.

Figure 6 shows the result of comparison between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error.

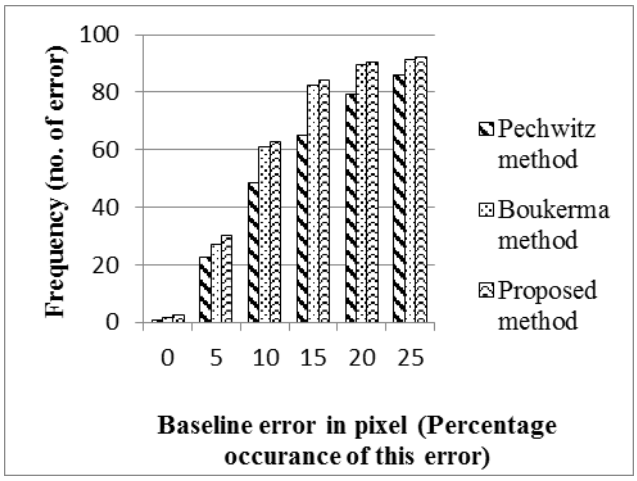


Fig. 6. Comparison between Pechwitz, Boukerma and proposed baseline estimation methods with different baseline error.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. CONCLUSION

In this paper, we proposed a novel baseline estimation method for Arabic handwritten text based on feature points. The proposed method aims to address the main challenges of Arabic handwritten baseline estimation process, and to overcome the failure and weaknesses in the existing methods. The proposed baseline estimation method establishes the baseline candidate regions using HPH first and followed by the selection of the region that containing the baseline depending on number of text pixels, after that using features points and linear regression the baseline has been estimated for each subwords. In general, the proposed baseline estimation method can be used in many methods and techniques as the main process such as text segmentation, skew normalization and features extraction. The experiment of the proposed Arabic handwritten baseline estimation method was performed on set_a of IFN/ENIT benchmark dataset images. In comparison with Pechwitz [10] and Boukerma [14] methods, the result of proposed method shows the best performance in terms of visual experiment and analytical experiments.

REFERENCES

[1] A. Gacek, Arabic manuscripts: A vademecum for readers. Vol. 98. Brill, 2009.

[2] F. Farooq, V. Govindaraju, M. Perrone, Pre-processing methods for handwritten Arabic documents, IEEE 8th Int. Conference on Document Analysis and Recognition (ICDAR'05), pp. 267-271, 2005.

[3] N. Arica, F.T. Yarman-Vural. Optical character recognition for cursive handwriting. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, no. 6, pp. 801-813, 2002.

[4] A.M. Zeki, The segmentation problem in arabic character recognition the state of the art. IEEE Int. Conference on Information and Communication Technologies, pp. 11-26, 2005.

[5] R. El-Hajj, L. Likforman-Sulem, C. Mokbel. Arabic handwriting recognition using baseline dependant features and Hidden Markov modeling. IEEE 8th International Conference on Document Analysis and Recognition (ICDAR'05), pp. 893-897, 2005.

[6] F. Lotfi, F. Nader, B. Mouldi. Arabic Word Recognition by Using Fuzzy Classifier. *Journal of Applied Sciences*. 6(3), pp. 647-650, 2006.

[7] F. Stahlberg, S. Vogel. Detecting dense foreground stripes in Arabic handwriting for accurate baseline positioning. IEEE 13th Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 361-365, 2015.

[8] B. Parhami, M. Taraghi. Automatic recognition of printed Farsi texts. 1980 Conference on Pattern Recognition, Elsevier, vol. 14, no. 1–6, pp. 395-403, 1981.

[9] B. Timsari, H. Fahimi. Morphological approach to character recognition in machine-printed Persian words, Int. Society for Optics and Photonics pp. 184-191, 1996.

[10] M. Pechwitz, V. Margner. Baseline estimation for Arabic handwritten words. Proceedings. IEEE, 8th Int. Workshop on Frontiers in Handwriting Recognition, pp. 479-484, 2002.

[11] M. Ziaratban, K. Faez. A novel two-stage algorithm for baseline estimation and correction in Farsi and Arabic handwritten text line. IEEE 19th Int. Conference on Pattern Recognition (ICPR), pp. 1-5, 2008.

[12] B. Houcine, M. Kherallah, A.M. Alimi. New algorithm of straight or curved baseline detection for short arabic handwritten writing. IEEE 10th Int. Conference on Document Analysis and Recognition, pp. 778-782, 2009.

[13] P. Nagabhushan, Alireza Alaei. Tracing and straightening the baseline in handwritten persian/arabic text-line: A new approach based on painting-technique. *Int. Journal on Computer Science and Engineering* vol. 2, no. 4, pp. 907-916, 2010.

[14] H. Boukerma, F. Nadir. A novel Arabic baseline estimation algorithm based on sub-words treatment. IEEE Int. Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 335-338, 2010.

[15] A. Papandreou, B. Gatos, S.J. Perantonis, I. Gerardis. Efficient skew detection of printed document images based on novel combination of enhanced profiles. *Int. Journal on Document Analysis and Recognition (IJDAR)* vol. 17, no. 4, pp. 433-454, 2014.

[16] S. Mahnaz, M. Sid-Ahmed. Skew detection and correction based on an axes-parallel bounding box. *Int. Journal on Document Analysis and Recognition (IJDAR)*, Springer, vol. 18, no. 1, pp. 59-71, 2015.

[17] A. Boukharouba. A new algorithm for skew correction and baseline detection based on the randomized Hough Transform. *Journal of King Saud University-Computer and Information Sciences*, 2016.

[18] A. Al-Khatatneh, S.A Pitchay. Compound binarization for degraded document images. *ARPN Journal of Engineering and Applied Sciences*, pp. 1819-6608, 2015.

[19] Linda G. Shapiro, G.C.S., Computer Vision2002: Prentice Hall, pp. 608.

[20] Fahmy, M.M.M. Haytham, El-Messiry H. Automatic recognition of typewritten Arabic characters using Zernike moments as a feature extractor. *Studies in Informatics and Control Journal*, 10, no. 3, pp. 227-236, 2001.

[21] W. Abu-Ain, S.N.N. Sheikh Abdullah, B. Bataineh, T. Abu-Ain, K.Omar, Skeletonization Algorithm for Binary Images. Procedia Technology 11, pp. 704-709, 2013.

[22] A. Kaban and S.A Pitchay, Single-frame image recovery using Pearson type VII MRF, *Journal Neurocomputing* 80, Elsevier· pp. 111-119, 2012.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)