ISSN 2278-3091
Volume 8, No.1.1, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4581.12019.pdf
https://doi.org/10.30534/ijatcse/2019/4581.120419

Arabic Handwriting Recognition Model based on Neural Network Approach

Manal Abdullah [#1], Afnan Agal [#2], Mariam Alharthi [#3], Mariam Alrashidi [#4]

[#]Computer Science Department, Faculty of Computing and Information Technology FCIT
King Abdul-Aziz University, KAU Jeddah, Saudi Arabia SA
[1]maaabdullah@kau.edu.sa, [2]aagal0001@stu.kau.edu.sa,
[3]malharthi0334@stu.kau.edu.sa, [4]malrashidy0006@stu.kau.edu.sa

ABSTRACT

Arabic language as the main language of more than millions of people all over the world has attracted researches in the handwriting script recognition field. Arabic scripts have many difficulties which make Arabic Handwriting recognition challenging. In this paper the model proposed aimed at recognizing Arabic words the IFN / ENIT dataset. The model is an OCR using Neural Network classifier preceded by a set of preprocessing techniques includes removing spaces between words, bolding the words, baseline estimation and correction and resizing the words images. The recognition rate of the proposed model is 70%. This result showed how much the proper selection of the preprocessing steps affects the recognition rate of handwritten words. The propo sed model differs from other suggested models in Arabic handwriting field, according, the unlike preprocessing steps applied on a model and a new approach in estimating and correcting words baseline.

Key words— OCR, Neural Network, baseline, PAWs, Arabic, Handwritten, Recognition

1. INTRODUCTION

Any person can read whatever is written by hand/displayed either one in ordinary handwriting or in printed layout. The same effort can be done by the machine where it is named handwriting recognition HR. Handwriting recognition has been a widespread field of research ever since a few years under the specialization of pattern recognition and image processing. HR can be divided into two types: off-line and on-line [1]. Off-line handwriting recognition receipts an image from a scanner (photographed images of the paper documents), digital camera, or other digital source. The input image is binaries within threshold technique depend on the color pattern is it colored or gray scale image, so that the image pixels can be 1 or 0. In on-line, the data is immediately receipt while user writing on the digital tablet. Fundamentally, it receives a string of (x, y) coordinates pairs from the digital tool such as electronic pen touching the digital tablet.

More than one reason recognizes Arabic language different from other languages according to its shape and the writing style [2]. Here are some of these reasons:

- First, Arabic letters/characters come in different forms depending on their location in the word. These locations are beginning of a word, middle, ending of a word and single shape when the letter is un-connected as shown in Table 1.

Table 1: Arabic characters form [2].

Letter	single	Beginning	Middle	Terminate
Alif	ا	ا	ـا	ـا
Baa	ب	بـ	ـبـ	ـب
Taa	ت	تـ	ـتـ	ـت
Thaa	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Haa	ح	حـ	ـحـ	ـح
khaa	خ	خـ	ـخـ	ـخ
Dal	د	د	ـد	ـد
Dhal	ذ	ذ	ـذ	ـذ
Raa	ر	ر	ـر	ـر
Zai	ز	ز	ـز	ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Sad	ص	صـ	ـصـ	ـص
dad	ض	ضـ	ـضـ	ـض
Taa	ط	طـ	ـطـ	ـط
Dhad	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Ghain	غ	غـ	ـغـ	ـغ
Faa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Haa	ه	هـ	ـهـ	ـه
Waw	و	و	ـو	ـو
Yaa	ي	يـ	ـيـ	ـي

- Second, numerous Arabic characters have similar main body and we can differentiate them from each other by particular extra strokes that may be dots or Hamza.

- Third, Arabic words written as joined characters or cursive way and the letters connected except if one of the following characters (أ, د,ذ,ر,ز,و) exist in the middle of the word to generate sub words .

- Fourth, a number of Arabic words may have overlap circumstance, especially in Arabic handwritten result from the style of the writer who writes a text.

The importance of constructing OCR system that be able to convert handwritten words written in Arabic into a computer format to process is making words editable and reusable by any user instead of rewriting each word manually. In this paper, a model is proposed for Arabic handwritten word recognition using Neural Network Classifier that is applied on and tested against the IFN / ENIT dataset. In Section Ⅱ, the paper presents an overview about Arabic language and OCR. In Section Ⅲ, some efforts in related domain are presented. Section Ⅳ describes the proposed model. Finally, Section Ⅴ and Ⅵ present result and conclusions.

2. BACKGROUND

Arabic language is the fundamental language of in excess of 280 million individuals as local language and around 250 million individuals are talking in Arabic as a second language. This language comes as the fifth position of the most ordinarily utilized dialects on the planet. There are some different dialects identified with the Arabic language as these dialects have a few similitudes with it whence the character shapes or from the articulation. These dialects are Jawi, Persian, Bengali, and different dialects [2].

A. Handwritten Recognition History

Optical character recognition system (OCR) has been researched for numerous past decades. During the year 1914 Emanuel Goldberg developed a system that reads handwritten numbers and letters and then transferred to telegraph code. During the same interval of time Edmund Fournier d'Albe evolved the Optophon, a handheld scanner that shot the printed page and gave the output. In 1985, structural methods were suggested to work beside the statistical methods. After 1990, a real advance is accomplished using additional techniques and methodologies in image processing and pattern recognition fields. In 1994, RCA Engineers Corporation offered the first simple computer type with optical character recognition to aid the blind people. It is proposed to transfer the handwritten statement into stamped cards as an input to the computer for assisting in handling delivery of million books [3].

In the current days there is a more effective computer and more precise tools like electronic pen, scanner, and tablets are worked and used for this purpose. A lot of approaches like Hidden Markov Model, Neural Network NN, back propagation algorithm, fuzzy neural network are employing to be familiar with handwritten documents. The computerized identification and recognition of a text on scanned images has permitted many applications, for example searching for words in huge documents that has a big amount of data, computerized arranging of postal mails, bank cheques and allow easily editing previously printed documents or files. The text in manuscripts is usually well-organized than text written by hand; the recognition mission is maybe easier. But still the image distortion, unexpected patterns, and unseen writing shapes and styles offer challenges to the recognition system [4], [5].

B. Recognition Engine

The acknowledgment gadget or motor can be rule-based, probabilistic, or a mix. The methodologies talked about in [4], utilize counterfeit neural systems, concealed Markov models, guidelines, or cross breeds of principles with factual techniques.

1) Overview of OCR

Optical Character Recognition is a method by which a person is able to transform printed text or scanned page in ASCII character that a computer can recognize and understand [5]. It goes through four stages as shown in Figure 1.

- First, the preprocessing stage, which includes thinning and bold a word, baseline estimation and correction, reducing space between parts of words and image resizing.
- Second, feature extraction were segmentation and figure size function used for this purpose.
- Third, it is classifying Arabic word included in the dataset by OCR includes training phase and testing phase.
- Finally is the analysis phase where the recognized word written in notepad.

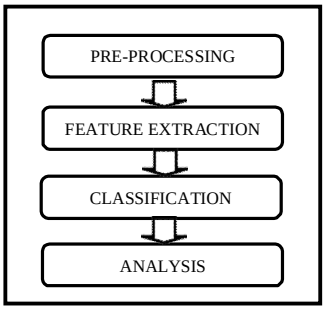

Figure 1: Stages of OCR Recognition System [6].

2) Design Approaches of OCR[1] ,[4]

Several approaches worked in the design of OCR systems can discuss briefly as follows:

- Matrix Matching: it transforms every single character into a pattern or style inside a matrix. After that, it follows by matching the pattern with an index of known characters.
- Fuzzy Logic: it is numerous value logic that permits intermediate values to be stated between traditional evaluations like yes/no, true/false, black/white, etc. This approach is benefited once solutions do not have a distinct true or false value and there is uncertainly included.
- Feature Extraction: this technique identifies each character by the existence or nonappearance of main features, involving some important features such as height and width of the character, loops, and additional character characteristics.
- Structural Analysis: structural Analysis classifies characters via testing some features including the shape or the general form of the tested image, also studying them with vertical and horizontal histograms.

- Neural Networks: this approach is biological inspired as it models and simulates the technique in which the human neural system operates. The advantage of this approach is the ability to identify characters via abstraction which is wonderful for faxed files and harmed text.

Artificial Neural Networks (ANNs), which we adopt in this research, consist of simple processing nodes and a high level of connection. The weights inside the nodes are educated and learn from training information or data. The nodes are structured into three layer types: input layer, intermediate (also called hidden layers), and the last is called output layer [1]

3. RELATED WORK

A lot of researches have been proposed to address OCR AHR problems. Using off-line recognition, AlKhateeb et.al. [6], proposed a model to recognize word using K-nearest Neighbor classifier. Researchers test the proposed system using the IFN / ENIT database by recognizing the entire word image instead of recognizing characters one by one with a recognition rate of 76%. Also, Sarfraz et.al in [7] proposed an off-line Arabic handwritten system to recognize word using Artificial Neural Networks which composed of three phases with a recognition rate of about 73%.

In [8], AL-Shatnawi and Omar categorized baseline detecting techniques into four categories: baseline detection, the word basic structure, text contour representation and lastly, principal components analysis (PCA). They also address the difficulties in detecting AHR baseline with the advantage and disadvantages of each technique. In [9], Khemiri proposed system includes three major phases: first was baseline estimation, then came features extraction phase followed by word classification. In baseline estimation, sequences of sub-baselines are taken into consideration. First, they extract by portions of the Arabic words and refinement of damh, kasrah and sokon that actually called diacritic signs. The baseline of each Part of Arabic Word can be predictable by the juxtaposition of its PAW baselines. In this way a structural feature can be extracted from the word. In [10], authors proposed OCR system for Urdu language used in Pakistani and Indian text to recognize individual characters. Where they use supervised learning to train and teach the feed forward NN with a recognition rate of 98.3%

In [11], Addakiri and Bahaj suggested an on-line recognition system to recognize the handwritten Arabic using NN. As ten writers tested the suggested system where, every writer wrote the Arabic 28 characters several times exactly was five times. The recognition rate was 83% for all characters, and they noted a system gives great results on the characters that include sharp edges. In [12], they offered an off-line recognition scheme for AHR using HMM. They used re-ranking to improve the accuracy and tested on the IFN/ENIT database. The results show that the system recognition rate with ranking is 83.55% and without ranking is 82.32%.

4. RESEARCH PROBLEM

Due to the less studies in the field of Arabic handwritten recognition, this study attempts to improve the recognition rate of AHR words. There are many factors that may affect the recognition of these texts that already mentioned in the introduction (Section I). NN is used with a set of steps in preprocessing stage to enhance the recognition rate while that applied on a specified dataset (IFN/ENIT).

5. PROPOSED SYSTEM MODEL

The proposed system includes four phases. First, the preprocessing stage, which includes reducing space between parts of words. Then image resizing is second phase. Thinning and bold a word then baseline estimation and correction are third and fourth phases respectively. The details about the system internal operations are shown in Figure 2. The second step is feature extraction where segmentation and figure size functions used for this purpose. The third step is classifying Arabic words using NN approach. This step includes training phase and testing phase on IFN/ENIT Dataset. Finally, the recognized word is written in notepad.

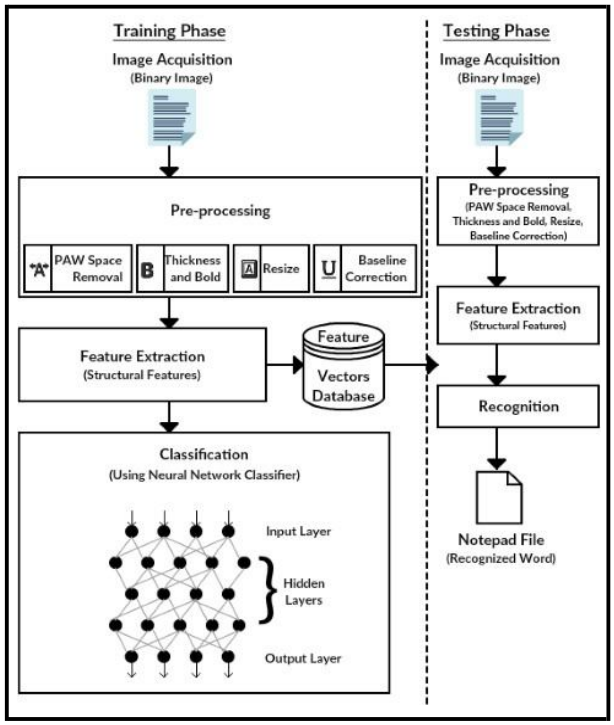

Figure 2: The Proposed System Model.

A. IFN/ENIT Dataset

IFN/ENIT is an open source database that is used in this paper and can be accessed freely. It was made by the Organization for Communications Technology at Technical University of Braunschweig (Institut f¨ur Nachrichtentechnik, IFN) and the l'Ecole Nationale d'In´egnieurs de Tunis. The

overall total of binary pictures of handwritten Tunisian city names exists is 26459. Names in a database were written by 411 persons, and they were categorized and labeled corresponding to 946 name class. The described database was benefited as an assessment tool to assess researcher's works in the field Document Analysis and Recognition [13].

This database, in fact, has a group of specials, which makes preprocessing is a challenging process. It has 24 classes of collecting names that having the hugest appearance regarding frequencies. Table 2 offers a summary of the total of images included in a database, its words, quantity of connected Part of Arabic Word, and quantity of characters in the IFN/ENIT database.

Table 2: Quantity of Town Name Images, Words, Paws, And Characters In Words [13].

Quantity of words in town names	Quantity of town name images	Quantity of PAW's	Quantity of characters
1	12992	40555	76827
2	10826	54722	98828
3	2599	20120	36004
4	42	188	552
Total	26459	115585	212211

Some statistics about the age of writers and career of each one are not shown in this research where it is not used.

B. Preprocessing

Preprocessing is the process of preparing the image for feature extraction then recognition. It includes a lot of operations that applied to the recognized word and remove the undesired features from the word image. In this research four preprocessing stages are applied as follows:

1) Remove spaces between words

The vertical projection of the info picture has been utilized to discover the spots where there are no pixels. Then, the image of the word is cropped into multiple segments, so that each segment contains part of a word (POW), means that no blank spaces are found in segments. Then, concatenation process is applied to join these segments. The word before space removal is shown on the right side of Figure 3, while word after space removal is on the left.

Figure 3: The result of removing spaces.

2) Image Resize

An imresize() method which is built-in method in Matlab tool is used in this preprocessing stage to make sure that all images either in training or testing sets have the same size.

3) Thinning and Bold

Thinning is the extraction of the skeleton of a word by reducing the thickness into 1 pixel. There are two approaches for thinning algorithm: iterative and non-iterative. The iterative method keeps processing the word image until no longer pixel to change while non-iterative method performs all operations once. In the proposed system, the iterative thinning algorithm is used. More deeply, the iterative approach can be classified into: parallel and sequential. The parallel algorithm is performed in this research. "Figure 4" shows before and after thinning algorithm.

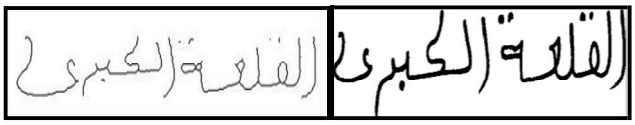

Figure 4: The result of thinning algorithm.

After thinning algorithm is applied, the word is bolded into 6 pixels. The purpose of thinning the word before bold it is to make sure that the entire word have the same number of pixels (1 pixel). The result of bolding process is shown in Figure 5.

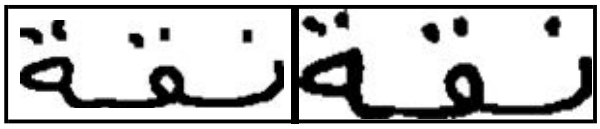

Figure 5: The result of bolded the word.

4) Baseline Estimation and Correction

AHR words are frequently written with a single baseline. This is due of its letter extensions and the form of its letter. Thus, the location of ascenders and descenders differs corresponding to the writing style used by the writer.

A series of sub-baselines are obtained [7] for each word. It begins getting parts of the Arabic words (PAWs) and filtering diacritic signs. For every word of n PAWs, the baseline of each PAW can be guessed using a baseline estimation algorithm. Then, the whole word baseline is designed through juxtaposition of its PAW baselines. The baseline correction technique presented in [8] is adopted with some changes to suit images rather than data. We extract the baseline for the entire word and then the extracted baseline is corrected to form a single baseline. Figure 6 shows the baseline of a word before and after correction.

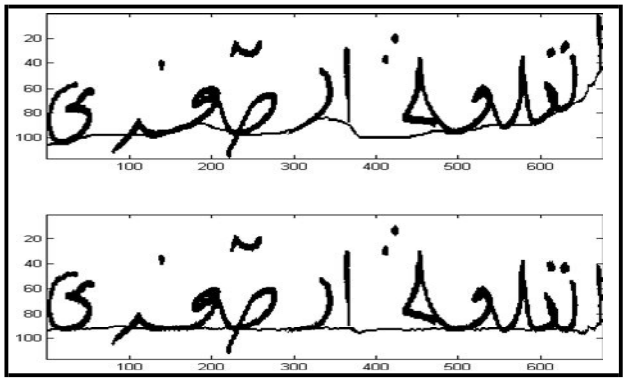

Figure 6: Baseline estimation and correction.

C. Feature Extraction

Selecting the suitable type of features is influenced by the characteristics of the text. Also, the kind of the system deal with, which may be online or offline. Nevertheless, feature types of recognizing any text written in any language can be classified into three principal types which are structural, statistical, and global feature transformation. Each type of them can be explained briefly as follows [9]

Statistical features and characteristics can easily be computed and can describe such features. Those features are zoning of pixels which are worked by partitioning the recognized text in the image into small areas and make use of the density of pixels in those small areas as features.

The proposed system extracts the features using figresize() method, which is built-in method in Matlab. The word image is cropped into several sub-images, where the sizes of these sub-images are not fixed. To standardize the sub-images and eliminate the spaces around the border of the character, the maximum row and column with 1s as well as the peak point are founded. This process will continue until reaching a line with all 0s, or simply a white space.

These sub-images are resized to 50x70 to make the input data for a network standard and feeding data properly. Then, calculate the average value for each 10x10 block. The size of each sub-image will become 5x7, where each will represent a feature. The 35 features or input are fed to the network. Mention that; the input indeed is a negative image of the figure. Means, 0 refers to black and 1 refers to white, while the value between them represents the intensity of the relevant pixel. "Figure 7" shows these steps.

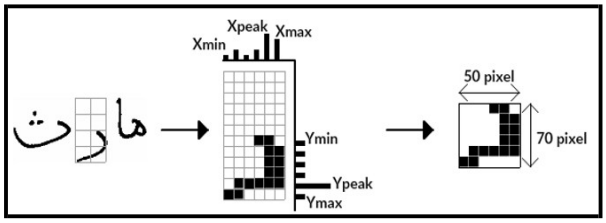

Figure 7: Feature Extraction Steps.

6. EXPERIMENT RESULTS

In order to evaluate the performance of the proposed AHR system that was programmed using Matlab version R2016b, multiple experiments were conducted on IFN/ENIT database. The collected words from the dataset were 569 words. Some words were ignored for the poor of writing and the difficulties to be recognized by human readers. The system was tested by 100 words. These experiments show that the system can recognize AHR with recognition rate of 70%. When testing the system in other words, not from the database but written in a clear manner, the recognition rate becomes higher than 90%. The training results in the system, state, regression, and also the mean squared error are shown in Figures 8, 9, and 10 respectively.

This test showed that the proposed Arabic Handwriting Recognition system has confusion and rejection in some words. Because of some characters with broken loops and character segmentation problem. Peak connection fixes several segmentation troubles and assists in providing superior accuracy. Observing that, the tested word is better recognized if the OCR system is trained using well-written instances. Also, the pre-processing steps worked to raise the recognition rate as well.

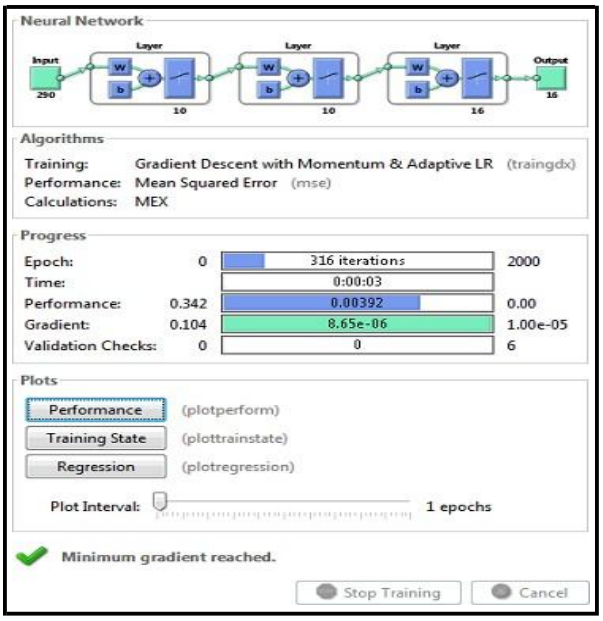

Figure 8: The Results of Training the Arabic HR.

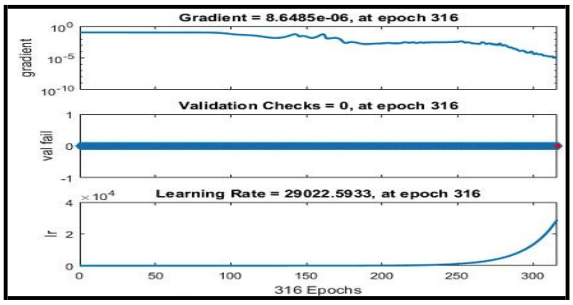

Figure 9: Arabic HR Training State.

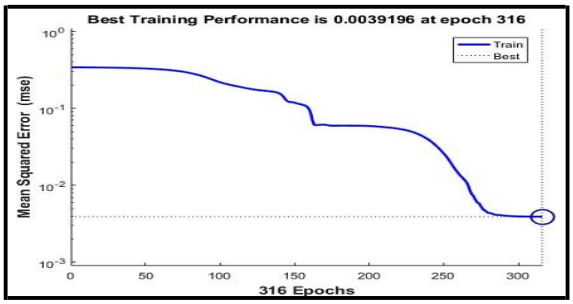

Figure 10: Arabic HR System Mean Squared Error (MSE) of the system.

Table 3 lists the final results of the Arabic HR compared to another system which has a lot of similarity to the proposed model from the aspect it was considering Arabic recognition text using NN. While it uses different techniques to extract features using a technique named moment invariant and different segmentation technique was used as discussed in related work section (Section 3). The study was offered in 2003, includes printed texts by computer and did not include those handwritten.

Table 3: System Recognition Rate Comparison.

System	Dataset	Recognition rate (%)
Offline Arabic Text Recognition system [12]	randomly printed text	73%
Proposed system	IFN/ENIT	70%
Proposed system	General printed and typed words	> 90%

As shown in Table 3 the proposed model tested on 100 words that written by hand. The system was able to recognize 70% of the words entered into the model. Also, the recognition rate reached greater than 90 %, while testing on set mixed of printed and typed words which outside the selected database. While in [12], the recognition rate was 73% of randomly printed text only.

7. CONCLUSION AND FUTURE WORK

A lot of methods and approaches can be used to recognize AHR words and texts. In this paper, we used OCR using Neural Network approach. As well, we used different preprocessing steps. Working on a preprocessing phase, it leads to inventing a novel method for baseline estimation and correction, with improving the recognition rate to more than 90%.

The NN classifier results in accurate results in Arabic HR through training the proposed model on a set of Arabic words in the IFN/ENIT dataset. The preprocessing phase is based on the following steps: deleting spaces between parts of a word, resizing an image, thinning followed by bolding the word and finally the baseline estimation and correction.

As future work, same preprocessing steps will be applied and tested using another type of classifier model such as Hidden Markov Model. Also, shortly we will compare the recognition rate while using the holistic approach which based on recognizing the entire word with an approach based on recognizing a single (character /letter) in Arabic handwriting to check the advantages and disadvantages of both approaches.

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, May 2006. https://doi.org/10.1109/TPAMI.2006.102

[2] C. J. Mathew, R. C. Shinde, and C. Y. Patil, "Segmentation techniques for handwritten script recognition system," *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, pp. 1–7, Mar. 2015. https://doi.org/10.1109/ICCPCT.2015.7159397

[3] J. H. AlKhateeb, F. Khelifi, J. Jiang, and S. S. Ipson, "A new approach for off-line handwritten arabic word recognition using KNN classifier," *2009 IEEE International Conference on Signal and Image Processing Applications*, pp. 191–194, Nov. 2009. https://doi.org/10.1109/ICSIPA.2009.5478620

[4] B. El Qacimy, A. Hammouch, and M. A. Kerroum, "A review of feature extraction techniques for handwritten Arabic text recognition," *2015 International Conference on Electrical and Information Technologies (ICEIT)*, pp. 241–245, Mar. 2015. https://doi.org/10.1109/EITech.2015.7162979

[5] R. Singh, C. S. Yadav, P. Verma, and V. Yadav, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network," *International Journal of Computer Science & Communication*, vol. 1, no. 1, pp. 91–95, 2010.

[6] J. H. AlKhateeb, F. Khelifi, J. Jiang, and S. S. Ipson, "A new approach for off-line handwritten arabic word recognition using KNN classifier," *2009 IEEE International Conference on Signal and Image Processing Applications*, pp. 191–194, Nov. 2009. https://doi.org/10.1109/ICSIPA.2009.5478620

[7] M. Sarfraz, S. Nawaz, and A. Al-Khuraidly, "Offline Arabic Text Recognition system," 2003 International Conference on Geometric Modeling and Graphics (GMAG'03), pp. 7695–1985, 2003. https://doi.org/10.1109/GMAG.2003.1219662

[8] AL-Shatnawi and K. Omar, "Methods of Arabic Language Baseline Detection – The State of Art," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 10, pp. 137–143, 2008.

[9] Khemiri, A. Kacem Echi, A. Belaid, and M. Elloumi, "Arabic handwritten words off-line recognition based on HMMs and DBNs," *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 51–55, Aug. 2015.

[10] Shamsher, J. Orakza, A. Adnan, and Z. Ahmad, "OCR For Printed Urdu Script Using Feed Forward Neural Network," pp. 1–4.

[11] K. Addakiri and M. Bahaj, "On-line handwritten arabic character recognition using artificial neural network," *International Journal of Computer Applications*, vol. 55, no. 13, pp. 42–46, Oct. 2012.

[12] J. H. AlKhateeb, J. Ren, J. Jiang, and H. Al-Muhtaseb, "Offline handwritten arabic cursive text recognition using hidden Markov models and re-ranking," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1081– 1088, Jun. 2011. https://doi.org/10.1016/j.patrec.2011.02.006

[13] L. Chergui and M. Kef, "SIFT descriptors for arabic handwriting recognition," International Journal of Computational Vision and Robotics, vol. 5, no. 4, p. 441, 2015. https://doi.org/10.1504/IJCVR.2015.072193