ISSN: 2706-9087

ISSUE 8 DESEMBER 2019

مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences

العدد 8 2019 ديسمبر

Arabic Language Character Recognition using Walsh-Hadamard Transform (WHT) vs. Discrete Fourier Transform (DFT)

EmadEddin Gamati **Anisa F. Elbokhare** **Amani M.Soliman**

Computer Science Department, Faculty of Education, University of Tripoli
E.gamati@uot.edu.ly *A.elbokhare@uot.edu.ly* *A.soliman@uot.edu.ly*

Abstract:

One of the common used methods for text recognition (especially with Arabic text), is the usage of character Databases for driving the training and validation (for all the different methods that are used for preprocessing, segmentation and recognition). There are no inclusive and dependable databases for all Arabic letters particularly when considering the four different shapes for each Arabic character (based on the character position inside the word).

In [1], the researchers presented a new Arabic Optical Character Recognition "AOCR" approach called "sliding window for printed AOCR" method (segmentation-free character recognition independent of a lexicon of words). It works based on matching the content of the targeted text image/document with a small pre-prepared database to find the positions of the recognized characters in the scanned image. The AOCR experiment is implemented using WHT/DCT and is applied using three different font types and nine different font sizes. In this paper, we tested the same proposed "AOCR" method using a different implementation (WHT/DFT).

Keywords— Arabic OCR, word extraction, pattern recognition, segmentation, segmentation-free, printed AOCR, sliding window, WHT, DCT.

Introduction:

The automation process development of 'extracting text from images' (called Optical Character Recognition (OCR) techniques/systems) has dominated the attention of researchers for its importance and applicability. Furthermore, The OCR Systems provide human-machine interaction, which is widely used and most needed in many daily applications involving converting a text image into an editable text format (automate the processing of: postal mail sorting, check/tag processing, restore old documents …).

Currently, the most advanced and accurate OCR systems and research have been done on the characters of the Latin languages, whereas other languages relatively still need some work (depending on the language complexity and the challenging in terms of

ISSN: 2706-9087
ISSUE 8 DESEMBER 2019
مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences
العدد 8 2019 ديسمبر

writing methodologies). These complexities and challenges could be overcome if hand-writing character recognition was considered.

Related Work:

In [1], Walsh Hadamard Transformation (WHT) and Discrete Cosine Transformation (DCT) have been used for the character image feature extraction to carry out a performance comparison of segmentation-free printed Arabic character recognition techniques. The comparison was based on the required time for recognition. The researchers used an Arabic character image database (three font types for each character, nine font sizes for each font type). The results shown for their proposed method of letter recognition was 99% for all printed Arabic characters (regards the used size and font type of the characters and its position inside the word). In addition, recognition time/speed using WHT was faster than DFT.

In [2], the researchers used Artificial Neural Networks techniques for automatic recognition of Arabic printed text. For testing purposes, the system experiments implemented different models and sizes of Naskh font. It showed stability, good performance and up to 73% recognition rate. In addition, a huge diversity in the respective image extracted features results between segmentation module and the training set in.

The presented system in [3] makes use of a hybrid of several holistic features that combine global word level Discrete Cosine Transform (DCT) based features and local block based features. The method has been tested using different sets of 1152 words with three different fonts and four font sizes and has achieved 99.3% Word Recognition Rate (WRR). It also has been tested using sets of 2730 words of recent computerized text and has attained more than 84.8% WRR.

In [4], the researchers started their approach by binarized images using global binarization. Then, by using horizontal and vertical transition techniques, they extracted characters' statistical features. Afterward, they applied K-Nearest Neighbor (KNN) and Random Forest Tree (RFT) to recognize printed Arabic characters. The results of their work show that KNN is the fastest, but RFT perfremed better. In terms of recognition rate, RFT (with 98% rate) is better than KNN (with 87% rate).

The used techniques in the proposed approach to automatic printed Arabic text recognition by the researchers in [5] were linear and ellipse regression techniques. The researchers started by collecting all possible forms of each character, then, they generated a unique code to represent each character form (Each code contains a sequence of lines and ellipses). The results indicate 86% recognition rate in a sample of 14,000 different Arabic words with different fonts.

ISSN: 2706-9087
ISSUE 8 DESEMBER 2019
مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences
العدد 8 ديسمبر 2019

Arabic Optical Character Recognition (Ar-OCR)

Arabic characters are considered one of the most difficult and challenging characters to be recognized using OCR technologies. Because of the writing rules of the language, the different shapes (up to 4 shapes/character) of the same character (based on its location inside the word) and the extra marks (up to 4 marks) which can be used over/under each character.

Challenges and motivation

Here are some of the challenges of dealing with the Arabic language scripts and its characters in OCR processing compared with other languages: ❶ The Arabic scripts are written from right to left. ❷ The language is composed of 28 Arabic characters. ❸ Each character can have up to four different shapes depending on its position in the word (isolated, beginning, middle, or end forms). ❹ Arabic characters (inside the word) are connected along a baseline. ❺ The total number of shapes for Arabic characters (28 character) reaches 100 (Table-I).❻ A pair of characters may be combined to form another character (called a ligature).

Table I: The four possible shapes/forms for Arabic characters

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

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

Main Feature (Baseline)

Arabic characters (inside a word) are connected along a baseline. The position of the baseline can appear clearly with horizontal projection, while vertical projection can be used for segmenting sub-words and characters, as illustrated in Figure1.

ISSN: 2706-9087
ISSUE 8 DESEMBER 2019
مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences
العدد 8 ديسمبر 2019

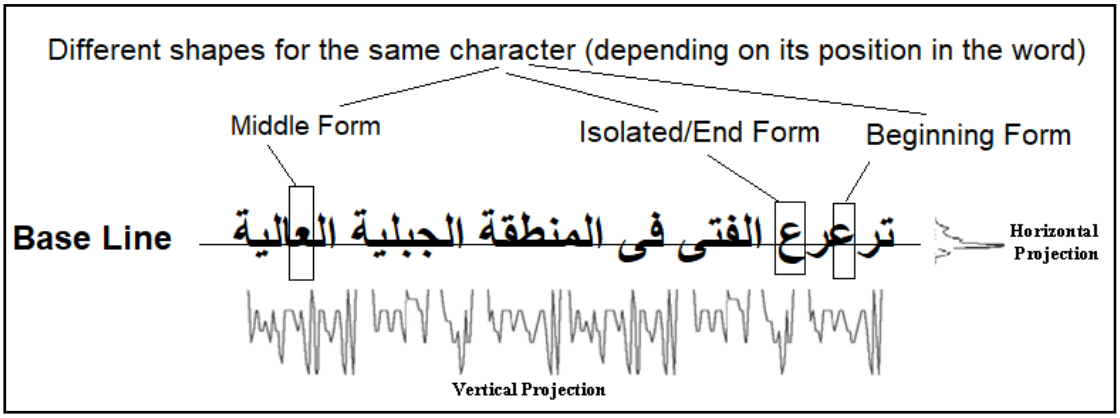


Figure1. Baseline and vertical and horizontal projections for the input text image

The proposed approach:

All the segments that were extracted from the scanned documents matched with the character images, which was stored in the character database. After aggregating all characters, the original text was collected (illustrated in the next section). This process was applied to all target text (regardless of font type or size). A comparison between WHT with FDT was applied regarding recognition rate vs. performance/time.

Walsh-Hadamard Transform (WHT)

With its base function "Walsh function" (which has only two values, +1 or -1), WHT shows good results for digital signals. Also, in terms of image compression, WHT-based is dependent on a couple of techniques to represent the image with minimum data [7]. The Hadamard matrix [10] was used to produce the Walsh transformation. WHT presents high efficiency in many applications, including processing (speech, images, and signals), analysis and filtering, coding.

Discrete Fourier transform (DFT)

The DFT is one of the important techniques used in image processing applications; it has a great degree of precision compared to similar techniques [8]. Also, it is used as: ❶ calculator for the two signal discrete-time convolution, ❷ filter design tool, and ❸ discrete-time signal measuring spectra [9].

In addition, if magnitudes, frequencies, and phases are considered, a sum of complex exponentials represents the image by The Fourier transform [11]. Normally, analysis of planar curve is done by using of a Fourier Transform. In the OCR techniques, the character boundary draws the curve. Therefore, analysis with DFT techniques would be perfect since the boundary of each character is a closed curve, and that the coordinate's sequence of (x; y) which draws the curve is periodic [12].

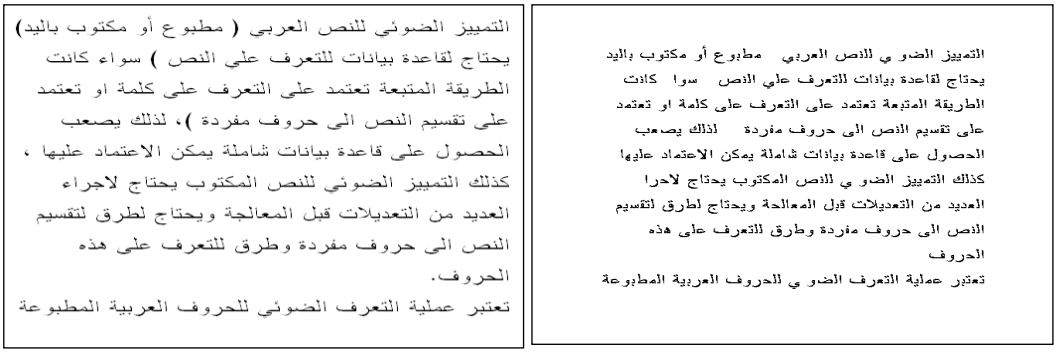
ISSN: 2706-9087
ISSUE 8 DESEMBER 2019
مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences
العدد 8 ديسمبر 2019

Characters Database

All character images (with their different shapes, sizes and font types) are collected in a structured database, which is a file of the character images (For each font size) that consists of the characters' structural and transform features (features vector). The structural features are height, width, the number of pixels above baseline for each character.

Results:

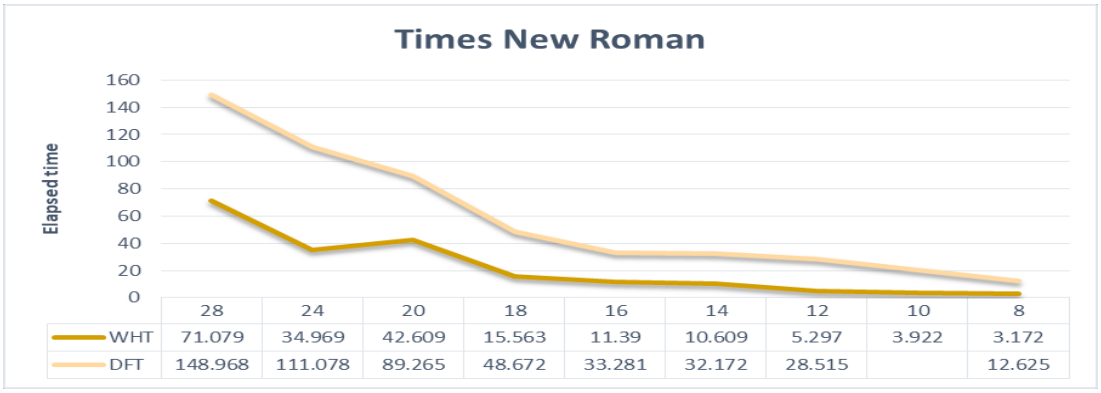
Figure 2.presents (for clarification) how the results of using the proposed method look like which represents the result of recognizing characters (the sample in Figure 2.a, the result in Figure 2.b). Also, the same sample was used in nine different font sizes (the selected sizes are 8, 10, 12, 14, 16, 18, 20, 24 and 28), and three different font types (Arial, Times New Roman and Simplified Arabic).



a: Original document image *b: Result of character reconitin*

Figure 2: Sliding window AOCR in Times New Rom Font size18

In this experiment, the recognition time (in seconds) which is used in the comparison is shown in Figures 3-5. The spectral domain features are extracted by transforming the character image under the moving window using WHT and DFT. These results clearly show that extracting the spectral domain features using WHT gives a faster recognition rate than the DFT.

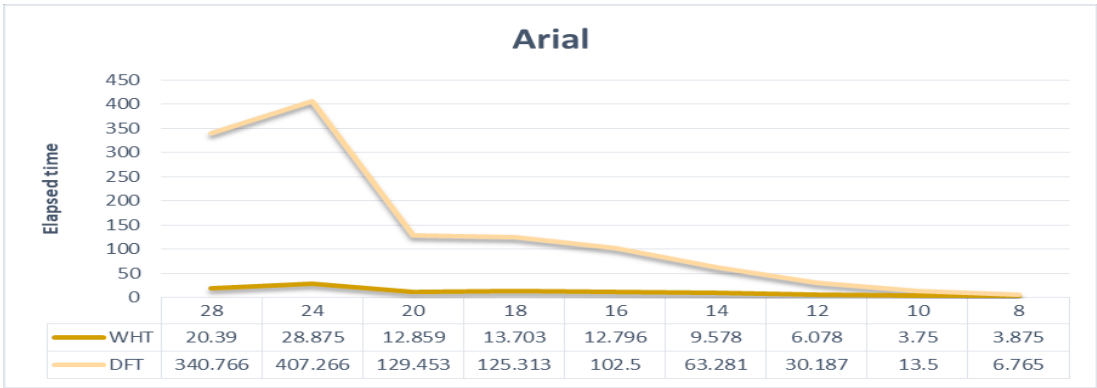


	28	24	20	18	16	14	12	10	8
WHT	71.079	34.969	42.609	15.563	11.39	10.609	5.297	3.922	3.172
DFT	148.968	111.078	89.265	48.672	33.281	32.172	28.515		12.625

Figure 3: Recognition time for Times New Roman font

ISSN: 2706-9087
ISSUE 8 DESEMBER 2019
مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences
العدد 8 2019 ديسمبر

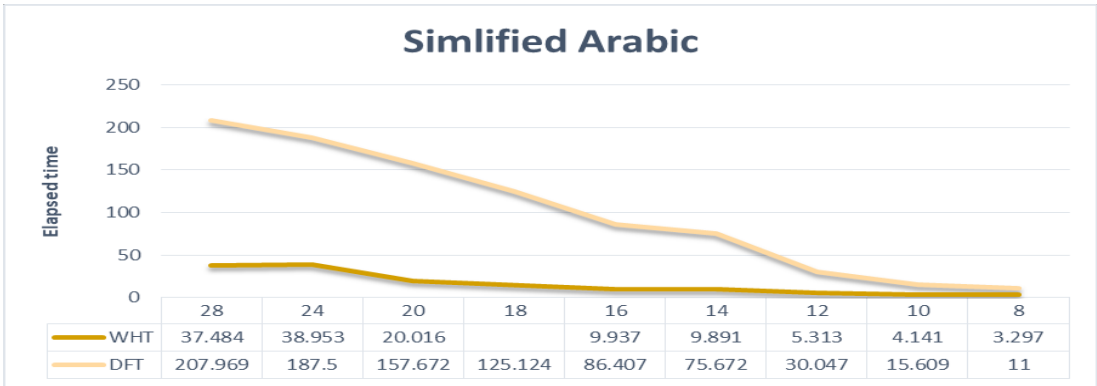
From Figures 3-5 and the attached data, it is clear that small font size characters can be recognized faster than large fontcharacters. This can be justified by two reasons: (1) large font size characters have different dimensions than the small size characters; (2) Also, the number of lines in large font size characters is greater than the number of lines in the small font size characters.



	28	24	20	18	16	14	12	10	8
WHT	20.39	28.875	12.859	13.703	12.796	9.578	6.078	3.75	3.875
DFT	340.766	407.266	129.453	125.313	102.5	63.281	30.187	13.5	6.765

Figure 4: Recognition time for Arial font

For the above mentioned reasons, the needed time for recognition is less for smaller font sizes. In addition, transforming the characters using WHT or DFT, to extract their features, requires a large transform matrix for large font sizes which in turn requires more time for transformation process.



	28	24	20	18	16	14	12	10	8
WHT	37.484	38.953	20.016		9.937	9.891	5.313	4.141	3.297
DFT	207.969	187.5	157.672	125.124	86.407	75.672	30.047	15.609	11

Figure 5: Recognition time for Simplified Arabic font

Conclusion:

To conclude, the results of the experiments carried out by using different implementation (WHT vs. DFT) to the "sliding window for printed AOCR" method [1] (WHT vs. DCT where used) are as follows:

1) It can be clearly noticed, the segmentation-free printed Arabic character recognition technique using Walsh Hadamard Transform (WHT) gives a much better

performance (required time for recognition) compared with Discrete Fourier Transform (DFT).

2) The experiment results of using the proposed method gives a very high rate of letter recognition (up to 99%) for all printed Arabic characters despite the used font size and type of the characters and its position inside the word.

3) When font size 8 was used as an exceptional case, less letter recognition rates were detected with specific characters (e.g. ر,و)

4) Improvement in recognition time is achieved as WHT produced better results than DFT with respect to recognition speed

Reference

[1] Besbas W., Elbokhare A., Sunni M., Gamati E., "Performance Comparison of WHT and DCT Transforms Used in Sliding Window Printed Arabic OCR" International Science and Technology Journal Vol. 10 April 2017.

[2] Delgutte B. and Greenberg J., "The Discrete Fourier Transform", Biomedical Signal and Image Processing", Spring 2005.

[3] Fazekas A. and HajduA., "Recognizing Typeset Documents using Walsh Transformation", Journal of Computing and Information Technology - CIT 9, 2001, 2, 101–112.

[4] Hopkins J., "Character Recognition using Fourier Description", A project submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science, Boise State University, April 2006.

[5] Jayathilake A., Perera A. , Chamikara M., " Discrete Walsh-Hadamard Transform in Signal Processing", IJRIT International Journal of Research in Information Technology, Volume 1, Issue 1, January 2013, Pg. 80-89 80 A.A.C.A.Jayathilake *et al*, IJRIT.

[6] McAndrew A., "An Introduction to Digital Image Processing with Matlab", Notes for SCM2511 Image Processing 1 Semester 1, 2004, School of Computer Science and Mathematics Victoria University of Technology.

[7] Nashwan F. , Rashwan M. , Al-Barhamtoshy H. , Abdou S., and Moussa A., "A Holistic Technique for an Arabic OCR System", Journal of Imaging 2018.

[8] Rashad M., and Semary N., " Isolated Printed Arabic Character RecognitionUsing KNN and Random Forest Tree Classifiers", A.E. Hassanien et al. (Eds.): AMLTA 2014, CCIS 488, pp. 11–17, 2014. Springer International Publishing Switzerland 2014 .

ISSN: 2706-9087

ISSUE 8 DESEMBER 2019

مجلة العلوم الإنسانية والتطبيقية
Journal of Humanitarian and Applied Sciences

العدد 8 ديسمبر 2019

[9] Sarfraz M., Nawaz S.N. and Al-Khuraidly A. (2003) " Offline Arabic Text Recognition system" *International Conference on Geometric Modeling and Graphics (GMAG'03) 0-7695-1985-7/03* .

[10] Shahin A., " Printed Arabic Text Recognition using Linear and Nonlinear Regression", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 1, 2017.

[11] Win K., Htwe N., Image Compression Based on Modified Walsh Hadmard Transform (MWHT) , *3rd ISERD International Conference, Singapore, May 2015*.

[12] Yeotikar V., Wanjari M., Dhore M., " Text Extraction from Document Images Using Fourier Transform Based Method", International Journal of Computer & Mathematical Sciences IJCMS, pp89-95, Vol. 3, Nov. 2014.