Journal of Computer Science 2 (12): 879-884, 2006
ISSN 1549-3636
© 2006 Science Publications

Probabilistic Artificial Neural Network For Recognizing the Arabic Hand Written Characters

[1]Khalaf khatatneh, [2]Ibrahiem M.M El Emary and [3] Basem Al- Rifai
[1,3]Prince Abdullah bin Ghazi College for Science and Information Technology
Al-Balqa Applied University, Al Salt, Jordan
[2]Factually of Engineering, Al-Ahliyya Amman University, Amman, Jordan

Abstract: The objective of this study was to present a new technique assists in developing a recognition system for handling the Arabic Hand Written text. The proposed system is called Arabic Hand Written Optical Character Recognition (AHOCR). AHOCR was concerned with recognition of hand written Alphanumeric Arabic characters. In the present work, extracted characters are represented by using geometric moment invariant of order three. The advantage of using moment invariant for pattern classification as compared to the other methods was its invariant with respect to its: position , size and rotation .The proposed technique was divided into three major steps : the first step was concerned with digitization and preprocessing documents to create connect components, detect the skew of characters and correct it .The second step deals with how to use geometric moment invariant features of the input Arabic characters to extract features . The third step focused on description of an advanced system of classification using Probabilistic Neural networks structure which yields significant speed improvement. Our final results indicate and clarify that the proposed AHOCR technique achieves an excellent test accuracy of recognition rated up to 97% for isolated Arabic characters and 96% for Arabic text.

Keywords: Optical characters recognition (OCR), hand written arabic character recognition (AHOCR), probabilistic neural network (PNN), geometric moment invariant (GMI)

INTRODUCTION

A computer technology sub-field which has potential to be useful in a plurality of settings is automated recognition of textual information. This filed has been referred to generally as Optical Character Recognition (OCR). In general ,an OCR machine reads machine printed /hand written characters and tries to determine which character from a fixed set of the machine printed/hand written characters is intended to represent. The task of recognized characters can be broadly separated into two categories: the recognition of machine printed data and the recognition of hand written data Machine printed characters are uniform in size, position and pitch for any given font. In contrast, hand- written characters are non-uniform, they can be written in many different styles and sizes by different writers and by the same writers. Therefore, the reading of machine printed writing is a much simpler task than reading hand writing and has been accomplished and marketed with considerable success[1,2] .

The work presented in this study tries to demonstrate a framework for giving good recognition accuracy for off-line hand written Arabic characters input by developing a new system that can deal with Arabic hand written characters. So, our work tends to over look the following phases:

1. Studying the various techniques used in recognizing the hand written Arabic characters.
2. Considering moment invariant (order 3) method and Neural Networks (probabilistic Neural Net work) for studying this problem.
3. Studying the problems of characters recognition techniques and make a little of comparisons between these techniques.
4. Developing a new recognition system technique to recognize hand written Arabic characters problems in order to overcome the problems that exist in the current technique.
5. Analyzing the recognized system from the proposed system with respect to the other recognized system obtained from other techniques.

Finally, the proposed approach is trying to prove that using Probabilistic Neural Networks for recognizing hand written Arabic characters is better than other techniques since it overcomes all the problems in the previous techniques and it is suitable for handling all kind of problems in different fields. At the end, Arabic Hand Written Optical Character Recognition (AHOCR) is developed in the devices that use Arabic character recognition to process many documents automatically. AHOCR aims to convert document images to symbolic for modification, storage, retrieval, reuse and transmission. It helps the transition

Corresponding Author: Dr.Ibrahiem M. M. El Emary, Faculty of Engineering, Al Ahliyya Amman University, Amman, Jordan

from book shelves and filing cabinets to the paperless world. Although there is no evidence yet of less paper, electronic document already abound[3, 4].

Also, this study aimed to document a frame work for giving good recognition accuracy to off-line hand written Arabic characters input. So, we propose an automatic prototype extraction method. Then, prototypes are used to train a document-specific OCR system to perform character recognition on document images of the same or similar quality and typesetting. Finally, we say that our contribution provides: a moment method, type of neural network algorithm probabilistic (PNN) which make use of as much information as possible from the character bit maps and labels to estimate character widths, character location, etc and match/ no match probabilities.

Related works of hand written arabic character recognition systems: Machine simulation of human reading has been the subject of intensive research for almost three decodes. A large number of researches are concerned with Latin, chiness and Japanese characters. However, little work has been conduct on the automatic of Arabic characters because of its complexity of text. The main objective of this section is to present the state of Arabic character recognition research throughout the last two decades. The previously related work can be described as:

Hierarchical rule based approach: Sheik and Al-Taweel[5] assumed a reliable segmentation stage which divide letters into the four groups of position (initial, media, final and isolated). The recognition system depends on a hierarchical division by the number of strokes. One stroke letters were classified separately from two stroke letters …etc. Ratios between extreme and position of dots in comparison to the primary stroke were defined heuristically on the data set to produce a rule-based classification.

Segmented structural analysis approach: Al -Emami and Usher[6] used a structural analysis method for selecting features of Arabic characters. The classifications use a decision tree. In preprocessing, some of the features extracted during the segmentation process were direction codes, slops and presence of dot flags. A new input needed to search three decision trees for the primary stroke and also for the upper and lower dots. The system was trained on 10 writers with a set of 120 postal code words with a total of 13 characters. They use one tester who had a recognition rate of 86%.

Structural and fuzzy approach: Amin and Bouslama[7] presented a hybrid system that combine structural and fuzzy techniques. Structural analysis discriminated between various letter classes to be recognized and fuzzy logic allowed for variability in people hand writing within the same class. Sampling was done on the same class and on the input data points by comparing tangent angles at various points along the line. End points were kept automatically. The first point that had a tangent difference bigger than a threshold became next sampled point. The authors chose basic shapes such as curves, loops, lines and dots as good feature for discrimination between letter classes. These were constructed using geometric and structural relationship between the sampled points. After fuzzifying the features, fuzzy " If-Then rules " were created heuristically by the authors followed by a study of the data set. These fuzzy rules could distinguish letters from combination of the fuzzy features and allowed for fuzzy membership to cover the variability in hand written between authors.

Template matching and dynamic programming approach: Alimi and Ghorbel[8] showed how to minimize error in an off-line recognition system for isolated Arabic characters using template matching and dynamic programming with assumed segmentation. The reference bank of prototypes was prepared after smoothing, normalization and coding the data coordinates into a parametric representation of angles. When new data was presented to the system, the distance between the prototype and the new data string was minimized using dynamic programming. The number of prototypes was varied to see the effect on recognition rates. More prototypes give better accruing.

Artificial neural networks classifiers: Haraty and El-Zabadani[9] present a system for recognition of handwritten Arabic text using neural networks. Their work builds upon previous work that dealt with the vertical segmentation of the written text. However, they faced some problems like overlapping characters that share the same vertical space. They tried to fix that problem by performing horizontal segmented and the second one performs the horizontal segmentation. Both networks use a set of features that are extracted using a heuristic program. The system was tested and the rate of recognition obtained was 90%. This strongly supports the usefulness of proposed measures for handwritten Arabic text.

Optical character recognition system: The goal of optical character recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of OCR involves several steps including: segmentation, feature extraction and classification. Firstly, for the classification process, there are two steps in building a classifier: training and testing. These steps can be broken down further into sub-steps as shown in Fig. 1[10].

From Fig. 1, we see that the training phase is structured from three sub phases given by pre-processing, feature extraction and model estimation. On the other hand, for the testing phase, this phase contains

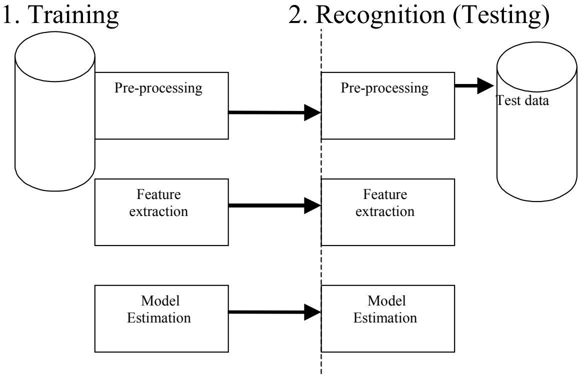


Fig. 1: Pattern classification process

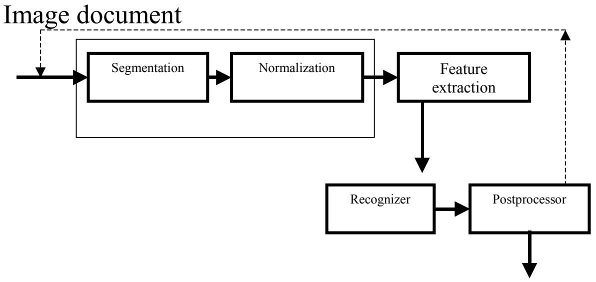


Fig. 2: Flow diagram for optical character recognition system

also three sub phases given by: pre-processing, feature extraction and classification. Secondly, for the OCR-Pre-processing process; preprocessing is primarily used to reduce variations of characters. In OCR systems, pre-processing includes the connection of segmentation and normalization as shown in Fig. 2. Pre-processing receives a first binary image of a plurality of characters. Pre-processing is generally consists of a series of image to image transformation. The pre-processing steps often performed in OCR are: Scanning, Binarization, Noise Removal, Segmentation and Normalization[11].

Thirdly, with regard to feature extractor, this step is the key issue of character recognition. Feature extraction abstracts high level information about individual patterns to facilitate recognition. A wide variety of approaches have been proposed to capture the distinctive feature of machine characters. These approaches fall into one of two categories: (1) global analysis which contains techniques as moments and mathematical transforms. (2) Structural analysis in which efforts are aimed at capturing the essential shape features of characters generally from their skeletons of contours. Fourthly, with regard to recognizer (classifier), there are three categories of character classifiers: neural network approach, statistical approach and structural approach.

Fifthly, for the final phase of OCR which is post-processor, the post-Processor is designed to supplement the recognition process to improve the accuracy of the process. The post-process examines the bitmap generated by the recognizer and makes a determination as to the validity of the selection made by the recognizer. In one embodiment, gross structural feature such as character strokes and contours contained in the bitmap fed-to the post processor by the recognizer are used for this purpose. The post-processor use various techniques to distinguish one character from another. For example, the geometry/topology of an image (e.g bitmap) can be used to distinguish characters represented by the image. Geometric / topological features include (i) loops such that appear in the handwritten letters ف،ق، و، ـه. (ii) Straight lines which appear in such handwritten letters as أ. (iii) endpoints and (iv) intersections.

The proposed handwritten Arabic character recognition system (AHOCR): Handwritten Arabic character recognition system (AHOCR) aims to convert images document to text. The main objective of this section is to introduce a novel method of off-line handwritten Arabic character recognition used to construct AHOCR. In AHOCR, after moving the document from a storage location to a digital scanner, for each image document scanned at 300 dpi, the output is formatted image with ('bmp': windows Bitmap (BMP), 'jpg' or 'jpeg', joint photographic expert group (JPEG) and adequate results are obtained with this quality, the scanner generates image document or a selected portion of the image document. The scanner then applies the image to a recognition system according to the invention.

The AHOCR recognition system as shown in Fig. 3 includes a Geometric moments Invariant (GMI) for feature extraction and Probabilistic Neural Network (PNN) for recognition. Recognizer can process the image received from the scanner to determine all of the character written on the document that was scanned. In general, the recognition system generates data representative of each character on the document and passes that data as output to a buffer memory. The central computer typically controls and /or initiates other aspects of the operations such as storage, retrieval, reuse and transmission. It helps the transition from bookshelves and filling cabinets to the paperless. The main components of AHOCR are: Pre processor, feature, extractor and recognizer, classifier.

The first stage of our proposed AHOCR is called preprocess. This stage is structured from various phases given by: Binarization, noise removal, segmentation and normalization. The second stage is called character feature extraction operation of AHOCR. This stage is subdivided into the following phases: Geometric moment invariant, moment invariant for Arabic character of AHOCR, neural network classifier operation of AHOCR, probabilistic neural network structure, probabilistic neural network classifier of AHOCR, Creating database for Arabic Handwritten Invariant moment of our proposed approach as well as operation concepts is shown in Fig. 4.

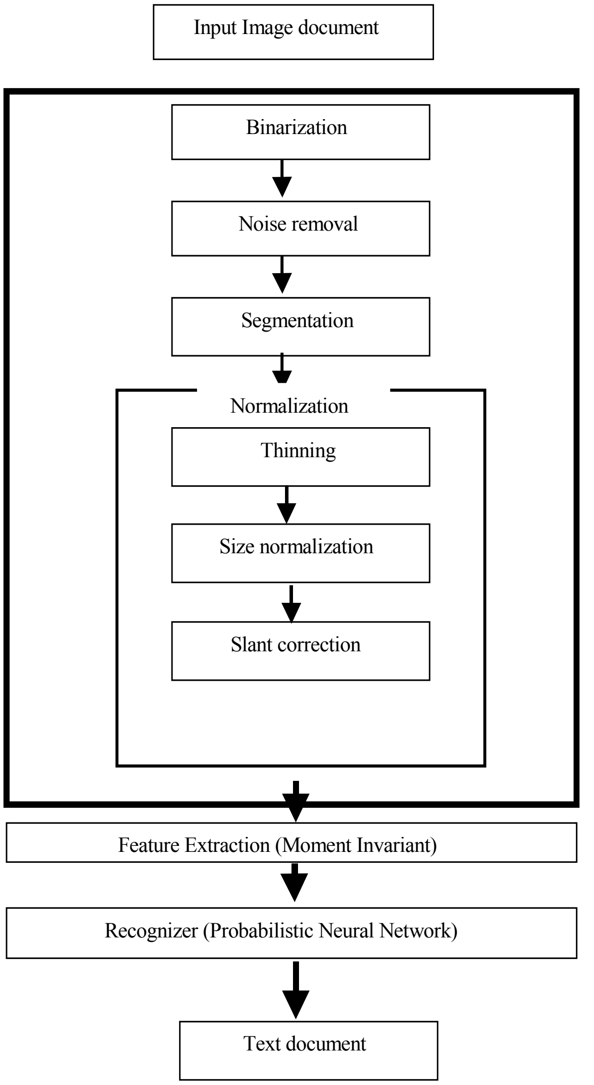


Fig. 3: Flow diagram for the proposed handwritten Arabic character recognition system (AHOCR)

Experimental results of AHOCR for handwritten Arabic text document: Our experiments were conducted on the Arabic handwriting of 25 independent writers document shown in Fig. 5. These documents were then processed. The experiments were done on 3 disjoint data sets given by:

1. Training (37800)= 20 volunteers x 5 iterations x 378 characters
2. Validation (3780)= 10 volunteers x 378 characters
3. Test (764) (5 volunteers with different number of characters in each document).

The Validation set was composed of characters that were written by the same authors that were totally seen in the training process. The test set was written by 5 authors that were totally unseen in the training process. There are two procedures done in our experiments: training and testing. The trial trained on the training set and tested on the validation set and test set.

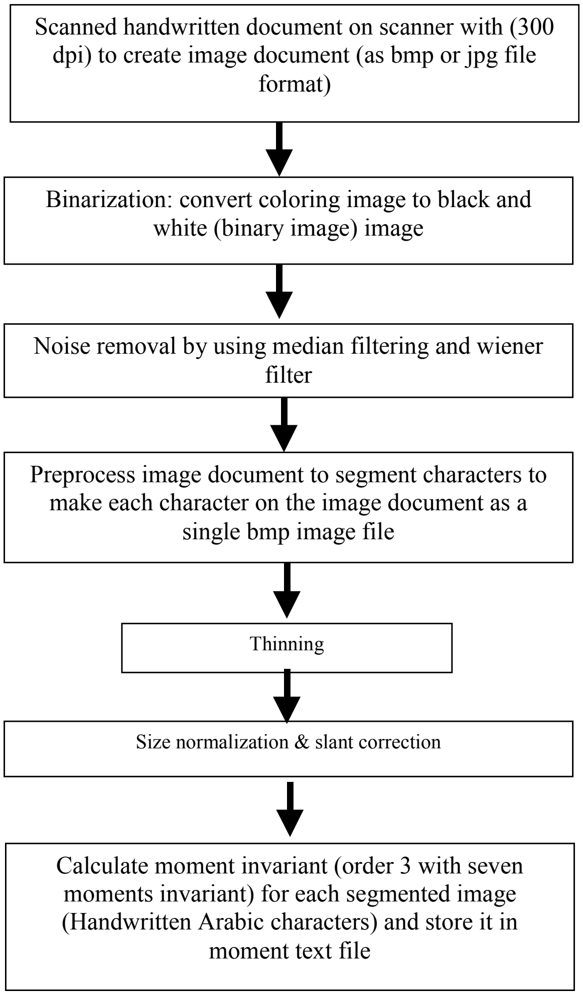
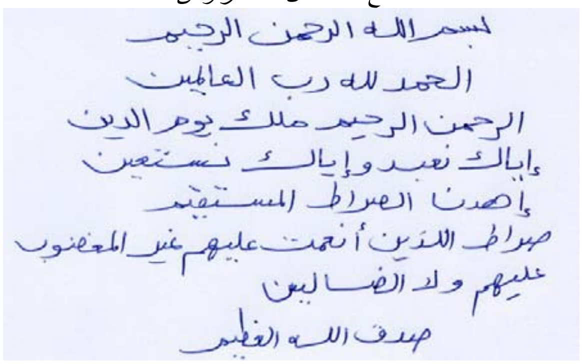

Scanned handwritten document on scanner with (300 dpi) to create image document (as bmp or jpg file format)
Binarization: convert coloring image to black and white (binary image) image
Noise removal by using median filtering and wiener filter
Preprocess image document to segment characters to make each character on the image document as a single bmp image file
Thinning
Size normalization & slant correction
Calculate moment invariant (order 3 with seven moments invariant) for each segmented image (Handwritten Arabic characters) and store it in moment text file

Fig. 4: Operations executed on image document to create database for Arabic handwritten invariant moment

سورة الفاتحة



بسم الله الرحمن الرحيم
الحمد لله رب العالمين
الرحمن الرحيم ملك يوم الدين
إياك نعبد وإياك نستعين
إهدن الصراط المستقيم
صراط الذين أنعمت عليهم غير المغضوب
عليهم ولا الضالين
صدق الله العظيم

Fig. 5: Arabic handwritten document written by the first test writer

The experiment was to simply train the probabilistic neural network architectures on the training set and test on all three sets. The results are become the baseline for letter optimization.

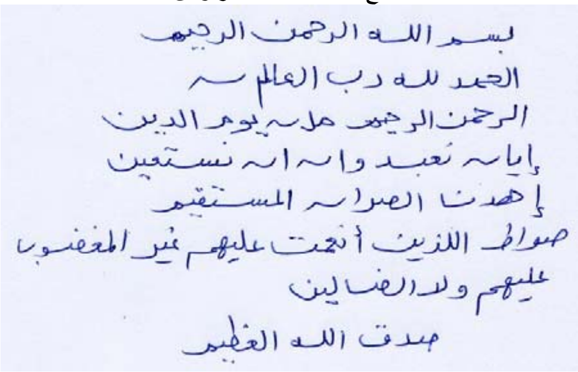
Table 1: Five test writers and recognition accuracy rate of them

Writer	Error rate	accuracy	Rate
21	4%		96%
22	5%		95%
23	3%		97%
24	3%		97%
25	5%		95%
Average test accuracy rate			96%

Table 2: Recognition accuracy results for text Arabic character

Train recognition	Validation recognition recognition	Test recognition	Average
Accuracy rate rate	Accuracy rate	Accuracy rate	Accuracy
22	5%		95%

س و ر ة ا لفا ت ح ة



بسم الله الرحمن الرحيم
الحمد لله رب العالمين
الرحمن الرحيم ملك يوم الدين
إياك نعبد وإياك نستعين
إهدنا الصراط المستقيم
صراط اللذين أنعمت عليهم غير المغضوب
عليهم ولا الضالين
صدق الله العظيم

Fig. 6: Recognized Arabic handwritten document by AHOCR

The baseline for recognition accuracy was defined as the average accuracy of the validation and test set of the best PNN architecture. Probabilistic neural network (PNN) has 7 input nodes in its input layer and 142 nodes in its output layer architecture. For this architecture, there are 7 nodes in the input layer (node for each of the seven moment value of input data (image character) ad described in Table 1, 142 output nodes (141 for Arabic characters, kart node for reject result). In the output layer, we take the advantage of the fact that final values range from 0 to 1. Given this, we can interpret each output of a node in the output layer as a confidence level for each possible output. So, each node in the output layer represents a possible Arabic character being recognized. The output layer is set to reorganize 142 possible outputs: (أ-ي), (0-9), (+, -, *, 1, =, !, ., ;, ?) and not recognized.

After running our algorithm with a learning rate of η=0.9, we found that choosing an error goal for probabilistic architecture during the 8000 epoch's probabilistic training time will improve: the accuracy across training, validation and test (Table 2). The best probabilistic for recognition accuracy is the probabilistic three layers neural networks architecture. For training, it recorded a recognition accuracy rate of 99%. For validation, it had recognition accuracy rate of 98%. For test, it had a recognition accuracy rate of 96% and the average recognition accuracy rate is 97%.

Looking at a breakdown of the test set results (Table 1), we notice that five test writers had a good accuracy rate as shown in Fig. 6.

CONCLUSION AND FUTURE WORKS

AHOCR is optical handwritten Arabic character recognition (OCR) software capable of producing a fully editable electronic document with current accuracy of 97% for isolated Arabic handwritten character recognition and 96% for Arabic handwritten document recognition. After running AHOCR software and experimentation on total of 15933 Arabic characters (for isolated Arabic characters and 42344 Arabic characters for Arabic text), these characters were then process the experiments on 3 disjoint data sets: training data set, validation data set and test data for isolated Arabic characters. We conclude the following assure points.

* Moment- invariant features for handwritten characters are tuned to produce relevant features for Arabic recognition from data coordinates while reducing the input space.
* Probabilistic network are tuned to recognize the 141 character classes in and easy and powerful recognizing way.
* Accurate recognition rate for AHOCR (isolated Arabic character) is 100% for the training data set, 99% for the validation data set and 97%, for the test data set and the average recognition rate is 98%. Accurate recognition rate for AHOCR (Arabic text) is 99%, for the training data set, 98% for the validation data set, 96% for the test data set and the average recognition rate is 97%.

Arabic handwritten recognition is a difficult problem but our hope is that the AHOCR system will be a step towards a neural network approach to robustly solve it. Now, it remains for further research to build on this foundation and work towards automatic recognition of Arabic document handwritten character, recognized Arabic motions.

REFERENCES

1. Amin, A., 2000. Recognition of printed arabic text based on global features and decision tree learning techniques. Pattern Recognition, 33: 1309-1323.
2. Amondon, R and Srihari, 2000. On-line and off-line handwriting recognition: A comparative study. IEEE Trans. Pattern Analysis and Machine Intelligence, 22: 63-84.
3. Amin, A. and K. Mandana, 1999. Automatic recognition of printed Arabic text using neural network classifier. IEEE Trans. Pattern Analysis and Machine Intelligence, 20: 1.

4. Amin, A. and S. Singh, 1999. Neural network recognition of hand printed characters. Neural Computing and Applications, 8: 76-76.
5. Klassen, T., 2001. Towards neural network recognition of handwritten Arabic letters. Master Theses, Dalousie University, Halifax, U.S.A.
6. Al Emani, S. and M. Usher, 1990. Off-line recognition of handwritten Arabic characters. IEEE Trans. Pattern Analysis and Machine Intelligence, 12: 704-710.
7. Bouslama, F. and A. Amin, 1998. Pen-based recognition system of Arabic character utilizing structural and fuzzy techniques. Second Intl. Conf. Knowledge-based Intelligent Electronic Systems, 21-23 Apr., Adelaide, Australia. Jain, pp: 76-85.
8. Alimi, A. and O. Ghorbel, 1995. The analysis of error in an off-line recognition system of Arabic handwritten characters. Proc. ICDAR, Montreal, Canada, pp: 890-893.
9. Harty, R.A. and H.M. El-Zabadani, 2005. An offline Arabic handwriting recognition system. Intl. J. Computers and Applications.
10. Farah, R.A., M.T. Khadir and M. Sellaim, 2005. Artificial neural network fusion: Application to Arabic words recognition. ESANN 2005 Proc., Eur. Symp. Artificial Neural Networks, Bruges (Belgium), pp: 27-29.
11. Yihong, X. and G. Nagy, 1999. Prototype extraction and adaptive OCR. IEEE Trans. Pattern Analysis and Machine Intelligence, 21: 12.