HandwrittenText Recognition System for Automatic Reading of Historical Arabic Manuscripts

M. S. Farag
Math Department, Computer Science
Al-Azhar University, Cairo, Egypt Egypt &
Universiti Teknologi Malaysia, Johor, Malaysia

ABSTRACT

This paper presents an Arabic handwritten text recognition system for historical Manuscripts using the Matlab software, the paper is composed from number of stages, the first stage giving a short description of related work in handwritten Arabic recognition systems, the second stage discuss the preprocessing methods which contain of filtering, a certain methods will be applied on samples of database images todetect the best filter, normalization and cropping text for feature extraction, the third stage is the text segmentationinto lines, words, detecting the dots and remove it from the word with saving its position before segmentation to its primitives, the fourth stage gives a practical approach to the character recognition using a proposed multimodal technique by applying three techniques of character recognition, artificial neural network, hiddenmarkov modeland alinear classifier, saving the result into an array choosing the mode of thedata stored in the array,finally giving some experimental results.

Keywords:-historical document, OCR, neural recognizer, Islamic Manuscripts, off-line characters recognition.

1. INTRODUCTION

Optical character recognition (OCR) is computer software designed to translate images of typewritten or handwritten text into machine-editable text encoded in a standard encoding scheme, Automatic recognition of handwritten words can be classified into two different approaches. The first approach, offline, uses the images as input for the recognition steps. The second approach, online, uses the trace of a pen for the classification and recognition of the input information. Offline recognition of handwritten cursive text is more difficult than online recognition because the former must deal with two-dimensional images of the text after it has already been written. Arabic belongs to the group of Semitic alphabetical scripts in which mainly the consonants are represented in writing [1-4], while the markings of vowels (using diacritics) is optional and is rarely used. Arabic is spoken by more than 300 million people and is the official language of many countries. Enormous amount of research has been undertaken in the field of recognizing typed and handwritten Latin, Chinese, and Indian letters. Little progress, however, has been made in the recognition of Arabic letters, mainly due to their cursive nature. Unlike most of the other languages, both typed and hand-written Arabic letters are cursive. Furthermore, Arabic letters can take more shapes than Latin letters.

Other problems facing Arabic letter recognition systems include:

a) The unevenness of Arabic fonts; i.e., a certain letter in a specific font can be misinterpreted as a different letter in another font. In Arabic, some letter pairs may be combined together to form another letter, that is often referred to as a ligature. The only mandatory ligature is the (Lam Alef). Other ligatures are optional. Ligatures greatly complicate the segmentation task of an Optical Character Recognition (OCR) system.

b) Arabic has 28 letters, each of which can be linked in three different ways or separated depending on the case. Therefore, each letter can have up to four different forms depending on its position.

c) Arabic letters have different heights, which puts an extra burden on the noise detection task of the OCR system.

d) Line mingling, a phenomenon exhibited by improperly spaced documents. Many applications require offline HWR capabilities such as bank processing, mail sorting, document archiving, commercial form-reading and reading historical document so we can use it . Historical document considered as an important part of Arabic cultural heritage. The automatic pre-processing of Arabic collections from different historical periods in order to restore and use, is a definite advantage which is confronted with many difficulties due to the storage condition, natural time consuming and the complexity of their content. Consequently, different degradations have led to resulting many artifacts and distortions on those documents such as show-through effects, interfering strokes, background spots and coins, humidity absorbed by paper, curvature effect. Moreover, the development of analysis system to process automatically this type of digitized document is difficult and ambitious due to the wide variety of metadata containing. In general, the noticed metadata is main text body, illustrations, punctuation, titles, written annotations, drawings, and rarely ornaments.

The paper is organized as follows: Section 2 introduces the literature review (previous studies); section 3 presents the system architecture; section 4 presents the preprocessing operations, cropping, filtering, normalizationand segmentation; section 5 introduces a neural recognition and a proposed algorithm for character recognition;section 6 presents the experimental results; section 7 for future research and, section 8 gives conclusions.

2 LITERATURE REVIEW

Hanan Aljuaid,Zulkifli Muhammad and Muhammad S. [5]

This paper present a recognition system to recognize Arabic characters the preprocessing and the segmentation stages must be done before the feature extraction and the recognition stages

handwriting recognition system using genetic algorithm, handwriting samples in AHPD-UTM has been used. This study shows that the Arabic database AHPD-UTM is possible to obtain interesting off-line Arabic handwriting recognition rate. The system was used to recognize Arabic characters in all their form. First of all, the projection and thinning method used to solve segmentation and feature extraction method. The conjunction method solve the over segmentation problem. The recognition problem solved by genetic algorithm which may yield a different solution for the same word in each time that depends in the number of population and iteration, but there are a big similarity between the original solution and the proposed. To solve that problem the system keep the three first solutions in the early iteration.

Gheith A. Abandah, Khaled S. Younis and Mohammed Z. Khedher (SPPRA 2008)[6]

This paper explores best sets of feature extraction techniques and studies the accuracy of well-known classifiers for Arabic letters. Depending on their position in the word, the system applied the Principal Component Analysis (PCA) as a preprocessing step to transform the data to a new space where the features are uncorrelated in this space, Among the 5 studied classifiers, the Quadratic Discriminant Analysis(QDA)classifier gives best accuracy with 20 features or less. The Linear Discriminant Analysis(LDA) classifier gives best accuracy with more features. The Diagonal Quadratic Discriminant Analysis(DQDA) and Diagonal Linear Discriminant Analysis (DLDA)classifiers' accuracies are about 10% lower than the LDA classifier's accuracy. The 3NN classifier has very low accuracy.

Using four classifiers each tuned for one letter form gives about 11% better accuracy than using one classifier for all forms. And, with small number of features, the four classifiers are also more accurate than one classifier with the additional letter form information.The final and initial forms are easier to recognize than medial and isolated forms. The highest recognition accuracy we have achieved is 87% using LDA classifier. This accuracy was limited by low recognition of some medial and isolated forms. Better feature extraction techniques are needed for letters with dots above the main body because of the variations in drawing these dots. Also better classification techniques are needed for these forms.

Wafa Boussellaa,Abderrazak Zahour,Bruno Taconet Adel Alimi, Abdellatif Benabdelhafid (PRAAD)[7]

Present a new system PRAAD for preprocessing and analysis of Arabic Historical Documents. It is composed of two important parts: Pre-processing and analysis of ancient documents. After digitization, the color or grayscale ancient documents images are distorted by the presence of strong background artifacts such as scan optical blur and noise, show-through and bleed-through effects and spots. In order to preserve and exploit this cultural heritage documents, we intend to create efficient tool that achieves restoration, binarisation, and analyses the document layout. The developed tool is done by adapting their expertise in document image processing of Arabic Ancient documents, printed or manuscripts. The different functions of PRAAD system are tested on a set of Arabic Ancient documents from the National library and the National Archives of Tunisia.

Ahmad M. Sarhan, and Omar I. Al Helalat[8]

In this paper, an Arabic letter recognition system based on Artificial Neural Networks (ANNs) and statistical analysis for feature extraction is presented. The ANN is trained using the Least Mean Squares (LMS) algorithm. In this paper, a new system for the recognition of typed Arabic letters is presented. The system is composed of two stages, a feature extraction stage, followed by an ANN classification stage. The performance of the system was compared to the traditional solution that bypasses the feature extraction stage and consists only of an ANN stage.A statistical analysis on the Arabic letters was performed and showed that the pixel values of the Arabic letters are highly uncorrelated.

Volker Märgner – Haikal El Abed – Mario Pechwitz [9]

The system in this paper uses Hidden Markov Models (HMM) for word recognition, and is based on character recognition without explicit segmentation. For recognition again basically a standard Viterbi Algorithm is used. The recognition process has to perform the task to assign to an unknown feature sequence a valid word from the lexicon. The basic way to do this is to calculate the probability that the observation was produced by a state sequence for each word of the lexicon. The sequence with the highest probability gives the correct word. The system using the frame-based HMM approach to recognize handwritten Arabic words are very promising. Nevertheless there is still a lot of work to do.

3-SYSTEM ARCHITECTURE

System architecture shows the stages of the system, see figure(1).

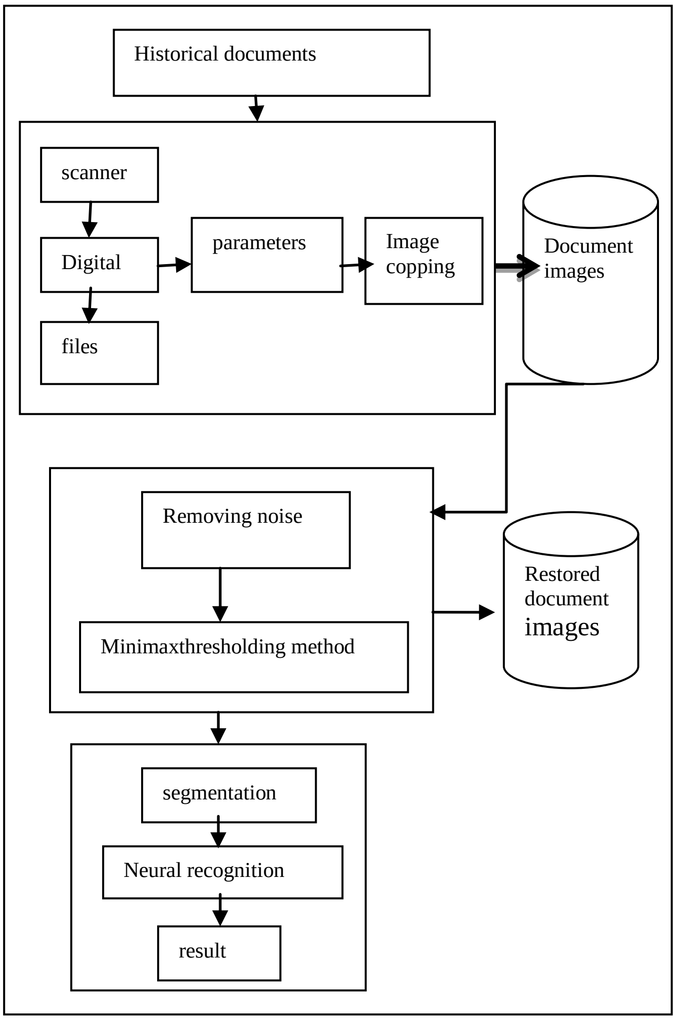


Fig.1. system architecture

4-PREPROESSING

In the following steps, basic pre-processing task such as filtering, normalization, feature extraction and image segmentation must be performed.

4.1- Cropping

Cropping refers to the removal of the outer parts of an image to improve framing, accentuate subject matter or change aspect ratio. Depending on the application, this may be performed on a physicalphotograph, artwork or film footage, or achieved digitally using image editing software. The term is common to the film, broadcasting, photographic, graphic design and printing industries.The Crop Image tool is a moveable, resizable rectangle that you can position interactively using themouse.When the Crop Image tool is active, the pointer changes to cross hairswhen youmove it over the target image. Using the mouse, you specify the crop rectangle by clicking and dragging themouse. You can move or resize the crop rectangle using the mouse. When you are finished sizing and positioning the crop rectangle.

4.2 Filtering

Removing noise is to remove information coming from the background such as show-through effects,interfering strokes due to seeping of ink during long period of storage, spots of humidity and curvature effect. Example for document before using filter method, see figure(2).

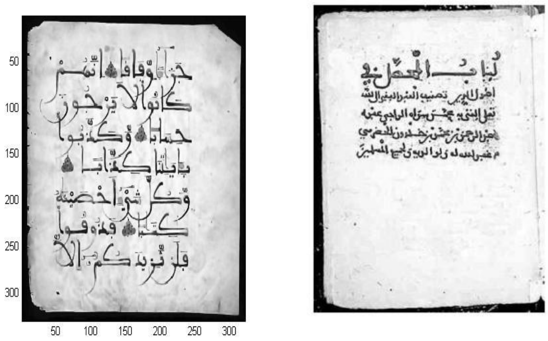


Fig.2. document before using filter

Applying the following methods to remove noise from historical Manuscripts, all filtering methodswill apply on Figure (2).

4.2.1Median Filtering

In median filtering, the neighboring pixels are ranked according to brightness (intensity) and the median value becomes the new value for the central pixel, see figure(3). the value of an output pixel is determined by the *median* of the neighborhood pixels.

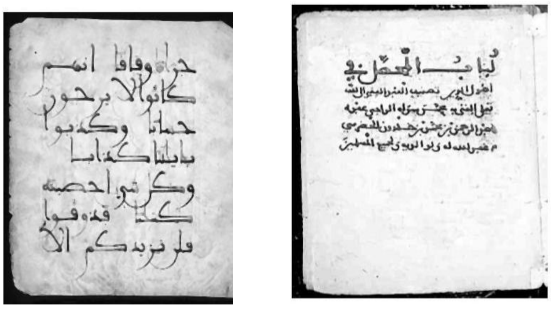


Fig.3. using median filter

4.2.2Adaptive Filtering[10]

A popular adaptive filter for reducing random noise in signals is the Wiener filter, see figure(4). Its two-dimensional version can be used to recover noisy images.

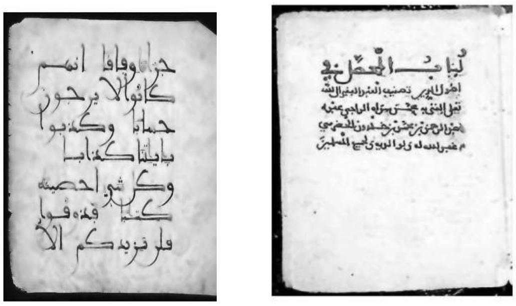


Fig.4. using the Wiener filter

4.2.3 Linear isotropic diffusion

Linear isotropic diffusion is mostly used for smoothing images, see figure(5).

$$\partial_t \mathrm{u} = \mathrm{div}\ (\mathrm{u})$$
$$\mathrm{u}(\mathrm{x},\mathrm{y},0) = \mathrm{I}(\mathrm{x},\mathrm{y})$$

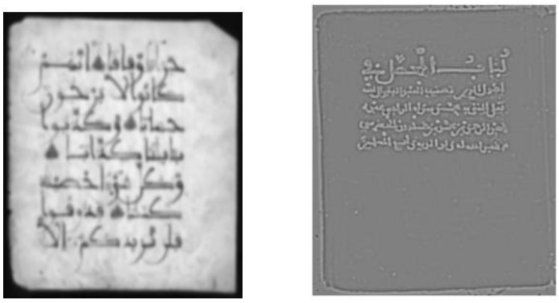


Fig.5.after using linear diffusion filter

4.2.4nonlinear isotropic diffusion filter[11]

which is a technique aiming at reducing image noise without removing significant parts of the image content, typically edges, lines or other details that are important for the interpretation of the image. Anisotropic diffusion resembles the process that creates a scale-space see figure(6), where an image generates a parameterized family of successively more and more blurred images based on a diffusion process. Each of the resulting images in this family are given as a convolution between the image and a 2D isotropic Gaussian filter, where the width of the filter increases with the parameter. This diffusion process is a *linear* and *space-invariant* transformation of the original image.

The filtering process consists of updating each pixel in the image by an amount equal to the flow contributed by its four nearest neighbors.

two functions have been suggested :

$$c_1(\bar{x},t) = \exp\left(-\left(\frac{|\nabla I(\bar{x},t)|}{k}\right)^2\right)$$

$$c_2(\bar{x},t) = \frac{1}{1+\left(\frac{|\nabla I(\bar{x},t)|}{k}\right)^{1+\alpha}} \propto > 0$$

kis referred to diffusion constant or flow constant diffusion equation 1, favours high contrast edges over low contrast ones.

Diffusion equation 2, favours wide regions over smaller ones. The greatest flow is produced when the image gradient magnitude is close to the valueofk. Therefore, by choosing k to correspond togradient magnitudes produced by noise, the diffusion process can be used to reduce noise in images. Assuming an image contains no discontinuities, object edges can be enhanced by choosing a value of k slightly less than thegradient magnitude of the edges.

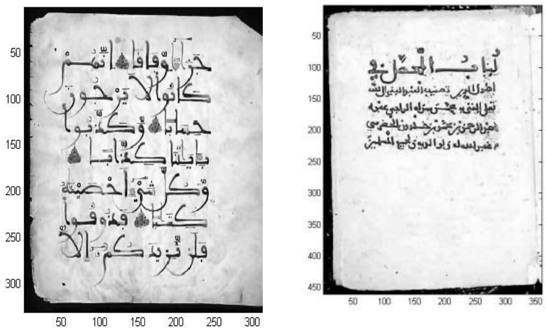


Fig.6. historical document after using non linear anisotropic diffusion filter

4.2.4 Advantage and disadvantage of filtering methods

This section shows the results of comparison among number of filters (advantages, disadvantages) see table 1, showing that the nonlinear isotropic diffusion is the best filter.

Table 1.Advantage and disadvantage of filtering methods

Filter name	**advantages**	**disadvantages**
1-Medianfilter	– No reduction in contrast across steps, since output valuesavailable consist only of those present in the neighborhood (no averages). – Median filtering does not shift boundaries, as can happenwith conventional smoothing filters (a contrast dependentproblem). – Since the median is less sensitive than the mean to extremevalues (outliers), those extreme values are more effectively removed.	Median filters can tend to erase lines narrower than ½the width of the neighborhood. They can also round offcorners.
2-Adaptive filter	- it is more selective than linear filter - preserves edges and other high-frequency parts of the image - less designing tasks are necessary	The Adaptive filter is more expensive to compute than a smoothing filter
3- Linear isotropic diffusion filter	•Continuously simplifying of the image •Reducing the noise in the image	•Linear isotropic diffusion does not only reduce noiseit also blues important features like edges •No a-priori knowledge is taken into account •Result:it makes edges harder to identify
4-nonlinear isotropic diffusion filter	Nonlinear anisotropic diffusion Combination of two features: •Non-linearity -The diffusion at border is much less than the diffusion elsewhere •Anisotropy -Diffusion should be perpendicular to edge -No diffusion over edges	--------

4.3Normalisation

Thresholding is the operation of converting a grayscale image into a binary image[12]. Thresholding is a widely applied preprocessing step for image segmentation. Often the burden of segmentation is on the threshold operation, so that a properly thresholded image leads to better segmentation.

There are mainly two types of thresholding techniques available:

a) Global.
b) Local.

4.3.1 Globalthresholding

Global thresholdingis a technique a grayscale image is converted into a binary image based on an image intensity value called global threshold see figure(7). All pixels having values greater than the global threshold values are marked as 1 and the remaining pixels are marked as 0.

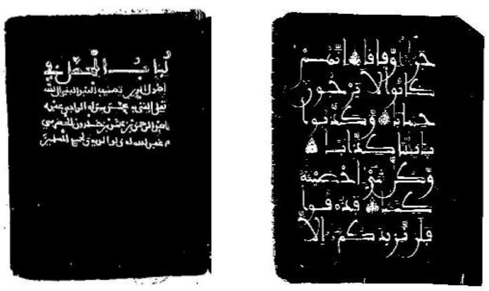


Fig.7. using global threshold method

4.3.2OTSUmethod [13]

OTSU proposed a global image thresholding technique where the optimal global threshold value is determined by maximizing the between–class variance with an exhaustive search. Although OTSU’s method remains one of the most popular choices for global thresholding techniques see figure(8), it does work well for many real world images where a significant overlap exists between the pixel intensity values of the objects and the background for un-even and poor illumination.

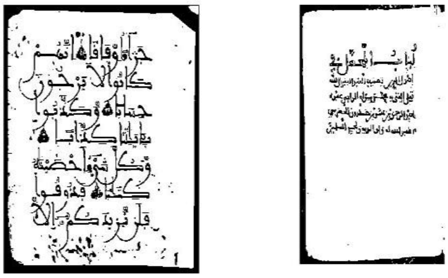


Fig.8. using OTSU method

4.3.3 Local thresholding[14]

Methodwhere the thresholding operation dependson local image characteristics is superior to the global ones for poorly illuminated images.Incidentally, it is noted that the local thresholding techniques have hand tuning parameters that need to be adjusted for differently illuminated images and the values of these parameters vary significantly for different images, see figure(9).

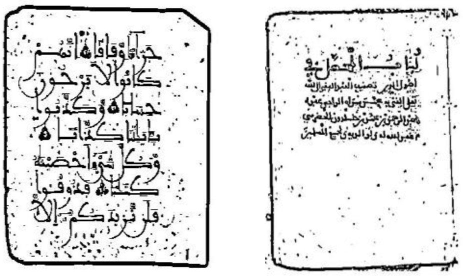


Fig.9. local adaptive method

4.3.4 MiniMax method[15]

Using here an automated adaptive local image thresholding method where no manually-adjusted, weighting parameter is present for the data and theregularization terms in the energy functional. We use a variational energy functional consisting of a non-linear combination of a data and a regularization term.

The energy functional is a function of the threshold surface, the image, as well as the weighting parameter, see figure(10).This makes a balance between the data and the regularization terms. A minimax solution of the proposed energy functional is obtained iteratively by alternating minimization and maximization of the energy functional respectively with regard to the threshold surface and the weighting parameter.

The novelty of this method is that from an image it automatically computes the weights on the data fidelity and the regularization terms in the energy functional, unlike many other previously proposed variational formulations that require manual input of these weights by laborious trial and error.

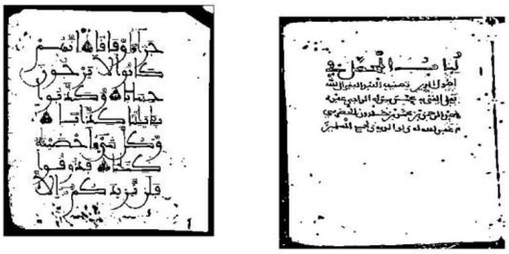
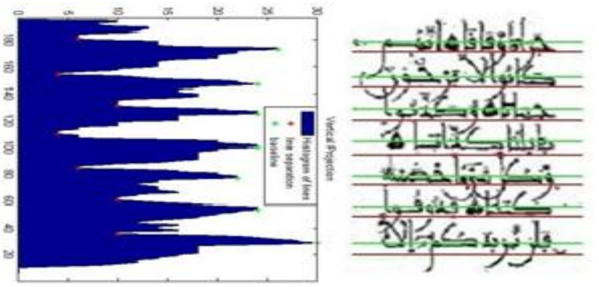


Fig.10. using minimax method

4.4- Segmentation:

Automatic segmentation can be done by dividing the paragraph into linesusing horizontal projection see figure (11), each line into words using drawing rectanglearound each word in the line see figure (12) and so every word segmented into its primitives using vertical projection.



(11 a) **(11 b)**

Fig.11. a. minimum and maximum peaks, b. minimum peakspresents the lines separation among text lines and the maximum peaks presents the baselines.

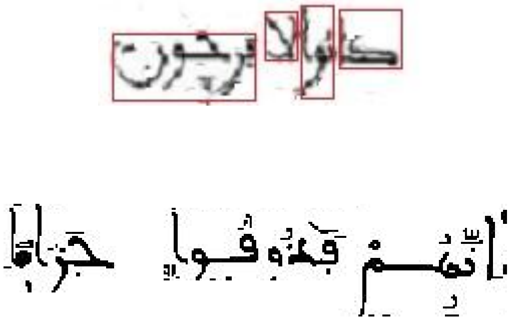


Fig.12.segmented word

Sample of the collected database see figure (13).

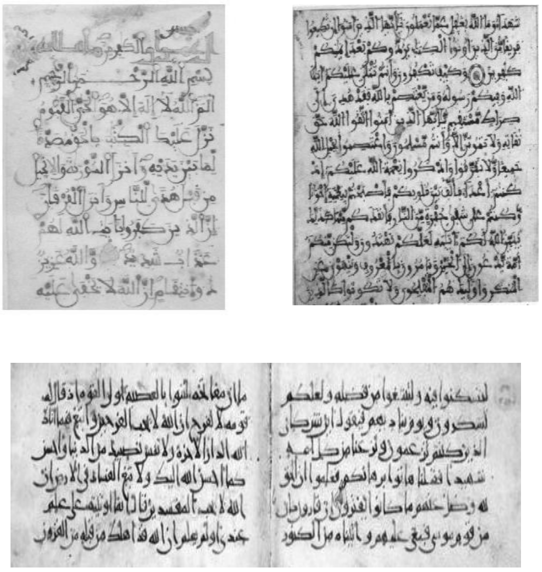


Fig.13. sample of the database

5-NEURAL RECOGNITION:

In this section, applyingthe algorithm with Optimized Error Based on Modified Gram Schmidt with Reorthogonalization MGSR[15-16], sample data of training characters as in figure(14).

The Algorithm:

The entire training procedure based on the MGSR deficiency is then summarized as follows.

1) Randomly initialize w^0 to orthogonal vectors of small random values using MGSR.

2) Forward-propagate function signals to obtain $y_{(i)} \quad (i = 1,2,\cdots,m), e$ as will as $E(w^k)$ in (1).

3) If $E(w^k) \le \varepsilon$ or the maximum number of iteration come then the trainingobjective is met and the training stops; otherwise go on.

4) Backpropagate error signals to obtain the local gradients of each node

a- $\delta_k = (t_k - y_k)y_k(1 - y_k)$ for each outputnode.

b- $\delta_j = y_j(1 - y_j)\sum_{k=1}^{m}\delta_k w_{jk}$ for each node in the hidden layer.

5) We apply the MGSR procedure from input to hidden layer and from output to hidden layer to be orthogonal matrices.

6) We update the weights

a- $w_{oh}(t+1) = w_{oh}(t) + \eta\delta_k y_j + \mu\Delta w_{oh}(t)$

for each output node

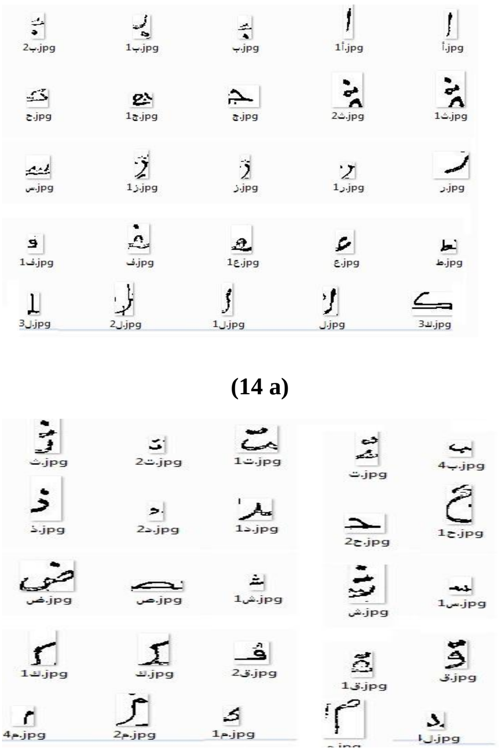
b- $w_{ih}(t+1) = w_{ih}(t) + \eta\delta_j x_i + \mu\Delta w_{ih}(t)$

for each node in the hidden layerwhere η is the learning term, μ is the momentum term,

and $\Delta w_{oh}(t) = w_{oh}(t) - w_{oh}(t-1)$,

$\Delta w_{ih}(t) = w_{ih}(t) - w_{ih}(t-1)$

7) go back to step 2



(14 a)

(14 b)

Fig.14.a sample of training data set

5.1- A proposed algorithm for character recognition

1- Apply the artificial neural network see section 5, hidden markov model and alinear classifier on the training data set.
2- Store the result into an array.
3- Select the mode in the array.
4- If the mode at least two, then the result is a high accuracy
Else goto step one, after changing the parameters in the Techniques to optimize the errors (increasing the number of epochs, decreasing the target error rate,…).

This algorithm can be apply in the segmentation stageon the image by applying different techniques, counting the number of lines and the number of words in each line, if the mode at least two continue to the next stage else repeat the segmentation stage another one with changing their parameters.

6- EXPERIMENTAL RESULTS

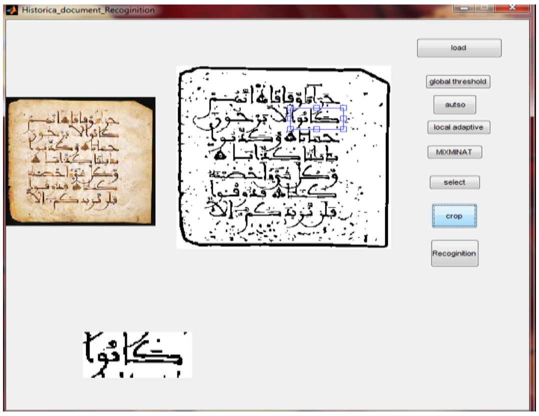


Fig.15. system model

The matlab system implementation as shown in figure (15):

- load button :load the historical document.
- global threshold button: thresholding using global threshold method after using non linear isotropic diffusion filter.
- OTSUbutton:thresholding using OTSU threshold method after using non linear isotropic diffusion filter.
- Minimax button:thresholding using minimax threshold method after using an isotropic diffusion filter.
- Crop button: detect the image and crop draw a bound box around words and automatically segmented forlines, words and characters.
- Recognition button: recognize charactersusing advanced neural-network training algorithm with optimized error training algorithm with optimized error based on modified gram schmidt with reorthogonalization and store the result in notepad file.

7-FUTURE WORK:

In the future segmentation process and Neural recognition approach to recognize handwritten Arabic wordsare very promising,nevertheless there is still a lot of work to do.

8-CONCLUSION

The system introduce a good way to recognize text from historical document after discussing a number of filtering technique and choosing the best method,non linear isotropic diffusion filter for removing noise, thresholding using OTSU method, automatically segmented image into lines using horizontal projection, lines into words using boundary box and words into characters using vertical projection with no overlap among text lines and finally recognize characters usingadvanced neural-network training algorithm withoptimized error based on modified gram schmidt withreorthogonalization and proposing a multimodal algorithm for increasing the performance of both segmentation and recognition.

REFERENCES

[1] Zaher Al Aghbari, Salama Brook, "HAH manuscripts: A holistic paradigm for classifying and retrieving historical arabic handwritten documents Original Research Article", Expert Systems with Applications, Volume 36, Issue 8, October 2009, Pages 10942-10951.

[2] Jin Chen, Daniel Lopresti, "Model-based ruling line detection in noisy handwritten documents", Pattern Recognition Letters, In Press, Corrected Proof, Available online 15 September 2012.

[3] Jun Tan, Jian-Huang Lai, Chang-Dong Wang, Wen-Xian Wang, Xiao-XiongZuo, "A new handwritten character segmentation method based on nonlinear clusteringNeurocomputing, Volume 89, 15 July 2012, Pages 213-219.

[4] . Ntirogiannis, B. Gatos, I. Pratikakis, "A combined approach for the binarization of handwritten document images", Pattern Recognition Letters, In Press, Corrected Proof, Available online 11 October 2012.

[5] HananAljuaid,Zulkifli Muhammad and Muhammad Sarfraz. "A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach"(Journal of Computer Science 6 (5): 496-501, 2010 ISSN 1549-3636 © 2010 Science Publications).

[6] Gheith A. Abandah, Khaled S. Younis and Mohammed Z. Khedher "Handwriting Arabic Character RecognitionUsingMultiple Classifiers Based On Letter Form "Fifth IASTED International Conference on Signal Processing, Pattern Recognition and ApplicationsPages 128-133 ACTA Press Anaheim, CA, USA ©2008.

[7] WafaBoussellaa, AbderrazakZahour, Haikal El Abed, AbdellatifBenAbdelhafid, Adel M. Alimi: Unsupervised Block Covering Analysis for Text-Line Segmentation of Arabic Ancient Handwritten Document Images. ICPR 2010: 1929-1932.

[8]Ahmad M. Sarhan, and Omar I. Al Helalat"Arabic Character Recognition using Artificial Neural Networks and Statistical Analysis"*Proceedings of World Academy of Science, Engineering and Technology* Volume 21 May 2007 ISSN 1307-6884.

[9] VolkerMärgner – Haikal El Abed – Mario Pechwitz."OfflineHandwritten ArabicWord Recognition Using HMM - a Character Based Approach without Explicit Segmentation “Eighth International Conference on Document Analysis and Recognition (ICDAR 2005), 29 August - 1 September 2005, Seoul, Korea.

[10] Holger R. Roth "ADAPTIVE FILTERS"17th March 2008.

[11] PietroPerona and Jitendra Malik1990 "Scale-space and edge detection using anisotropic diffusion"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12 (7): 629–639.

[12] N. Ray and B. Saha, “Edge sensitive variational image thresholding,” proceeding of: Image Processing, 2007. IEEE International Conference on, Volume: 6.

[13]M. Sezgin and B. Sankur (2003). "Survey over image thresholding techniques and quantitative performance evaluation". *Journal of Electronic Imaging* 13 (1): 146–165.

[14] J. Sauvola and M. Pietikainen, “Adaptivedocument image binarization,” Pattern Recognition 33(2),pp. 225–236, 2000.

[15] I. A. Ismail, M. S. Farag 2006. " Advanced Neural-Network Training Algorithm with Optimized Error Based on Modified Gram Schmidt with reorthogonalization", International Journal of Intelligent Computing and Informational Sciences, Vol. 6, pp 69-74.

[16] Chaivatna Sumetphong, Supachai Tangwongsan 2006 "Modeling broken characters recognition as a set-partitioning problem", Pattern Recognition Letters, Volume 33, Issue 16, 1 December 2012, Pages 2270-2279.