A SURVEY ON ARABIC HANDWRITTEN CHARACTER RECOGNITION TECHNIQUES

Ashiq V M
Research Scholar, Department of CS, CA & IT,
Karpagam Academy of Higher Education,
Coimbatore, Tamilnadu, India.
Email: vmashiq@gmail.com

E.J. Thomson Fredrik
Professor, Department of CS, CA & IT,
Karpagam Academy of Higher Education,
Coimbatore, Tamilnadu, India.
Email: thomson500@gmail.com

Abstract: *This paper presents an overview of off-line handwritten Arabic character recognition and summarizes the main technical challenges and characteristics of Arabic. This also discuss about the latest trends in Arabic handwritten character recognition and their challenges. Natural language processing with Optical character recognition technique is most commonly used in the area of recognition of characters. In Hand written text there is no restraint on the writing style. Hand written characters are challenging to recognize due to diverse human handwriting style, variation in angle, size and shape of letters. Various attitudes of hand written character recognition are discussed here along with their performance.we discuss the preprocessing , feature extraction and post processing steps in this*

Keywords: Deep Neural Network, Character recognition, Optical character recognition, Etc.

1. INTRODUCTION

Handwritten character identification is mainly classified into two.

- Online character recognition
- Offline character recognition

In online character recognition when a user writes on a digitalized tablet which is electromagnetic or pressure sensitive. so the movements of the pen are converted into electronic signals and recorded in the computer.

In offline character recognition is more challenging.

It includes automatic conversion of text into an image and then compared and form the texts which are used by the text processing applications.

Handwritten characters difficult to recognize because of its different size, different font and different styles. Here we discusss the different methodologies for handwritten character recognition.

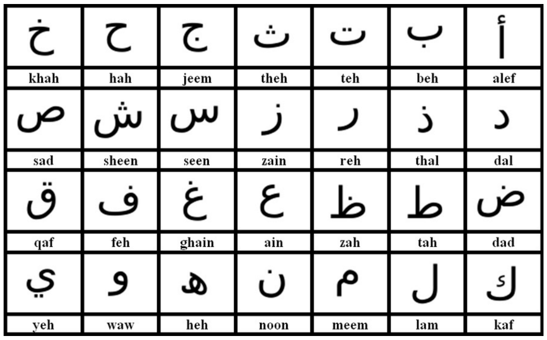
Fuzzy logic was first introduced by Lotfi Zadeh. It was developed for solving decision making problems through the use of "IF-THEN" rules. It was used later to model uncertainty and imprecision in data management. Fuzzy logic is an

easy way to reach definitive conclusions based on vague, ambiguous, imprecise and noise information .

Neural Network approach is an emerging technique in the area of handwritten character recognition through the use of Artificial Neural Network (ANN) implementations were networks employs specific learning rules to update the links (weights) among their nodes. Such networks can be fed with data from an input picture and trained to output characters in one or more forms. There are many structures of ANNs including, Perceptron, Adaline, Madaline, Kohonen, Back Propagation and many others. Back Propagation ANN is the most commonly used since it is effective and very simple to implement.

2. ARABIC LANGUAGE CHARACTERISTICS

Arabic is a native language for more than 250 million people. It is the third largest international language used by over one billion Muslims in their different religious activities. In addition to the Arabic language, there are several languages that use the Arabic alphabet, such as Urdu, Farsi (Persian), Pashto, Jawi, and Kurdish. Although a conclusion may review the main points of the Arabic text is written from right to left and is always cursive in both machine printed and handwritten text . The Arabic alphabet set is composed of 28 basic letters which consist of strokes and dots. Dots, above and below the characters, play a major role in distinguishing some characters that differ only by the number or location of dots e.g. Ba (ب), Ta (ت), and Noon (ن). The shape of an Arabic letter changes according to its location in the word, as shown For each character, there can be two to four different shapes: isolated, connected from the left (beginning of a word), connected from the left and right (middle of a word), and connected from the right (end of a word). Out of the 28 basic Arabic letters, six can be connected from the right side only while the other 22 can be connected from both sides. These six characters are: Alef (ا), Dal (د), Thal (ذ), Ra (ر), Zy (ز), and Waw (و) .These six characters have only two shapes, the isolated shape and the end shape, whereas the rest of the alphabets can appears in any of the four shapes mentioned above [13]. Consequently, each word may form one or more sub-words, where a sub-word is one or several connected characters,



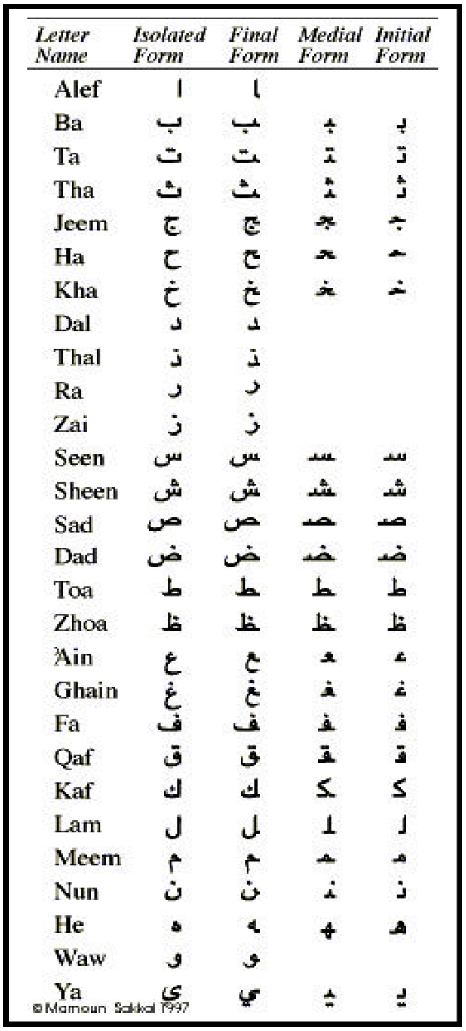
خ	ح	ج	ث	ت	ب	أ
khah	hah	jeem	theh	teh	beh	alef
ص	ش	س	ز	ر	ذ	د
sad	sheen	seen	zain	reh	thal	dal
ق	ف	غ	ع	ظ	ط	ض
qaf	feh	ghain	ain	zah	tah	dad
ي	و	ه	ن	م	ل	ك
yeh	waw	heh	noon	meem	lam	kaf

Fig: Arabic characters.

for example
مس, نـور تــش فــی , and مناس بــات . Moreover, in certain fonts, several characters can be vertically combined to form a ligature, especially in typeset and handwritten text. Ligatures can be formed out of two, three, or four characters. Characters in a word may also vertically overlap without touching.

The use of special stress marks called diacritics is another distinguishing characteristic of Arabic. Diacritics such as Fat-ha (◌َ), Dhammah (◌ُ), Shaddah (◌ّ), Maddah (~), Sukun (◌ْ), and Kasrah (◌ِ) may change the pronunciation and the meaning of the word. The diacritics significantly affect the OCR performance. There are nine Arabic letters; Sad (ص), Dhad (ض), Tah ط)), Dha (ظ), Fa (ف), Qaf (ق), Meem (م), Ha (ه), and Waw و)) that have closed loops. This makes the closed loop an important feature in recognizing Arabic characters. One of the important characteristics of Arabic text is the presence of a baseline which is an imaginary horizontal line running through the connected portions of the text. If the script is handwritten, the baseline is not straight, and may only be estimated. Another feature of Arabic characters is that they do not have a fixed width or size, even in printed from. The character size varies according to its shape which is, in turn, a function of its position in the word. In addition to the 28 characters, Arabic has additional non basic characters such as Hamzah (ء) and Ta marboota ة)). Hamzah can be isolated, on Alef (أ), on Waw (ؤ), or on Ya (ئ). Ta marboota is a special form of the letter Ta(ت) that only appears at the end of words.

Arabic characters depending on the shapes are shown in the table below



Letter Name	*Isolated Form*	*Final Form*	*Medial Form*	*Initial Form*
Alef	ا	ـا		
Ba	ب	ـب	ـبـ	بـ
Ta	ت	ـت	ـتـ	تـ
Tha	ث	ـث	ـثـ	ثـ
Jeem	ج	ـج	ـجـ	جـ
Ha	ح	ـح	ـحـ	حـ
Kha	خ	ـخ	ـخـ	خـ
Dal	د	ـد		
Thal	ذ	ـذ		
Ra	ر	ـر		
Zai	ز	ـز		
Seen	س	ـس	ـسـ	سـ
Sheen	ش	ـش	ـشـ	شـ
Sad	ص	ـص	ـصـ	صـ
Dad	ض	ـض	ـضـ	ضـ
Toa	ط	ـط	ـطـ	طـ
Zhoa	ظ	ـظ	ـظـ	ظـ
'Ain	ع	ـع	ـعـ	عـ
Ghain	غ	ـغ	ـغـ	غـ
Fa	ف	ـف	ـفـ	فـ
Qaf	ق	ـق	ـقـ	قـ
Kaf	ك	ـك	ـكـ	كـ
Lam	ل	ـل	ـلـ	لـ
Meem	م	ـم	ـمـ	مـ
Nun	ن	ـن	ـنـ	نـ
He	ه	ـه	ـهـ	هـ
Waw	و	ـو		
Ya	ي	ـي	ـيـ	يـ

©Mamoun Sakkal 1997

3. DEEP CONVOLUTIONAL NEURAL NETWORK (DCNN)

In 2017 December Khaled s younis peoposed an deep convolutional neural network for recognition of Arabic handwritten characters.

This is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics.

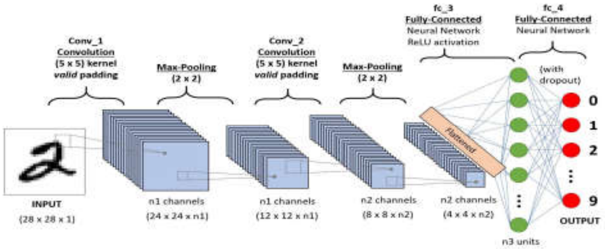


Fig: Convolutional network model

He used two datasets for the testing of accuracy of the characters.

AHCD(Arabic handwritten character data set)

The dataset is composed of 16,800 characters written by 60 participants; the age range is from 19 to 40 years and 90% of participants are right-handed. Each participant wrote each character (from "Alef" to "Yeh") ten times. The forms were scanned at a resolution of 300 dpi. The database is partitioned into two sets: a training set (13,440 characters to 480 images per class) and a test set (3,360 characters to 120 images per class)

AIA9K (AlexU isolated alphabet dataset)

This dataset introduces a compact 9K novel dataset of 28 classes that represent isolated Arabic handwritten alphabet of 32x32 pixels . AIA9K dataset was collected from 107 volunteer writers, between 18 and 25 years old, who are B.Sc. or M.Sc. students. The writers were 62 females and 45 males. Each writer wrote all of the Arabic letters 3 times. The total valid number of collected characters is 8,737 letters; this novel dataset can be requested from the authors of the paper mentioned in. A sample of the dataset is shown in Figure 3. These are 75 characters that were misclassified in one of the experiments

4. LINEAR CORRELATION

Abd Elhafeez ,Hamid Mariam M. Musa ,Mona M. Osman ,Moahib M. Alamien ,Marwa M. Elsaied proposed a linear co efficient method to recognize the character.

In this method there are five major stages.

- Image acquisition
- Pre processing
- Segmentation
- Feature extraction
- Classification

Image acquisition

In this step just convert the scanned or stored image into binary image.

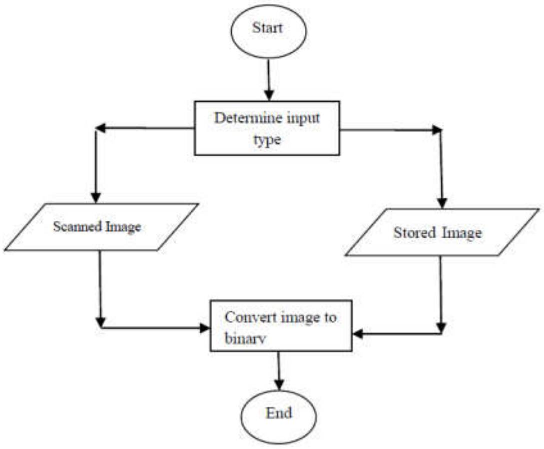


Fig: Image acquisition algorithm

Pre processing

In this process remove all the unwanted elements in the image. Binarization , cleaning and smoothing are included in the preprocessing.
In binarzation it takes only the information requires for the task is general shape , or outline information.
In smoothing it fills the superfluous point of the contour image.
In cleaning it remove the noise that could not be eliminated by cleaning.
Segmentation is done by three steps.
First the document is divided into lines by using
histogram, then calculating the number of dots in each horizontal pointed row.

Dividing the lines of the document into Characters, according to the shape of the Character, based on rules and information that owned by the system

Extracting the features by collecting the dots in each row separately, and also to the columns, then studying and analyzing the characteristics such as Character height and width, in preparation to identify the Crafts

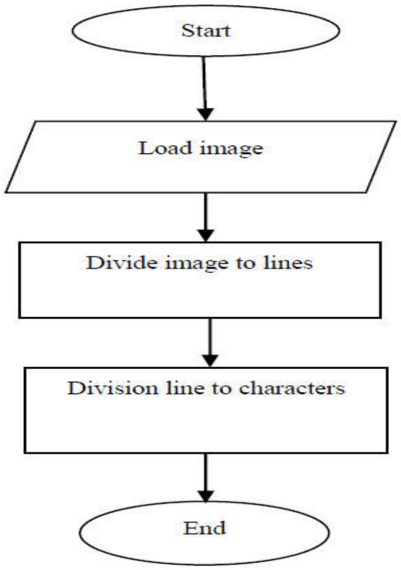


Fig: Segmentation algorithm

Feature extraction

This process extracts the features of the characters that are most relevant for classifying at recognition stage. This is an important stage as it can help avoid misclassification, thus increasing recognition rate. In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements.

Classification

There are variety of techniques used to classify the charecters.linear correlation algorithm is used here to classify the characters.

Equation are shown below.

$$x = \sum_m \sum_n \left(A_{mn} - \overline{A}\right)\left(B_{mn} - \overline{B}\right)$$

$$y = \sqrt{\left(\sum_m \sum_n \left(A_{mn} - \bar{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \bar{B}\right)^2\right)}$$

$$r = \frac{x}{y}$$

Where:
r = correlation value.
A = initial matrix (for character want to identify it).
B = template matrix (for stored character).

$\overline{A}$ = mean of the initial matrix (character want to identify it).

$\overline{B}$ = mean of the template matrix (for stored character).

Hence we find the peak signal to noise ration(PSNR) , Correlation factor.

5. DEEP BELIEF NEURAL NETWORKS

Mohamed Elleuch, Najiba Tagougui propose a technique called deep belief neural networks for recognizing the Arabic handwritten character recognition.
Deep Belief Networks are a graphical representation which are essentially generative in nature i.e. it produces all possible values which can be generated for the case at hand. It is an amalgamation of probability and statistics with machine learning and neural networks. Deep Belief Networks consist of multiple layers with values, wherein there is a relation between the layers but not the values. The main aim is to help the system classify the data into different categories.

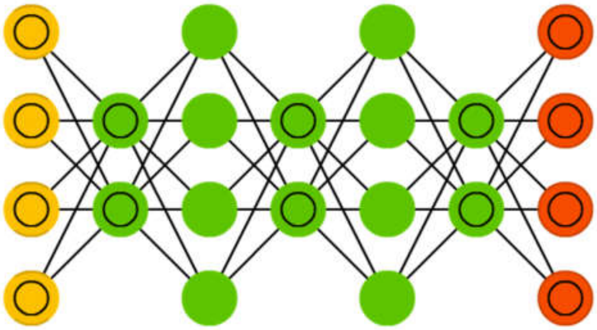


Fig:Deep belief neural networks

In this paper feature learning based approach by the DBN networks for handwritten Arabic word/characters recognition problem. This family of networks manages large dimensions input, which allows the use of raw data inputs rather than to extract a feature vector and learn complex decision border between classes. For the character level one, the results were promising with an error classification rate of 2.1% for the HACDB database. Unlike, the word level one with the ADAB database where the error rate exceeded the 40% and had to be improved.

Wavelet energy and extreme learning machine

Binu P. Chacko , V. R. Vimal Krishnan proposed a technique using wavelet energy and extreme learning technique for character recognition. Since WEF is robust to some extent of rotation and translation of the images so there is no need to implement skew correction methods. This is mainly used for multi-resolution images.

6. CONCLUSION

Now a days there are so many application in the field of character recognition and natural language processing. As a part of machine learning so many techniques are applied for recognition process.In this paper we discussed the difficulties, challenges in Arabic handwritten character recognition.we also discussed the techniques like wavelet transform, artificial neural network, Convolutional neural network,and other preprocessing, feature extraction, post processing techniques.Now the research mainly focus on the advanced deep learning techniques.

REFERENCES

[1] Amara N, Bouslama F (2003) Classification of Arabic script using multiple sources of information: State of the art and perspectives.

[2] International Journal on Document Analysis and Recognition, 5(4): 195-212.

[3]Lorigo LM, Govindaraju, V (2006) Off-line Arabic Handwritten Recognition:

[4] A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. EEE Trans. Pattern Analysis and Machine Intelligence 28(5): 712–724.

[5] G. Dimauro, S. Impedovo, G. Pirlo, A. Salzo (1997). Automatic bankcheck processing

[6] A new engineered system. In S.Impedovo et al, editor, International Journal of Pattern Recognition and Artificial Intelligence, World Scientific,.

[7] S. L. Xie, M. Suk (1988). On machine recognition of hand-printed chinese character by feature relaxation. Pattern Recognition, 21(1)

[8] D. Guillevic, C. Y. Suen (1995). Cursive script recognition applied to the processing of bank cheques. In Proc. of 3th International Conference on Document Analysis and Recognition, Montreal-Canada, August, pp

[9] L. Mico, J. Oncina (1999). Comparison of fast nearest neighbour classifier for handwritten character recogniton. Pattern Recognition Letters, 19(3-4):351-356.

[10] R. O. Duda, P. E. Hart, D. G. Stork (2001). Pattern Classification. John Wiley and Sons,