

A NEW MODEL OF ARABIC HANDWRITTEN RECOGNITION USING COMBINATION BETWEEN DWT WITH DATA REDUCTION METHOD

¹Dr. ANWAR YAHYA EBRAHIM, ²Dr. ASMAA SHAKER ASHOOR

¹Dr. in Babylon University, Babylon, Iraq

²Dr. in College of Education for Pure Science, University of Babylon, Babylon, Iraq

E-mail: ¹anwaralawady@gmail.com, ²asmaa_zaid218@yahoo.com

ABSTRACT

The objective of this research was to improve an automatic writer recognition technique for off-line Arabic handwritten text. This objective was met by developing a new method which not only addressed the problems in existing techniques but also realized better outcomes than the current solutions on this issue. The projected method is based on Discrete Wavelet Transform (DWT) to extract feature. Also important features for Arabic handwritten alphanumeric character recognition to aid the verification step. This scheme involves of four phases: preprocessing, feature extraction, important features and finally writing recognition. Important features are constructed by data reduction method. After collecting all important features of each character. To recognize fonts, which improves the classification performance. Then decision tree classifier was used for writing recognition. In this paper provide good accuracy dealing with important attributes. The suggested ideal is fast and reliable. The proposed technique has been achieved very promising results, with a validation accuracy of 98.45% for HACDB database.

Keywords: *Discrete Wavelet Transform (DWT), Data Reduction Method, (DT) Decision Tree Classifier, Writing Recognition*

1. INTRODUCTION

Every individual has some unique identities or characteristics which distinguish the person from others. These intrinsic identities or characteristics, which may be physical or behavioral, can be used for the identification of persons and is known as biometrics. Handwriting, whether alphabetical or pictographic based, has existed as an essential means of communication since the beginning of civilization and has evolved over time. Writing styles developed according to the local culture, geographical location, historical background, and temporal circumstances [1]. Early record keeping of handwriting comes from China over 2000 years ago, when they devised first inks and papers. At that time, these early handwriting and all subsequent handwriting followed some standard writing model. Depending upon personal preferences, writers generally do not strictly follow these writing models and the handwriting starts to deviate. These individual writer characteristics serve to distinguish one writer to another writer. Nevertheless, the problem of recognizing writing still is an active area of research and still has many

challenges. Arabic has many characteristics that make the process of recognizing Arabic writing as a difficult task.

This study suggests an Arabic handwritten verification and validation method for Arabic handwritten; based on two stages: The first stage Discrete Wavelet Transform (DWT). This scheme starts with image preprocessing [2]. In this step the noise is removed to eliminate unwanted data that negatively influences accuracy of verification and validation. The second stage is Arabic writing recognition scheme using Sparse principal components analysis (SPCA), where SPCA was applied to avoid difficulties of representing line segments and to reduce the dimensionality of the audit data and arrive at a classifier that is a function of the principal components which represents as input for classification technique [3],[4]. Decision Tree (DT) classifier using to Arabic writing recognition [5].

Experimental outcomes illustration that the suggested scheme has high accuracy compared with other systems; The rest of this research is organized as follow: Section 2 discusses the background and

related studies of the research, Fragment 3 details the projected work, followed by the experimental outcomes in Section 4, verification stage in segment 5, conclusions finally presented in part 6.

2. BACKGROUND AND RELATED STUDIES

Among different biometric identities, handwriting carries significant information about the person in terms of behavior, mental, physical, and emotional states. The way a writer draws strokes reveals the personality while the stability of writing style helps experts to distinguish writing of different writers (Baggett, 2004). Recently, many methods were introduced to verify handwriting. This research mainly focuses on writer recognition of a handwritten document which has some applications in various domains.

In writer recognition process, the main concern is to know about the writer, instead to know what is written, so it is related to the physical manner in which lines and loops are produced. The proposed methodology divides the text according to the proposed technique to find the feature selection from writing and their connected strokes. After division the text, important features are then selected which, are used to characterize the writer. Writer recognition performs two tasks: writer identification and verification. In writer identification task, for a handwritten questioned sample, a list of writers having most similar features is found from a database of known n writers. On the other hand, in writer verification task, given two handwritten samples to determine whether these samples are written by same writer or not. These tasks also help to improve the efficiency while dealing with large data sets.

This research addresses the problem of writer recognition from digitized handwritten samples. Among well-known contributions on this problem, grapheme based writer recognition has shown effective performances but to extract the graphemes, segmentation is required which itself is a very tedious, challenging and time consuming process. The window based approach proposed in (Siddiqi and Vincent, 2010) was intended to address this issue by segmenting the text into small square windows in an attempt to model a writing by small strokes rather than graphemes.

the window based approach represents a very small scale, where each window carries very limited information which, in some cases, may not even be meaningful to characterize the writer. Using the

grapheme level global codebook (Bulacu and Schomaker, 2007) achieve 87% identification rate on 900 writers. The codebook and contour features were combined together (Siddiqi and Vincent, 2010) and evaluated on (650 writers), RIMES (375 writers) and also combination of both the databases (1,025 writers) achieving identification rates of 89%. A single codebook considers only one aspect of writing by capturing information on the primary strokes only without inappropriate features selection result in high number of features, which are not considered as the best solutions for lower value of equal error rate, also for high accuracy and efficiency. Considering the size of real world forensics databases and the performance of existing writer identification systems on standard datasets, the identification rates of these systems needs to be improved to make them more reliable for practical applications.

In this paper, we review a technique that we consider as particularly appropriate for the goal called Discrete Wavelet Transform (DWT), This technique works by extracting the vertical projection feature from writing images. matching has been reported by Lei & al. (2004) [9], where a particular implementation (different from ours) of DWT. The implementation of DWT is described, subsequently.

However, the information of the feature set is usually rather complex, making processing difficult and slow. An essential step in processing writing features is to reduce the number of features of high dimensions. This is needed to make interpretation easier for users. One such method is called principal component analysis (PCA). PCA achieves the reduction by ignoring treat as any feature fewer than a fixed threshold. Unfortunately, such a procedure may also lead to serious errors because of lost information. An alternative to PCA restricts the features to a smaller number of possible values in the derivation of linear functions. (Zou, 2006) proposed sparse principal component analysis (SPCA) to estimate PCs with sparse loadings. This penalty can be specifically made as part of the regression criterion. The result is a PCA with loadings which may be sparse. And speed to select best features [3],[4]. Decision Tree (DT) which uses class trees and regression trees to make decision on how to respond to obtained data. To do so, choices are made to move along the tree, start from the root node to a leaf node, selecting a branch at each step. The leaf node arrived at is the response. A regression tree yields numeric response, checking at each step the value of one predictor (variable). The responses for

classification trees are nominal, such as 'true' or 'false'[10].

Decision tree induction has been widely applied in extracting knowledge from feature-based examples classification and decision making. C4.5 is the algorithm suggested by R. Quinlan in 1993 [5] for building a DT. The C4.5 DT divide data items into subsets, based on feature. If a feature maximizes the gain ratio when dividing information into classes, it is considered useful for producing a DT. This paper is concluded with a discussion and future research issues.

3. PROPOSED SYSTEM

This study addresses writer recognition from off-line Arabic handwritten text by utilizing the proposed technique. The main contributions of this research work include:

- To propose a new scheme for division of Arabic writing to extract useful local information which could eventually better characterize the writer.
- Development of a novel technique by representation of feature selection of each Arabic writing using data reduction methods by using sparse principal components analysis technique (SPCA).
- To classify and adopt the best suited method to boost accuracy rates, able to measure the performance of the classification procedure from the features selected which can achieve better performance in comparison to existing offline Arabic writing verification methods.

Our proposed system is mainly divided into four steps which are summarized in Figure 1.

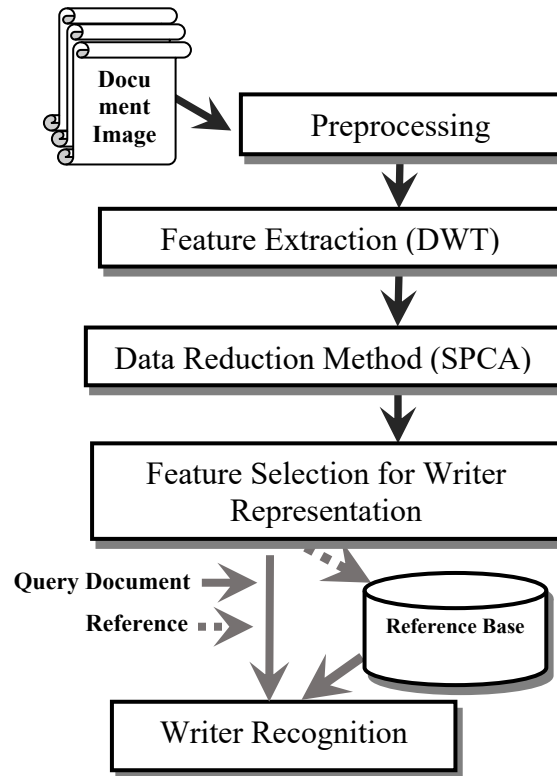


Figure 1: An overview of the proposed methodology

3.1 Preprocessing

The input raw data has been preprocessed based on techniques are: "binarization, noise removal and resize. The proposed work started with converting all writing images into grayscale images and performed binarization on all images using a global thresholding method [11]. As digital images are more likely to get noise, a median filter is also applied on all writing images to enhance image quality. Input numeral images may have lots of variation in size as writing style varies from one individual to another". Therefore, normalization is used to equate the size of each numeral image so that features of all numeral images can be extracted on the same scale.

3.2 Feature Extraction Using Discrete Wavelet Transform (DWT)

Feature extraction is a very crucial step in any classification problem. There feature extraction technique has been used to extract features from numeral images in this paper. (DWT) "is an elastic matching technique and represents feature extraction. DWT deals with breaking up of a signal into shifted and scaled versions of original (mother)

wavelet. In DWT, scales and positions are discrete. The results of DWT are only one direction.

It yields two components- approximation and detail coefficients. A 2d transform is a combination of two 1d transforms in both horizontal and vertical directions and thus generates 4 sub-bands. The approximation coefficients after 1d DWT are equivalent to combination of "LL and LH sub-bands. Similarly, the detail coefficients after 1d DWT are equivalent to combination of HL and HH sub-bands. Therefore, 1d DWT has an advantage over 2d DWT with respect to Arabic writing recognition because it incorporates both global description and horizontal edge details. However, it fails when there is a lot of variation in the expressions. To overcome this problem, we have proposed a novel feature selection technique.

Each sample decompose using DWT into four images [12], the first image represents the low pass values, while the other three images represent the high pass in vertical, diagonal and horizontal directions, respectively. Classifiers using DWT-based on data reduction method have been shown to be well suited for handwriting recognition task.

DWT is mainly used to extract the features from the image [13],[14]. The proposed technique uses the high pass images to extract the necessary information for the Arabic writing verification.

3.3 Data Reduction Method Using (SPCA)

The high pass information collected from DWT is reduced to a feature matrix that represents the main bone of Arabic writing. "The reduction is done by reduction data method" The reduction is used to reduce the information in a small area to simplify the calculations, without affecting the results. Feature selection is the technique of selecting a subset of relevant features for building robust learning models. By removing irrelevant and redundant

features from the data, feature selection helps improve the performance of Arabic handwriting recognition systems by improving generalization capability, minimizing redundancy and speeding up the learning process. Hence, feature selection forms an integral part of an Arabic writing recognition system. "it is essential to use, feature reduction for lessening the dimensionality of the features and selects important features of window fragments separately" For Arabic handwriting verification, feature reduction is normally a process which is carried out after feature extraction. In this study, (SPCA) technique is employed for feature selection. Since features dimensions" of features selection are heavy with

redundant coefficients "reduction of their size requires to choose the most significant ones as shown in Figure 2a,2b. In the presence of many features, select the most relevant subset of (weighted) combinations of features. Consider the problem of analyzing gene expression dataset and to iteratively select the most important genes in an Arabic handwriting data set. Show that adaptive weighted method produces better results when adaptive weighted of features is used. Adaptive weighted of features with sparse PCA is optimal minimizing method to the squared error, it is sensitive to outliers in the data that produce errors SPCA tries to avoid. It therefore is common practice to remove outliers before computing PCA.

The general SPCA algorithm "Define the matrix sparsity as the fraction of zero elements over all columns. Then the sparsity of A. In variable selection, A is some row changing of the identity matrix and the zero matrix. Therefore, a variable selection method is a highly sparse method because error will be low (Zou et al., 2006).

Algorithm 1

Step 1: Suppose A beginning at $V [1: k]$, the loadings of the headmost k (PCs).

Step 2: Assumed a constant $A = [\alpha_1 \dots \alpha_k]$, , fix the next elastic net issue

Step 3:
$$\beta_j = \left(|\alpha_j^T X^T X| - \frac{\lambda_1 \lambda_2}{2} \right) + \text{sign}(\alpha_j^T X^T X)$$
, $j=1, \dots, k$ (1)

Step 4: For a fixed $B_{pxk} = [\beta_1, \dots, \beta_k]$, PCA can be found via compute the SVD of the features matrix, calculate the SVD of $X^T X B = U D V^T$, then update $A = U V^T$.

Step 5: reiterate Steps 4-5, until concourse"

Step 6: Normalization:
$$\tilde{V}_j = \frac{\beta_j}{\|\beta_j\|}$$

In step 1, where "PCs are the linear combinations of all original features, V is the response non-zero components, V is less than or equal to k , given an integer k with $1 \leq k \leq p$. In step third, assumed the variables X be a $(n \times p)$ matrix, where n rows represents an independent feature from features (number of observations) and p is the number of variables (dimensions) respectively, β_j is spare coefficients, j be the predictors for nonzero entries, α_j is features vector, $X^T X$ is represent (covariance matrix) transpose for vectors matrix by row vector of features, $\text{sign}(\alpha_j^T X^T X)$ represents 1- norm in

the constraint, if X is standardized, then apply the (features) correlation matrix, which is chosen when the scales of the features are various. In current research, in order to find the optimal number of features, λ is penalty by directly imposing a constraint on "PCA and if $p > n$, a positive λ is required in order to obtain exact PCA when the sparsity constraint (the lasso penalty) disappears ($\lambda_{i,j} = 0$), call SPCA technique. $\beta = [\beta_0, \beta_1, \dots, \beta_n]^T$ where β is regression coefficients represents the optimal minimizing.

3.3 Classification Technique

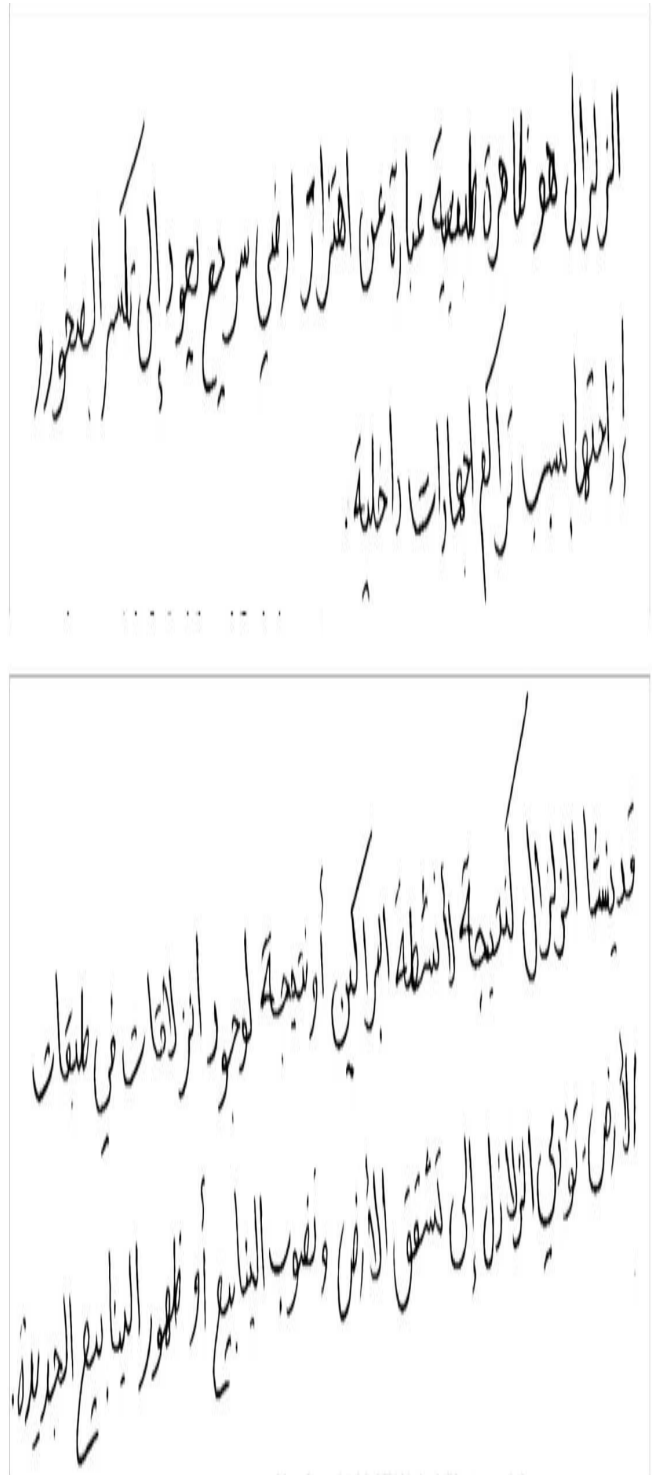
Classification technique involves identifying which of set of important features a new observation belongs, on the basis of training set of Arabic writing containing observations. Here, we use Decision Tree (DT) as the classifier. DT classifier calculates the distance between two corresponding points. This technique is used to measure the similarity between the important features of the test image with the important feature gallery obtained during the training process.

Classification trees and regression trees [16] predict responses to data. To predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. Classification trees give responses that are nominal, such as 'true' or 'false'. Regression trees give numeric responses. Each step in a prediction involves checking the value of one predictor (variable). The classification tree and the regression tree methods perform the following steps to create decision trees: 1. Start with all input data, and examine all possible binary splits on every predictor. 2. Select a split with best optimization criterion. 3. If the split leads to a child node having too few observations (less than the minimum leaf parameter), select a split with the best optimization criterion subject to the minimum leaf constraint. Impose the split. Repeat recursively for the two child nodes.

4. EXPERIMENTAL RESULTS

4.1 Preprocessing Stage

For Arabic writer identification on Arabic text, a HACDB database has been employed. Figure 3. shows a sample page of HACDB dataset. Two paragraph in Arabic script (a) written by the same writer and (b) few samples of Arabic dataset written by different writers.



(a)

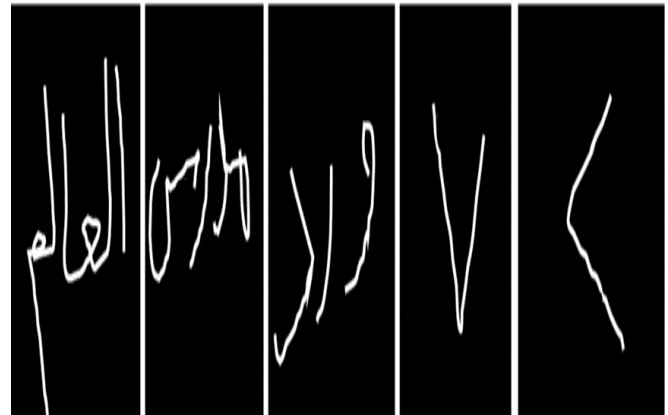
الزلازل هو ظاهرة طبيعية متبادرة عن اهتزاز أرضي سريع يعود إلى تكسر الصخور وإزاحتها بسبب تراكم الإجهادات داخلية.

الزلازل هو ظاهرة طبيعية عبارة عن اهتزاز أرضي سريع يعود إلى تكسر الصخور وإزاحتها بسبب تراكم الإجهادات داخلية.

الزلازل هو ظاهرة طبيعية متبادرة عن اهتزاز أرضي سريع يعود إلى تكسر الصخور وإزاحتها بسبب تراكم الإجهادات داخلية.



(a) "Original Arabic writing image"

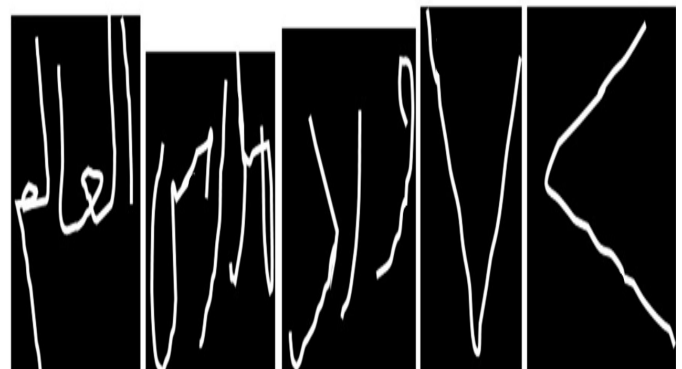


(b) "Scanned and converted from grayscale"

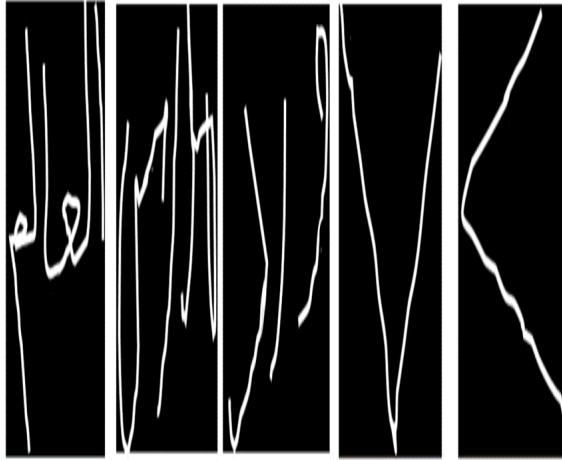
(b)

Figure 3: Two paragraph of Arabic text written by the same writer (a) and Some samples of Arabic text written by different writers

In this phase, the original Arabic writing image (Figure 4a) was scanned and converted from grayscale (Figure 4b)" to binary image (Figure 4c) so as to remove the background noise for improved identification process. "The resized image was easy component identification (Figure 4d).



(c) Remove the background noise



(d) Resized image

Figure 4: Preprocessing stages for Arabic writing

4.2 Feature Extraction Stage

The DWT technique was used to feature extraction as shown in (Figure 5). This helps improve the accuracy of the feature extraction process which is a very vital input to the performance of the entire Arabic writing identification and verification process and makes computation much easier and with minimal error.

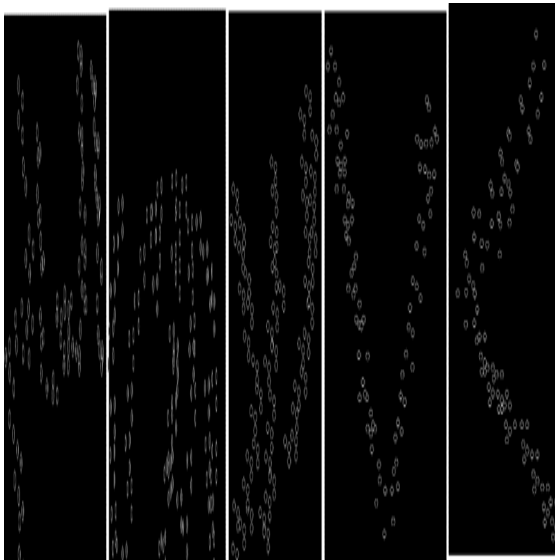


Figure 5: Feature extraction process using DWT

4.3 Feature Selection Stage

The SPCA technique was used to features selected, as shown in (Figure 6). The better categorized the generated data reduction will be, implying a better identification result.

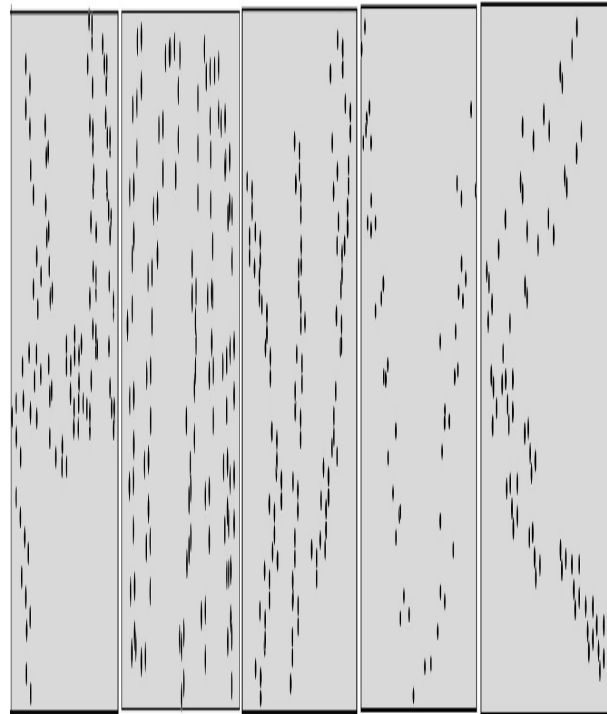


Figure 6: Feature selection process using SPCA

5. VERIFICATION STAGE

This is the final phase where the tested input signature is verified against the sample Arabic writing stored in the database. The HACDB dataset [17] contains 52 basic shapes of characters, and 8 shapes of overlapping characters for a total of 66 shapes of Arabic characters written by 50 people. Each person wrote each character twice. The number of character shapes collected totaled 6600 shapes of Arabic text characters.

The Feature extraction process could be the most challenging stage in Arabic writing recognition. We compared the use of DWT with combination DWT + SPCA to capture the features that are used to discriminate Arabic handwritten texts. The coefficients used in both techniques that have been used for classification were obtained from the

implementation of DT classifier. Analyzing the results of the experiments demonstrated that a DWT+ SPCA based important features provides better recognition results than DWT. Figure 7 shows the results obtained from the experiments, applied to a samples of HACDB database, the recognition accuracy reached the lowest pixel which is the better case (64x64) pixels for DWT+ SPCA 90.90% and DWT 74.11%.

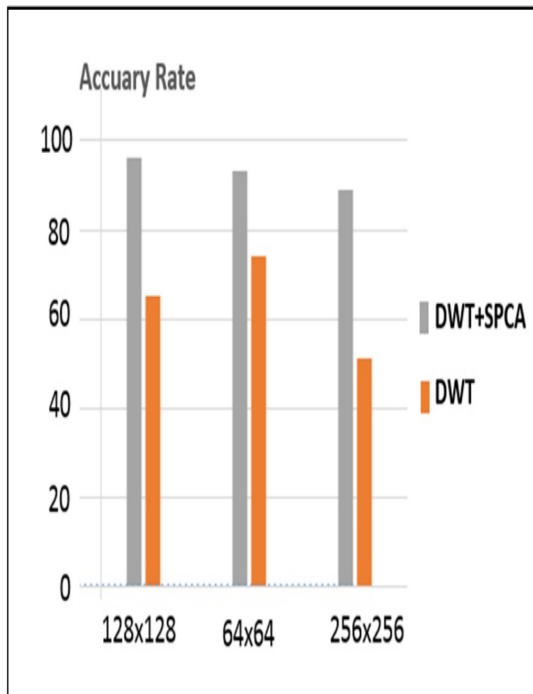


Figure 7: Comparison between using DWT with combination DWT + SPCA in HACDB database

For writer verification, the proposed technique simply compares the distance between two samples with predefined decision threshold. "Two documents in question are assumed to be written by the same individual if the distance between them is less than the threshold. The verification results are reported by varying the decision threshold. The overall performance is quantified by computing the accuracy rate". By presenting accuracy rate together in our experiment (see Table 1), we could conclude that our DWT with features reduction technique and tree classification was a credible and reliable technique for verification of Arabic writing. Table 1. provides a performance comparison of the proposed approach with other works using the same

database. (Mudhsh et al, 2017) used VGG/CEE technique and achieved classification accuracies equal to 97.32%. The (Elleuch et al., 2015) applied DBN technique and obtained classification accuracies equal to 96.36%. Also (Elleuch et al., 2016) CNN and SVM classifier and achieved classification accuracies equal to 94.17%. The (Elleuch et al., 2016) applied Deep SVM Classifier and obtained classification accuracies equal to 91.36%. Whilst the suggested method achieved classification accuracies equal to 98.45%. These are so far the best results on HACDB dataset.

6. CONCLUSIONS AND FUTURE WORK

This This research presents a practical solution to some of the fundamental problems encountered in the design of off-line Arabic handwriting verification, the limited number of users and, the large number of features from writing, and the lack of forgeries as counter examples. A new approach for feature selection is proposed for accurate design of off-line handwriting verification systems. It combines feature extraction, feature selection. This study presented a method for selecting the best features for offline handwriting verification by using SPCA technique. As a conclusion, the method of selecting the best features among a huge feature will help to improve the performance of handwriting verification.

In some cases, current feature does not improve the capability, these features are too many (high dimensions), which reduce classification process efficiency. Also feature selection methods applied for three causes: facilitation of methods to make them easier to explain by writers, improves the performance of the machine learning model and Finally reducing the time required storage space. This paper was aimed to examine the performance of proposed writer recognition techniques on the standard HACDB dataset which contains samples of Arabic handwritten text. The effectiveness of the proposed method was first evaluated on DWT approach with important features technique and appreciable recognition rates were achieved. Besides displaying test results, various tests and debates of the results were also expressed in this study. It shows that the suggested technique introduces unusually best outcomes as compared to current work in this field. Later, the evaluations were carried out on important features with DT

classifier and achieved results which were very promising, so far the best results on this dataset. The proposed method achieved classification accuracies equal to 98.45% in this paper.

ACKNOWLEDGEMENTS

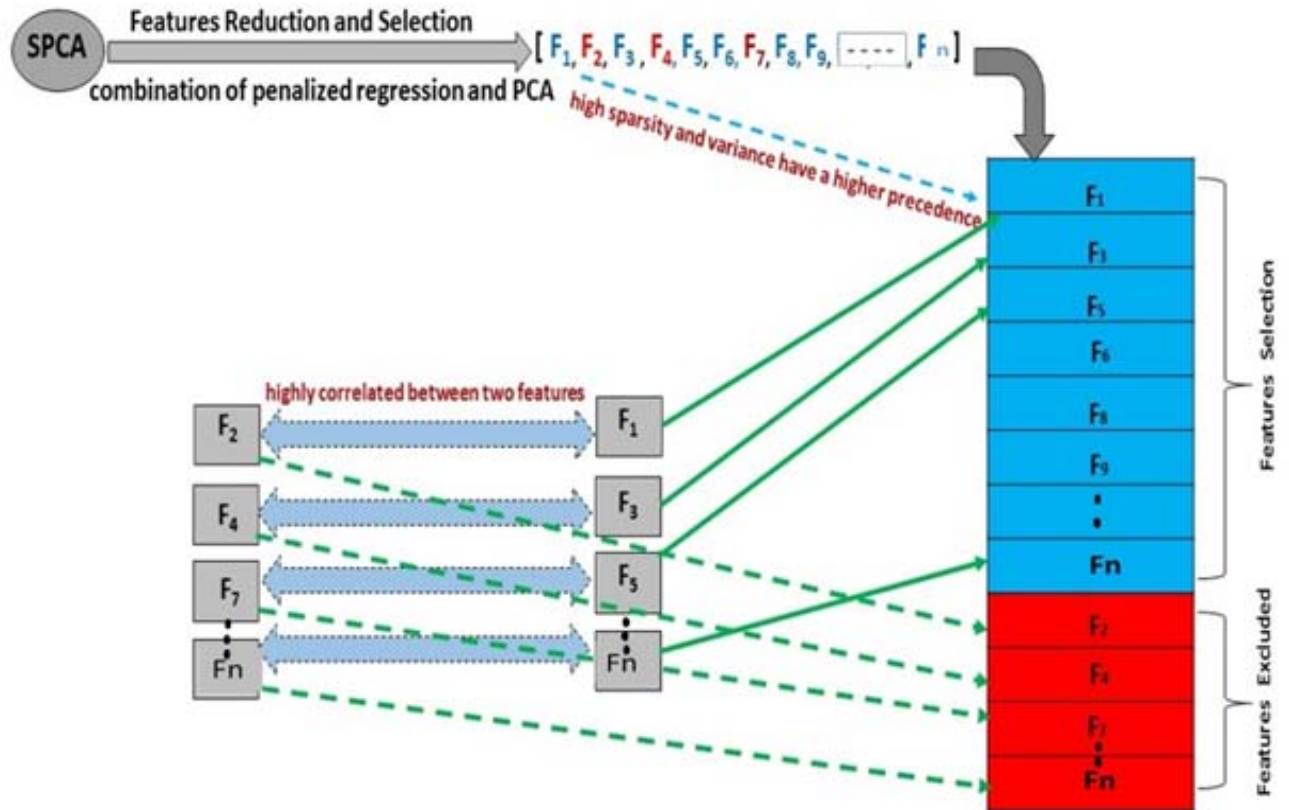
The writers are grateful to the Babylon University, Iraq for submitting study means in achieves this project.

REFERENCES:

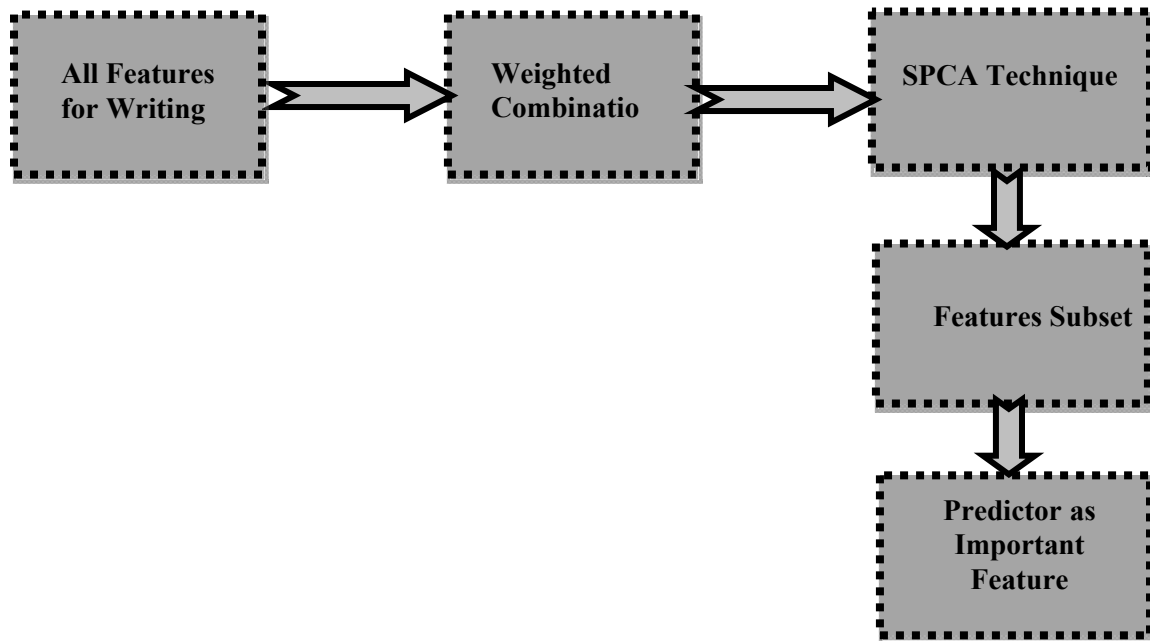
- [1] Plamondon, R. and Lorette, G., " Automatic signature verification and writer identification – the state of the art", *Pattern Recognition*, 22(2), 1989, 107-131.
- [2] W. Tian, Y. Qiao, and Z. Ma, "A New Scheme for Offline Signature Verification Using DWT and Fuzzy Net", *Proceedings of the International Conference on Software Engineering, Artificial, Intelligence, Networking, and Parallel/ Distributed Computing, IEEE Computer Society, Qingdao, China*, Vol. 3, 2007, pp. 30-35.
- [3] Zou, Hui, Trevor Hastie, and Robert Tibshirani., "Sparse principal component analysis", *Journal of computational and graphical statistics* 15.2 (2006): 265-286.
- [4] Ebrahim, A.Y., "Detection of breast cancer in mammograms through a new features and decision tree based, classification framework", *Journal of Theoretical and Applied Information Technology*, (2017b) Vol. 95, No. 12, ISSN: 1992-8645.
- [5] J.R. Quinlan. "C4.5 :Program for Machine Learning Morgan" Kaufmann Publishers, 1993.
- [6] Baggett, B. A. " Handwriting Analysis 101 – The Basic Traits", *Empressé Publishing*, 2004.
- [7] Siddiqi, I. and Vincent, N. "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features". *Pattern Recognition*, (2010), 43(11), 3853-3865.
- [8] Bulacu, M. and Schomaker, L. "Automatic handwriting identification on medieval documents". *Proceedings of Document Analysis*, 2007, Modena, 279-284.
- [9] Lei, H. & Palla, S. & Govindaraju, V. " ER2 : An Intuitive Similarity Measure for On-Line Signature Verification", *In Proc. IWFHR9*, Kimura F. & Fujisawa H. eds, Tokyo (Japan), 2004, October 26- 29, pp. 191–195.
- [10] V.N.Vapnik, V.N, "Classification-trees-and-regression-trees", *online Referencing, The Nature of Statistical Learning Theory. Springer* .1995.
- [11] Ebrahim, A.Y., "Classification of Arabic autograph as genuine and forged through a combination of new attribute extraction techniques", *Journal of University of Babylon*, Vol. 25, No. 5, 2017 pp.1873–1885.
- [12] Anwar Yahy Ebrahim, Ghazali Sulong, "Offline Handwritten Signature Verification Using Back Propagation Artificial Neural Network Matching Technique", *JATIT & LLS*. (2014), 31st July. Vol. 65 No.3. All rights reserved.
- [13] R. Gonzalez and R. Woods, "Digital Image Processing", 3rd edition, *New Jersey, Prentice- Hall, USA*, 2008.
- [14] Deepu V., "On-line writer dependent handwriting character recognition," *Master of Engineering project report*, Indian Institute of Science, India, Jan. 2003.
- [15] Mark S. Nixon & Alberto S. Aguado, " Feature Extraction and Image Processing", *Elsevier Limited, Third Edition*. 2008.
- [16] Hiremath, T. R., Patil, S. M., & Malemath, "Detection and Extraction of Text in Images using DWT", *Int. J. Adv. Res. Comput. Commun. Eng.*, (2015). 4(6), 533-537.
- [17] <http://www.mathworks.com/help/stats/classification-trees-and-regression-trees.html>.
- [18] Lawgali, A.; Angelova, M.; Bouridane, "HACDB: Handwritten Arabic characters Database for automatic character recognition", *In Proceedings of the 4th European Workshop on Visual Information Processing*, Paris, France, 10– 12 June 2013; pp. 255–259.
- [19] MA Mudhsh, R Almodfer - Information, "Arabic Handwritten Alphanumeric Character Recognition Using Very Deep Neural Network", *Information* (2078-2489), 2017 – search. Ebscohost.com.
- [20] Elleuch, M.; Tagougui, N.; Kherallah, "Towards Unsupervised Learning for Arabic Handwritten Recognition Using Deep Architectures", *In Proceedings of the International Conference on Neural Information Processing*, Istanbul, Turkey, 9–12 November 2015; pp. 363–372.



- [21] Elleuch, M.; Maalej, R.; Kherallah, "New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition". *Procedia Comput. Sci.* 2016, 80, 1712–1723.
- [22] Elleuch, M.; Kherallah, "An Improved Arabic Handwritten Recognition System using Deep Support Vector Machines", *Int. J. Multimed. Data Eng. Manag.* 2016, 7, 1–20.



(a)



(b)

Figure 2: (A) SPCA Features Selection Vector, (B) WSPCA Technique Process



Table 1: An Evaluation Table Comparing The Proposed Method With Other Previously Known Methods On HACDB Data Set For Arabic Writing

Authors	Method	Accuracy
Proposed,(2018)	WDT with SPCA + DT Classifier	98.45%
Mudhsh, (2017)	VGG/CEE Technique	97.32%
Elleuch, (2015)	DBN Technique	96.36%
Elleuch, (2016)	CNN and SVM Classifier	94.17%
Elleuch (2016)"	Deep SVM Classifier	91.36%