Last Teen Pixels for Arabic Font Size and Style Recognition

https://doi.org/10.3991/ijoe.v17i12.25065

Abdelouahed Ait Ider, Said Nouri, Abdelkrim Maarir(✉)
Sultan Moulay Slimane University, Beni Mellal, Morocco
a.maarir@ya.ru

Abstract—Arabic printed script segmentation and recognition techniques change from font to other i.e. each font has particular properties calligraphic and structural which differ with other. Majority of segmentation system suffer in word or sub word segmentation into characters because they consider one algorithm to segment all kind of Arabic printed font, style and size. The goal of this work is to prepare a system of word or sub word Optical Font Arabic Recognition (OFAR) for different font size and style of Arabic printed script, in order to integrate it in global Arabic Optical Character Recognition (AOCR) to choose preferred and good segmentation algorithm. APTI database was used to extract last ten pixels for each word or sub word to build new database of last 10 pixels for each word; OFAR is based upon this new database and our extraction approach called Pixels Continuity (PC) algorithm in different matrix direction and some histogram statistics to extract 20 features. Three KNN classifiers with K=5 and three different distances using Cityblock, Euclidean and Correlation based upon majority-vote are used to evaluate the system robustness. This classifier is compared in the first time with Back propagation Neural Network and Steerable Pyramid (SP) algorithm to re cognize three font families, then in the second time with Gaussian Mixture Models (GMMs) to recognize font and size. The average recognition results obtained was 99.55% about font and size and 98.17% for font, size and style recognition.

Keywords—word, sub-word, characters, APTI, last teen pixels, pixels continuity (PC)

1 Introduction

Several Arabic script segmentation and recognition system suffer in segmentation stages [1]–[3] which adopt one algorithm to segment different text of different font, style and size into characters. This hypothesis considers that different font has the same structure and calligraphy.

In document preprocessing system [4], segmentation and recognition stages are indispensable to convert document image (.PNG, .JPEG …) to editable file (.TXT, .DOC, .DOCX), in order to find other approach to solve complexity of AOCR system

by fragment multifont segmentation and recognition problem into omnifont problem. Recently another intermediate system for font recognizing (OFAR) was proposed [5], [6] that suffer from similarity between certain fonts. Other approach based on fixed-length window sliding from right to left on the word image to estimate font category likelihoods using Gaussian Mixture Models (GMMs) [7]. Other approaches were based on stochastic approach for font and size identification evaluated on ultra-low resolution Arabic word images, which show clearly the importance and potential of font recognition followed by mono-font word recognition system. The character and word recognition error could be reduced by over 70% when using font recognition first, followed by word recognition [8], [9]. Arabic font recognition based on a priori approach using steerable pyramid with 6 orientations give high recognition rates than Gray Level Co-occurrence Matrix, Gabor Filter and Wavelets evaluated using the Arabic printed text image dataset multi-font of APTID/MF database and Bp ANN for classification [10], [11].

Several authors recognize just font and size like in [8], [10]. To keep source document styles Arabic optical font recognition must recognize font, size and also styles.

In this work we propose a new approach of optical Arabic font recognition based on last 10 pixels extraction for each word or sub-word of Arabic printed text Image APTI database of low-resolution in multiple fonts, sizes and styles with different degradation conditions [12].

The proposed system is based on three steps: preprocessing, features extraction and classification. In the preprocessing, we use the thresholding technic to convert input image into binary image. Some statistical histogram measure and pixels continuity of horizontal, vertical, diagonal and antidiagonal direction in the same matrix are calculated to extract 20 features to identify font style and size for each word or sub-word. The suggested system performance was tested using APTI database and K-Nearest Neighbor classification technical using three different distances and majority vote. The organization of this paper is as follows: In the section 2 a brief explication about font recognition system. The section 3 deals with the preprocessing step. The section 4, features extraction method is described. The Section 5 deals with the classification approach which the KNN classifier is used. In section 6, experimental results and discussion are given. The paper is ended by a conclusion and perspectives.

2 Recognition system

Figure 1 illustrates the proposed font recognition system. In the first step Arabic printed word image of APTI database was transformed into binary image then the word was localized based on horizontal and vertical histogram. In the next step twenty features are extracted from last binary image from last 10 pixels of word or sub-word to build the training dataset and to classify fonts, sizes and styles of each word using three classifiers of K Nearest Neighbor and three different distances based on majority vote.

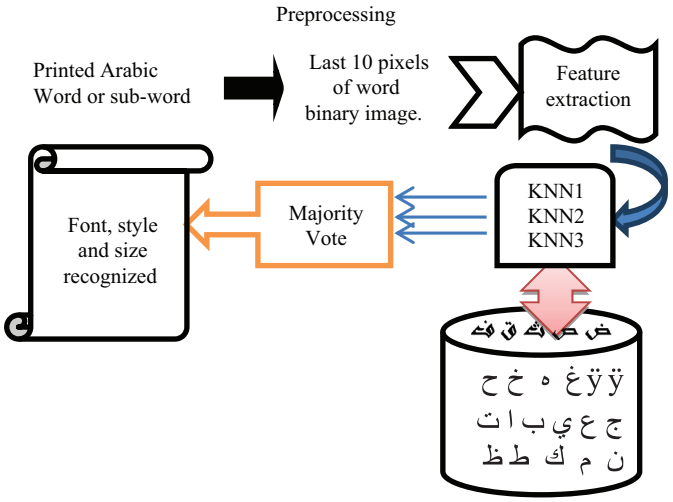

Fig. 1. Font Arabic recognition system

2.1 Font identification

In Arabic printed documents, text is presented using one or more font, styles and sizes. In [6], [8], [13] the authors consider; Font identification of some character in the paragraph is sufficient to identify dominant font in the same paragraph.

In our approach, we consider that a document or paragraph can be edited by more than one font, style and size, for this reason font detection in the paragraph was processed word by word to keep source fonts, styles and sizes. This hypothesis was tested on APTI database.

2.2 Arabic font

Arabic font is one of complex scripts written from right to left, it consists of 28 characters that contain dots below or above and each character can take four different structures Figure 2 according to position, styles and sizes. Rich styles in Arabic font calligraphy give more complex horizontal and vertical overlapping as shown in Figure 2 and 3.

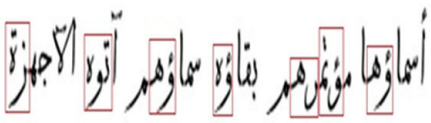

Fig. 2. Horizontal overlapping in Arabic word from APTI database

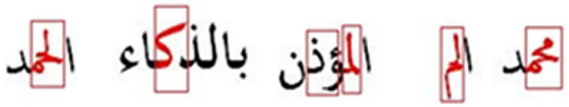

Fig. 3. Words presenting vertical and horizontal ligature

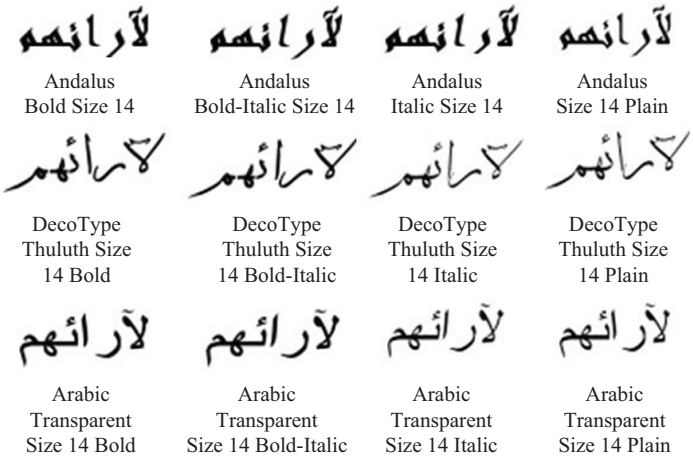

Fig. 4. An Arabic word with three fonts and four styles and size 14

Figure 4 presents the same word "لآرائهم" with three fonts Arabic Transparent, DecoType Thuluth and Andalus using different styles Bold, Bold-Italic, Italic and Plain with same size 14.

As shown in this Figure Similarity between styles of same font is more significant, this is one of the problems in font recognition process that influence in the OFAR system performance.

3 Preprocessing

Most of the styles are intuitive. However, we invite you to read carefully the brief description below in the preprocessing step input image for each word was transformed into binary image [14]–[16], then the position of word was localized based on histogram vertical and horizontal projection Figure 5.

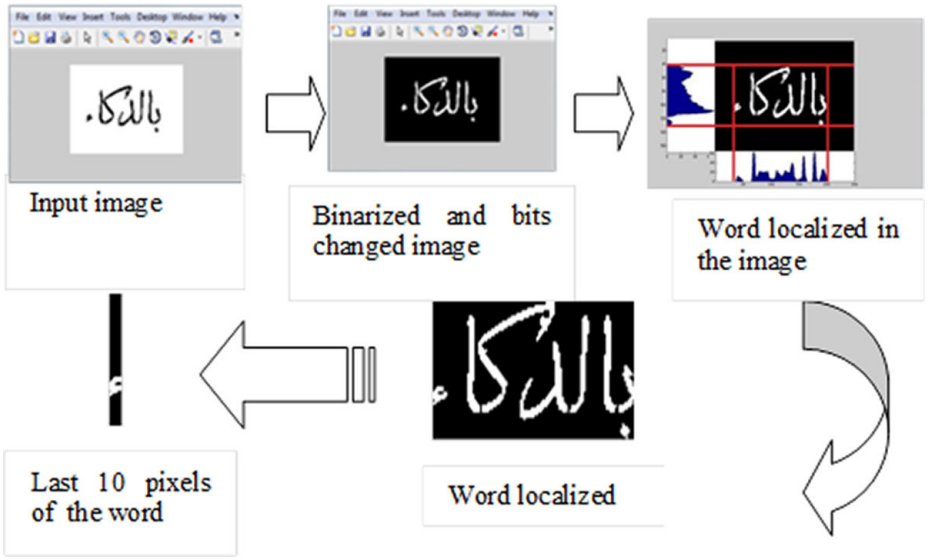

Fig. 5. Preprocessing steps

4 Feature extraction

In this study, the binary word image was represented by pixels of value are 1. Pixels Continuity (PC) technique for font feature extraction aims to give information about pixels distribution in four directions: horizontal, vertical, diagonal and antidiagonal. This technique was based on accounting the number of successive non zeros pixels. Maximum, minimum and mean of pixels length chain of four directions were calculated to obtain 12 features to characterize pixels continuity for each font. Horizontal and vertical length of continued pixels was extracted to identify Plain and Bold fonts.

Last ten pixels (Figure 6) of word or sub-word was analyzed to extract some features: minimum, maximum, mean of horizontal, vertical, diagonal and antidiagonal pixels continuity, other features were calculated based on histogram statistical. All those techniques are described below.

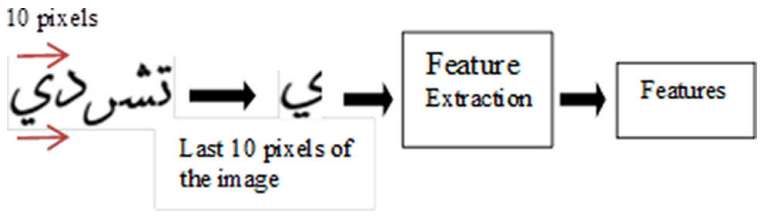

Fig. 6. Feature extraction process

If word or sub word image length after localization process is less than ten pixels the full-size length was considered as shown in Figure 7.

Fig. 7. Last ten pixels of word لأرائهم of font Arabic Transparent with size 14

4.1 Horizontal pixels continuity

To calculate pixels continuity, each raw in the matrix was analyzed to compute the length of continued pixels chain given in Figure 8.

For example, we consider the following matrix to calculate pixels continuity

Fig. 8. Horizontal pixels selection

As shown in Figure 8, in the first row there is two chains of pixels, the length of first is one pixel and second is three pixels; in the second row we have one chain with seven pixels as length and the last row present tow chains of length equal three pixels. Minimum, maximum and mean of horizontal pixels continuity was considered as features.

- Horizontal pixels continuity(M) = [1; 3; 7; 3; 3]
- Max (Horizontal pixels continuity) = 7
- Min (Horizontal pixels continuity>1) = 3
- Mean (Horizontal pixels continuity) = 3.4

4.2 Vertical pixels continuity

In this process, pixel continuity is calculated on columns to find vertical chains pixels continuity (Figure 9).

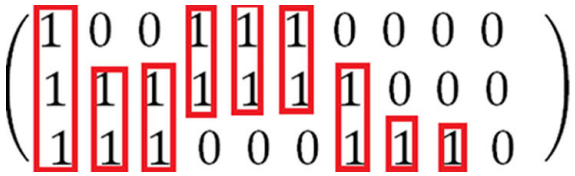

Fig. 9. Vertical pixels selection

- Vertical pixels continuity(M) = [3; 2; 2; 2; 2; 2; 2; 1; 1]
- Max (Vertical pixels continuity) = 3
- Min (Vertical pixels continuity>1) = 2
- Mean (Vertical pixels continuity) = 1.88

4.3 Diagonal pixels continuity

In these process pixels continuity is calculated on diagonal pixels to find diagonal chains length of pixels continuity as shown in Figure 10.

Fig. 10. Diagonal pixels selection

- Diagonal pixels continuity(M) = [1 2 3 1 1 2 3 3 1]
- Max (Diagonal pixels continuity) = 3
- Min (Diagonal pixels continuity>1) = 2
- Mean (Diagonal pixels continuity) = 1.87

4.4 Antidiagonal pixels continuity

This is calculated on antidiagonal pixels to calculate minimum, maximum and mean of antidiagonal pixels continuity (Figure 11).

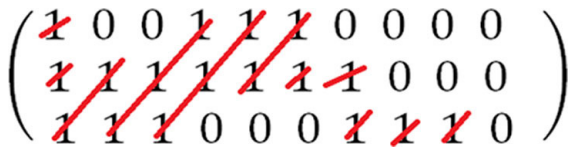

Fig. 11. Antidiagonal pixels selection

- Antidiagonal pixels continuity(M) = [1; 1; 2; 3; 3; 2; 1; 1; 1; 1; 1]
- Max (Antidiagonal pixels continuity) = 3
- Min (Antidiagonal pixels continuity>1) = 2
- Mean (Antidiagonal pixels continuity) = 1.6

Finally, we obtain a vector of 12 features to identify word or sub-word font matrix V= [7; 3; 3,4; 3; 2; 1,88; 3; 2; 1,87; 3; 2; 1,6]

4.5 Others features

Beside on the above-mentioned features, the following features are used.

- Minimum of horizontal histogram which is superior to 1.
- Maximum of horizontal histogram.
- Mean of horizontal histogram.
- Minimum of vertical histogram which is superior to 1.
- Maximum of vertical histogram.
- Mean of vertical histogram.
- White pixels density rate in last 10 pixels in word image.
- Height of last 10 pixels in word image.

5 Classification technique

Collected data from features extraction stage was ranged into training and testing dataset, each one represents one font with particular size and style for example font called "DecoType Thuluth_12_Bold-Italic" represent DecoType Thuluth font, size equal 12 and Bold-Italic as style. This class was labeled using number from 1 to 36.

Supervised classification technical was adopted to classify data using K-Nearest Neighbor classifier with three different distances and majority vote selection (Figure 13 and 14).

K Nearest Neighbor [17], [18] is one of famous supervised classification technical based on distances. Font classification was based on measuring distance between different classes and considering class that have minimum distance. In this work three metric was used: Cityblock distance, Euclidean distance and Correlation distance [19] (Figure 12).

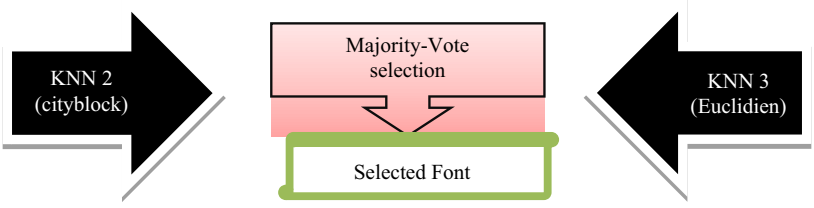

Fig. 12. Majority vote selection process

```
result1 ← knn1 result
result2 ← knn2 result
result3 ← knn3 result
Selected ← 0
    if (result3 equal result2)
            Selected ← result3
    end
    if (result2 equal result1)
            Selected ← result2
    end
    if (result1 equal result3)
            Selected ← result 1
    end
Selected   %variable presents final result of selection
```

Fig. 13. Vote majority selection algorithm

6 Experimental results and discussion

To evaluate our approach, we have considered three font families with three sizes (10, 12, and 14) and four styles: Bold, Italic, Bold-Italic and Plain style.

Table 1, illustrates number of test and training samples used from Arabic Printed Text Image APTI database to evaluate the proposed system. 3000 samples for testing and 9000 training were used for each sample 36 particulars fonts.

Table 1. Test and training dataset

Font Family	Test Samples	Training Samples
Arabic Transparent	36000	107999
Andalusia	36000	108000
DecoType Thuluth	36000	108000
Total	108000	323999

Classification results are obtained using three KNN classifications technical with three different distances and pixel continuity algorithm besides eight statistical parameters and majority vote principal to characterize each word by 20 features and classify testing data.

In our approach, font recognition process was divided into three steps: in the first step font family recognition was performed, next size recognition step and in the last, font style was identified.

6.1 Font family recognition results

We report in the Table 2, results obtained for each font family and average rate recognition using proposed algorithm Pixel Continuity (PC) besides eight statistical features (SF) and majority vote of three KNN classifiers with neighbor number K=5.

Table 2. Font family recognition rate

Size	Arabic Transparent	Andalus	DecoType Thuluth
10	99.50	99.29	99.56
12	99.60	99.70	99.72
14	99.55	99.70	99.35
Recognition rate	99.55	99.57	99.54
Total rate	99.55		

The Table 2 illustrates that, proposed system shows more performance about size 12 for three font family used.

Figure 14 illustrate the comparisons between proposed system using pixel Continuity (PC) and Statistical Features (SF) based on majority vote of three KNN and Steerable Pyramid using in the first time KNN classifier and in the second time Back propagation Neural Network [10].

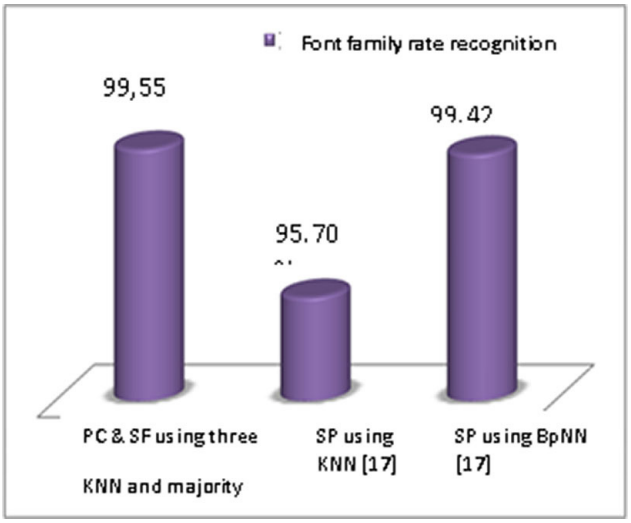

Fig. 14. Rate recognition of three font family

The obtained results about Font family recognition improve that, the proposed system gives best results than Steerable Pyramid technical used with back propagation and KNN in [10] about three font family tested, due to performance of Pixel Continuity (PC) algorithm which analyzes word image in four directions and select frequently chain length that characterize each font family.

6.2 Size recognition results

Table 3, illustrate rate recognition of font family by size obtained using Pixel Continuity and Statistical Features (SF) and three K-Nearest Neighbors with k=5 and vote majority selection. Size recognition rate was calculated for three sizes 10, 12, and 14 for each font family.

Table 3. Size recognition rate

Size	Arabic Transparent	Andalus	DecoType Thuluth	Mean Rate %	Total rate %
10	99.50	99.29	99.56	99.45	99.55
12	99.60	99.70	99.72	99.67	
14	99.55	99.70	99.35	99.53	

The analyses of results of Table 3, shows that our approach is more efficient with size 12. The total recognition rate obtained about font and size is 99.55%.

The following Figure illustrates the comparison by font and size between our approach "Pixels Continuity and Statistical Features (SF)" and Gaussian Mixture Models (GMMs) used in [8].

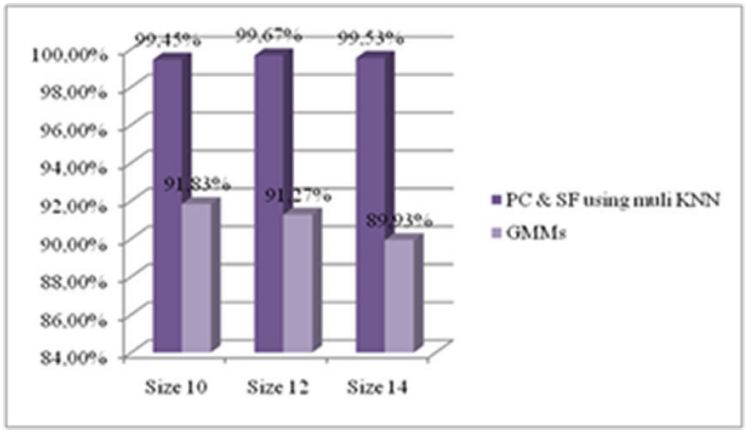

Fig. 15. Recognition rate of three Font family used by size of PC & SF using mulyi KNN compared with GMMs [8]

The analysis of the results presented in Figure 15, shows the performance of Pixel Continuity and statistical features compared to Gaussian Mixture Models due to ability of pixel continuity to estimate chain length of pixels in four directions, which changes from size to other of the same font family.

6.3 Font style and size recognition results

Figure 16 summarizes recognition rate for three fonts: Arabic Transparent, Andalus and DecoType Thuluth with four styles, Bold, Italic, Bold-Italic and plain and also three sizes10, 12 and 14. Best obtained results were 100% for Andalus font of size14 with two styles Bold and Italic.

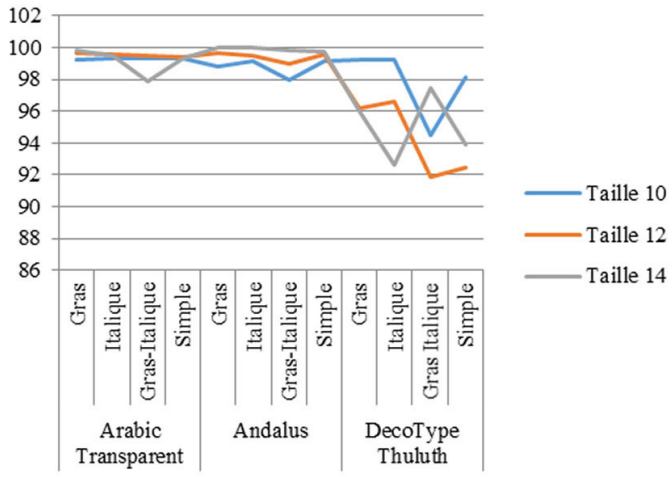

Fig. 16. Font family recognition rate by style and size

The analysis of style recognition results shows that: the proposed system recognizes Bold, italic and Plain style than Bold-Italic style, that is duo to fort similarity between styles and exactly about same size, for example: Arabic Transparent font of size 14 present more than 73% of similarity between Bold-Italic and Italic style and more than 77% of similarity between Bold and plain style, Beside of calligraphy of each font family which doesn't change significantly between style and size when font change.

7 Conclusion

Experiences are proved that omni-font system gives more performance than multi-font system, but lot of time OCAR system analyzes multi-font documents, so we need an OFAR (Arabic Optical Font Recognition). In this work, OFAR was performed and presented to resolve complexity of AOCR (Arabic Optical Character Recognition) system. New feature extraction algorithm called Pixels Continuity and 8 features histogram statistical to obtain 20 features was evaluated on 108000 samples using Dual-Core PC of 2.00 GHz, RAM of 2.00 Go and Matlab software and programming language.

100% of recognition rate for some font for example Andalus Bold of size 14 was obtained. Bold-Italic style has given less recognition rate comparing to other styles due to high similarity between styles for same font and last ten pixels for some word gives few information about font family, style and size. To resolve those particular problems, particular study for each font styles and size is necessary to build more performance OFAR.

8 Acknowledgment

In the end of this work, we give gratitude thanks for APTI database Auteurs that helped to test performance of our approach.

Publication has been made possible by financial support of the International Association of Online Engineering (IAOE).

9 References

[1] F. M. Fakir, 'Segmentation and Recognition of Arabic Printed Script', *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 2, no. 1, Art. no. 1, Mar. 2013. https://doi.org/10.11591/ij-ai.v2i1.1236

[2] M. S. Khorsheed, 'Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)', *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1563–1571, Sep. 2007, doi: https://doi.org/10.1016/j.patrec.2007.03.014

[3] B. Bushofa and M. Spann, 'Segmentation and Recognition of Printed Arabic Characters', in *Procedings of the British Machine Vision Conference 1995*, 1995, p. 54.1–54.10. doi: https://doi.org/10.5244/C.9.54

[4] D. Sporici, E. Cuşnir, and C.-A. Boiangiu, 'Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing', *Symmetry*, vol. 12, no. 5, Art. no. 5, May 2020, doi: https://doi.org/10.3390/sym12050715

[5] H. A. Al-Muhtaseb, S. A. Mahmoud, and R. S. Qahwaji, 'Recognition of off-line printed Arabic text using Hidden Markov Models', *Signal Processing*, vol. 88, no. 12, pp. 2902–2912, Dec. 2008, doi: https://doi.org/10.1016/j.sigpro.2008.06.013

[6] I. S. I. Abuhaiba, 'Arabic Font Recognition using Decision Trees Built from Common Words', *CIT. Journal of Computing and Information Technology*, vol. 13, no. 3, Art. no. 3, Oct. 2004, doi: https://doi.org/10.2498/cit.2005.03.04

[7] F. Slimane, S. Kanoun, A. M. Alimi, R. Ingold, and J. Hennebert, 'Gaussian Mixture Models for Arabic Font Recognition', in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 2174–2177. doi: https://doi.org/10.1109/ICPR.2010.532

[8] F. Slimane, S. Kanoun, J. Hennebert, A. M. Alimi, and R. Ingold, 'A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution', *Pattern Recognition Letters*, vol. 34, no. 2, pp. 209–218, Jan. 2013, doi: https://doi.org/10.1016/j.patrec.2012.09.012

[9] D. Huang and J. Gao, 'On-line Signature Verification Based on GA-SVM', *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 11, no. 6, Art. no. 6, Nov. 2015. https://doi.org/10.3991/ijoe.v11i6.5122

[10] F. J. Kallel, S. Kanoun, and V. Eglin, 'Arabic font recognition based on a texture analysis', in *ICFHR, International Conference on Frontiers in Handwriting Recognition*, Heraklion, Crète, Greece, Sep. 2014, pp. 673–677. doi: https://doi.org/10.1109/ICFHR.2014.118

[11] C. A. Oliveira Gonçalves, R. Camacho, C. T. Gonçalves, A. Seara Vieira, L. Borrajo Diz, and E. Lorenzo Iglesias, 'Classification of Full Text Biomedical Documents: Sections Importance Assessment', *Applied Sciences*, vol. 11, no. 6, Art. no. 6, Jan. 2021, doi: https://doi.org/10.3390/app11062674

[12] F. Slimane, R. Ingold, S. Kanoun, A. M. Alimi, and J. Hennebert, 'A New Arabic Printed Text Image Database and Evaluation Protocols', in *2009 10th International Conference on Document Analysis and Recognition*, Jul. 2009, pp. 946–950. doi: https://doi.org/10.1109/ICDAR.2009.155

[13] I. Abuhaiba, 'Arabic Font Recognition Based on Templates', vol. 1, no. 0, p. 7, 2003.
[14] T. Y. Kong, Ed., *Topological Algorithms for Digital Image Processing*. 2011.
[15] A. Rosenfeld and J. L. Pfaltz, 'Sequential Operations in Digital Picture Processing', *J. ACM*, vol. 13, no. 4, pp. 471–494, Oct. 1966, doi: https://doi.org/10.1145/321356.321357
[16] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, 'Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification', *Applied Sciences*, vol. 10, no. 23, Art. no. 23, Jan. 2020, doi: https://doi.org/10.3390/app10238631
[17] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice/Hall International, 1982.
[18] J. T. Tou, T.-C. Tou, and R. C. Gonzalez, *Pattern Recognition Principles*. Addison-Wesley Publishing Company, 1974.
[19] S. Nouri and M. Fakir, 'Printed Arabic Character Classification Using Cadre of Level Feature Extraction Technique', *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 2, Art. no. 2, 43/27 2013, doi: https://doi.org/10.14569/SpecialIssue.2013.030209

10 Authors

Abdelouahed Ait Ider, He received the PhD degree in Computer Science in Medical informatics and Decision Support Systems. He received his master's degree in business intelligence from Sultan Moulay Slimane University, Morocco. His ORCID iD is: https://orcid.org/0000-0003-0201-9672. Email: a.aitider@usms.ma

Said Nouri, has received PhD degree in Computer Science from the University of Sultan Moulay Slimane Beni-Mellal Morocco, on a thesis titled "Arabic printed texts automatic recognition" (2018). Laboratory of Information Processing and Decision Support, Faculty of Sciences and Technics Sultan Moulay Slimane University Beni-Mellal Morocco. His research focuses on patterne recognition, image processing, machine learning, image segmentation techniques. Email: said_maths@yahoo.fr

Abdelkrim Maarir, he received the PhD degree in Computer Science from Sultan Moulay Slimane University, Morocco, and MS degree in Business Intelligence from the University Sultan Moulay Slimane in 2013, Dr. Abdelkrim is currently a temporary professor in the Higher School of Technology, Sultan Moulay Slimane University. His research is involved in Image processing, Artificial intelligence, Data base management, etc. Email: a.maarir@ya.ru

Article submitted 2021-06-25. Resubmitted 2021-08-01. Final acceptance 2021-08-03. Final version published as submitted by the authors.