Handwritten Arabic Character Recognition: Which Feature Extraction Method?

A. Lawgali, A. Bouridane, M. Angelova, Z. Ghassemlooy
School of Computing, Engineering and Information Sciences
Northumbria University, Newcastle upon Tyne, UK
ahmed.lawgali@northumbria.ac.uk

Abstract

Recognition of Arabic handwriting characters is a difficult task due to similar appearance of some different characters. However, the selection of the method for feature extraction remains the most important step for achieving high recognition accuracy. The purpose of this paper is to compare the effectiveness of Discrete Cosine Transform and Discrete Wavelet transform to capture discriminative features of Arabic handwritten characters. A new database containing 5600 characters covering all shapes of Arabic handwriting characters has also developed for the purpose of the analysis. The coefficients of both techniques have been used for classification based on a Artificial Neural Network implementation. The results have been analysed and the finding have demonstrated that a Discrete Cosine Transform based feature extraction yields a superior recognition than its counterpart.

Keywords: *Arabic character, DCT, DWT.*

1 Introduction

Automatic off-line recognition of text, which is the ability of the computer to distinguish characters and words, can be divided into the recognition of printed and handwritten characters. Printed characters have one style and size for any given font. However, handwritten characters have styles and sizes which vary both for the same writer and between different writers. Many languages such as the Persian, the Urdu and the Jawi use Arabic characters [1]. The Arabic script is written from right to left and is composed of 28 characters with no capital or lower cases. Each character has two or four shapes where the shape of each character depends on its position in the word. The dots play a significant role in Arabic characters. The shape of some characters is similar but the difference arises with the position and the number of dots such as (ب, ت, ث), which can take place either above or below the characters. Some handwritten characters may appear to be similar although they are different and it is difficult for the human eye to spot the difference [2]. The length and width of Arabic characters can also be different; for example (ا, ب). The same character can be written differently in various forms; for example (ع, ء) [2]. In an automatic recognition system, the selection of the feature extraction method might be the most important step for achieving a high recognition accuracy. For example, Discrete Cosine Transform (DCT) can be used to extract the features of handwritten Arabic words. Alkhateeb *et. al.* [3] presented a technique for recognition of handwritten Arabic words where

DCT is used for extracting features of the word. These features are then fed into a neural network for classification. Mowlaei *et. al.* [4] introduced an approach for recognition of isolated handwritten Arabic/Farsi characters. They categorized isolated characters into 8 different classes, where each class contains similar characters. Likewise, a Discrete Wavelet Transform (DWT) can also be used to extract the features. In this paper, both DCT and DWT are used and compared for feature extraction of all the shapes of handwritten Arabic characters using an ANN in the classification stage. This is an important step for the recognition of processes after segmentation. The purpose of the work is to ascertain the effectiveness of each technique to capture useful information and hence resulting in more accurate recognition results. The organization of the paper is as follows. Section 2 describes data acquisition and pre-processing . Section 3 discusses the two methods used for features extraction. Section 4 describes the classification stage. Section 5 discusses the results and their analysis. Section 6 concludes the paper.

2 Data Acquisition and Pre-processing

Due to the non availability of a reference database of Arabic handwriting characters, a new database containing 5600 Arabic handwritten characters has been developed. The database is designed to cover all shapes of Arabic characters and is written by 50 writers, their age ranging from 14 to 50 years. The forms were scanned in colour mode and at resolution of 300 dpi. The images of the characters were converted into a binary format with small objects considered as noise and removed. The dots and marks, such as "hamza", are removed from the characters since they can affect the classification and also in order to reduce the number of classes used. The images of characters are resized as *64x64*, *128x128* and *256x256* for normalization purposes.

3 Feature Extraction

In printed and handwritten text, the features capture the information extracted from the characters. This information is passed onto the matcher to assist in the classification process. In this research, DCT and DWT are adopted to extract the features of the characters. Both DCT and DWT are widely used in the field of digital signal processing applications [5].

3.1 Discrete Cosine Transform (DCT)

DCT is a technique to convert data of the image into its elementary frequency components [5]. DCT clusters high value coefficients in the upper left corner and low value coefficients in the bottom right of the array*(m,n)*. DCT coefficients *f(u,v)* of *f(m,n)* are computed by:

$$f(u,v) = \alpha(u)\alpha(v)\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) cos\left[\frac{(2m+1)\,\pi u}{2M}\right] cos\left[\frac{(2n+1)\,\pi v}{2N}\right] \qquad (1)$$

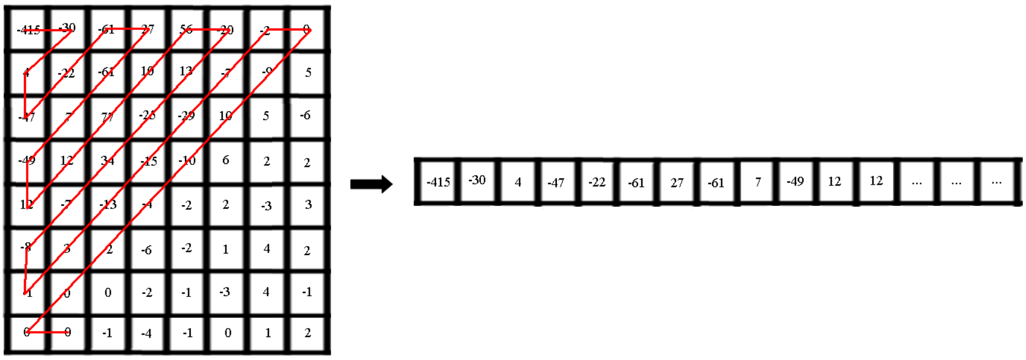
where

$$\alpha(u) = \begin{cases} \frac{1}{\sqrt{M}}, & u = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le u \le M-1 \end{cases}$$

and

$$\alpha(v) = \begin{cases} \frac{1}{\sqrt{N}}, & v = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq v \leq N-1 \end{cases}$$

The higher value DCT coefficients are then extracted in a zigzag fashion and stored in a vector sequence, see Fig. 1. By applying DCT, an image of a character is represented by this vector. Thus, one of the main characteristics of DCT is its ability to convert the energy of the image into a few coefficients [3]. The numbers of DCT coefficients chosen in the classification stage were set to 200 coefficient of an image of size *64x64*, 250 coefficients for *128x128* and 400 coefficients for *256x256* rather than all coefficients for all images. Extensive experiments were carried out using different values of coefficients and it was found that these coefficients were the most appropriate. The numbers chosen was determined by empirical testing in order to reconstruct perceivable characters. Fig. 2 depicts the original image and the image reconstruction with 250 coefficients. These features are utilized for recognition in classification stage.



-415	-30	-61	27	56	-20	-2	0
4	-22	-61	10	13	-7	-9	5
-47	7	77	-25	-29	10	5	-6
-49	12	34	-15	-10	6	2	2
12	-7	-13	-4	-2	2	-3	3
-8	3	2	-6	-2	1	4	2
-1	0	0	-2	-1	-3	4	-1
0	0	-1	-4	-1	0	1	2

-415	-30	4	-47	-22	-61	27	-61	7	-49	12	12	...	...	...

Figure 1. Rearranging DCT coefficients from zigzag order into one vector

(a) Original image

(b) Image reconstruction by using 250 coefficients

Figure 2. The original image and image reconstruction by using 250 coefficients

3.2 Discrete Wavelet Transform (DWT)

DWT is another technique used to extract the features of the characters where, at each decomposition level, a low-pass filter (LPF) and a high-pass filter (HPF) are applied to each row/column of the image to decompose into one low-frequency sub-band (LL) and three high frequency sub-bands (LH, HL, HH) [6]. Fig. 3 shows decomposition of DWT at one level.

The LL is known as an approximation of the coefficients and it represents the horizontal and vertical low frequency. The sub-band HL is known as horizontal sub-band; it represents the horizontal high and vertical low frequency. The horizontal low and vertical high

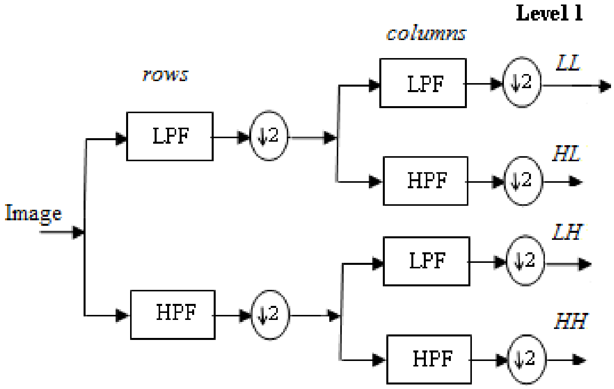
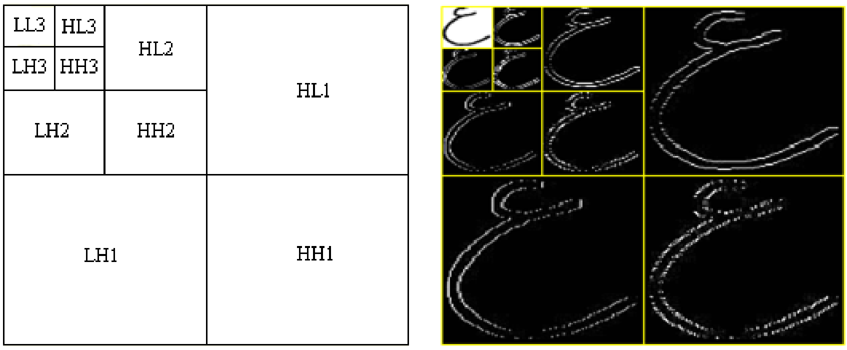


Figure 3. DWT decomposition at one level

frequency represent LH and are known as the vertical sub-band. Finally, the sub-band HH is known as diagonal coefficients; it represents horizontal and vertical high frequency components. The image can be decomposed on more than one level. Fig. 4 depicts a decomposition with three levels. There are various types of wavelet transforms which can be applied, such as Haar and Biorthogonal, etc. Each of them has its particular features. From practical experience, Haar transform has achieved the best result in Arabic handwriting recognition [7,8]. Each character is decomposed into three levels by the Haar wavelet. The low frequency coefficients are close to the original image and they contain full details of the image [6, 7]. Therefore, these coefficients are used to detect the features of the character image.



(a) DWT decomposition at three levels (b) Decomposed image at three levels

Figure 4. Three levels decomposition of character image

4 Classification

ANN is a nonlinear system which is used widely for problems which are not explicitly formulated, such as the pattern classification [1,9]. It has been used to deal with the features that have been extracted from the characters. ANN consists of processing elements with weights which are learned from the training data. Three layers were used in this present research for the architecture of the network: the input layer, the hidden layer and the output layer. Fig. 5 depicts example of the architecture 3-layer ANN. The input layer is

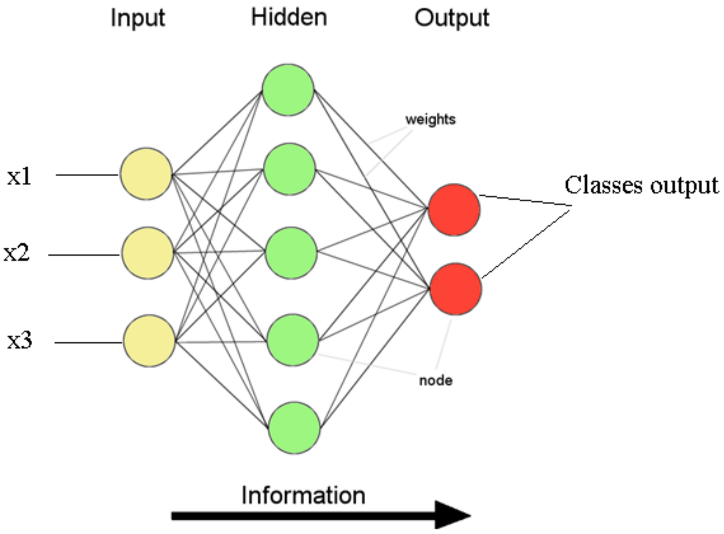


Figure 5. Example of the architecture ANN with 3-layer

fed by the features of the characters. Therefore, the number of nodes in this layer depends on the number of input features of the network. The last layer is called the output layer and the number of its nodes is based on the desired outputs. The hidden layer lies between the input and output layers. The feed forward network multi-layer perception (MLP) back propagation (BP) with supervised training algorithm is used in this work. It is the best known paradigm of training the neural network to classify patterns [9]. A classifier is used to identify the characters by using their features obtained by applying DCT and DWT. These are then compared and saved as models for the training stage.

5 Experimental Results

Experiments were carried out using a database containing 5600 Arabic handwritten characters which covers all the shapes of the Arabic characters. Both methods (DCT and DWT) were used and compared in terms of the information content of features extracted from the characters with the same ANN structure in the classification stage. To allow for a fair comparison, the sizes of images were set to *64x46*, *128x128* and *256x256* for both methods. The experiments were carried out in two steps. In each step, three sizes of images were applied. The first step was applied on a database containing 1600 isolated shapes of Arabic characters. In the DCT technique, 200 coefficients from size *64x64*, 250 coefficients from size *128x128* and 400 coefficients from size *256x256* were used to recognize the character.

ا	ب	ح	د	ر	س	ص	ط
ع	ف	ل	م	ن	ه	و	ي

Table 1. Arabic isolated character forms

These coefficients were used by the ANN . On the other hand, in the DWT technique, LL coefficients of level three were used and fed to the ANN. Table 2 reports the results obtained by using these different methods. Fig. 6 illustrates the difference of recognition by using different techniques for each size. The second experiment was carried out by using 5600 Arabic handwritten characters which covered all the shapes of Arabic characters. The same number of DCT and DWT coefficients were used in the first and second experiments. A decrease in the performance was noted in the second experiment when compared with the first one due to some of the shapes being similar. The results achieved are summarized in Table 3. Fig. 7 illustrates the difference of recognition rate for all shapes by using different techniques for each size. The results has shown that the feature extraction based on DCT yields a higher recognition rate than the DWT counterpart. A reason may be that DCT is more efficient, if the localization of changes is significant [10], on another hand its ability to compress data of the image makes it more efficient for pattern recognition applications [11]. However, the errors are mainly due to some Arabic handwritten characters appearing to be similar and it is thus difficult to recognize them. In some cases, the classification depends on contextual information used to recognize the character.

	DCT	DWT
64x64	94.87%	59.81
128x128	96.56%	57.00
256x256	92.18%	39.62

Table 2. Recognition rate by using different techniques

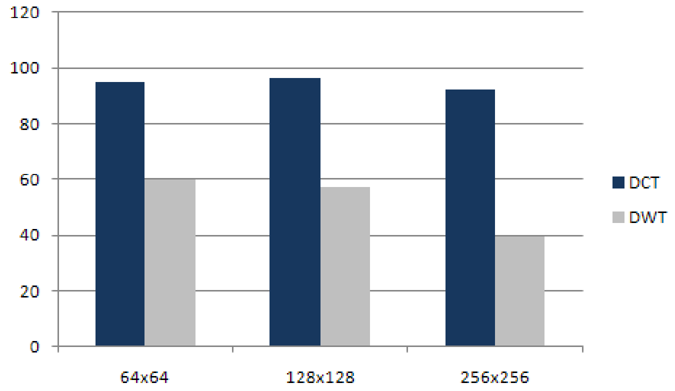


Figure 6. The difference of recognition by using different techniques for 1600 shapes

	DCT	DWT
64x64	79.87%	40.71
128x128	78.82%	27.85
256x256	73.82%	10.57

Table 3. Recognition rate for all shapes by using different techniques

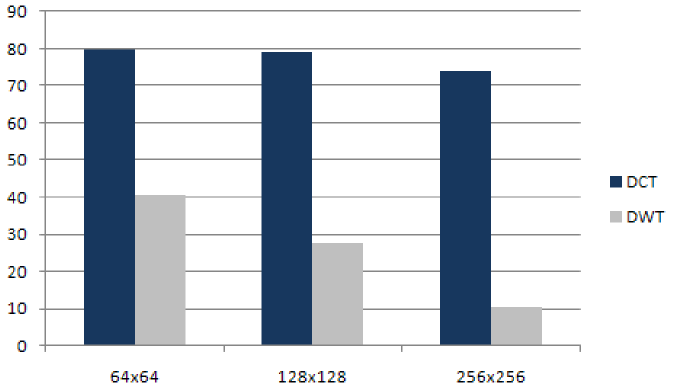


Figure 7. The difference of recognition by using different techniques for 5600 shapes

6 Conclusion

This paper has compared two techniques for features extraction of handwritten Arabic characters. First of all, a new database containing 5600 Arabic handwritten characters has been developed. The database has been designed to include all shapes of Arabic characters. DCT and DWT techniques have been used for features extraction of the characters. Coefficients of both techniques have been used in ANN for classification of the characters. The experiments were carried out in two steps. The first step was applied on a database containing 1600 isolated shapes of Arabic characters. The second experiment was carried out by using 5600 Arabic handwritten characters which covered all the shapes of Arabic characters. The results have demonstrated that features extraction by DCT has a higher recognition rate for Arabic handwritten text.

References

[1] L. M. Lorigo and V. Govindaraju, "Offline arabic handwriting recognition: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 712 – 724, 2006.

[2] A. M. Zeki, "The segmentation problem in arabic character recognition the state of the art," in *First International Conference on Information and Communication Technologies, ICICT 2005.*, pp. 11–26.

[3] J. H. AlKhateeb, Jinchang R., Jianmin J., S. S. Ipson, and H. El-Abed, "Word-based handwritten arabic scripts recognition using dct features and neural network classifier," in *5th International Multi-Conference on Systems, Signals and Devices*, 2008, pp. 1–5.

[4] A. Mowlaei, K. Faez, and A.T. Haghighat, "Feature extraction with wavelet transform for recognition of isolated handwritten farsi/arabic characters and numerals," in *14th International Conference on Digital Signal Processing, DSP 2002.*, 2002.

[5] A. Al-Haj, "Combined dwt-dct digital image watermarking," *Journal of Computer Science 3(9)*, pp. 740–746, 2007.

[6] M. Jiansheng, L. Sukang, and T. Xiaomei, "A digital watermarking algorithm based on dct and dwt," in *Proceedings of the 2nd International Symposium on Web Information Systems and Applications (WISA 2009)Nanchang, China*, pp. 104–107.

[7] J. Alkhateeb, *Word Based Off-line Handwritten Arabic Classification and Recognition*, Ph.D. thesis, School of Computing, Informatics and Media, University of Bradford, 2010.

[8] Z. Razak, N. A. Ghani, E. M. Tamil, M. Y. Idris, N. M. Noor, R. Salleh, M. Yaacob, M. Yakub, and Z.B. Yusoff, "Off-line jawi handwriting recognition using hamming classification," *Information Technology Journal 8(7)*, pp. 971–981, 2009.

[9] A. K. Jain, Jianchang M., and K. M. Mohiuddin, "Artificial neural networks: a tutorial," *IEEE Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.

[10] J. Dowling, B. Planitz, A. Maeder, J. Du, B. Pham, C. Boyd, S. Chen, A. Bradley, and S. Crozier, "A comparison of dct and dwt block based watermarking on medical image quality," in *Proceedings of the 6th International Workshop on Digital Watermarking, IWDW 2008*, pp. 454–466.

[11] A. M. Sarhan, "Iris recognition using discrete cosine transform and artificial neural networks," *Journal of Computer Science*, vol. 5, pp. 369–373, 2009.