ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1101022021.pdf
https://doi.org/10.30534/ijatcse/2021/1101022021

Recognition of Handwritten Sheba Character using Artificial Neural Network

Ibrahim M. G. Alwayle
Dept. of Comp. Science, Faculty of Science and Arts, Sharurah, Najran University, Saudi Arabia.
Dept. of Math. Science, Faculty of Science and Arts, Sada'h University, Yemen.
alwayle1@yahoo.com

ABSTRACT

From the Last twenty years, the computer-based mechanism has an essential process in daily life and research-oriented applications, the whole world be attracted by computers and approximately all the main processing is being completed automatically. Recognition of handwriting now it is an eye-catching and tough study analysis in image processing and pattern identification field in the today's world. To beat this problem Optical Character Recognition system (OCR) is practice and concentrated research has been carrying on OCR. Numerous OCR systems are existing in the market other than mainly of this system working for Japanese, English, Chinese, Roman letters. There is no adequate work on Arabic language particularly Sheba characters. In proposed paper, an OCR system for Sheba recognition of character depend on neural network is presented. After analysing various method for segmentation and pre-processing A and B which are used for image pre-processing and segmentation respectively to enhance performance for projected framework. All the issues and challenges for OCR system and relevancy or accuracy of proposed method are discussed and analysed briefly.

Key words : Sheba Character; Image pre and post-processing; Segmentation; Feature Extraction; Pattern analysis; Neural Network.

1. INTRODUCTION

Over the past decade, the computer-based work and analysis became the primary selection of everyone. It is an extremely computational power and development in image analysis algorithms has authorized the improvement of powerful computer based logical approach to normal languages .

Since the sixties character recognition [11] is a research difficulty which is in progress. At present it is a lively field of research due to the difficulty and its nature is complex. It is the necessary module for analysis document system. OCR helps in image natural language processing and pattern recognition. It is the recognition of written by hand or printed wording with the help of PC. Optical character recognition is calculated as the mainly doing well merits in the area of pattern recognition and AI. Numerous marketing organization are existing for carrying out OCR that presents for number of usages, while the apparatus also till now not capable to participate among abilities of human analysis. Forms contain characters images and scanner is scanned images and after that recognition engine of the OCR system understands the images and converted it of handwritten or print letters into ASCII data (MCR). This recognition gives a major advantage in terms to fill the space among communication of machine and man communication. The study of document and recognition has playing a most important responsibility in research of recognition pattern. In common, the study is on progress on OCR for scripts of Indian. However, there's no resolution has been presented that fix the issues perfectly and capably. The procedure of recognition of letters may be separated into further two parts, handwritten and printed character recognition. The printed documents also further divided in given printed documents is in better quality and printed documents that are spoiled.

Transformation of scanned or printed text images is called Optical character Recognition (OCR), handwritten text into text that is editable for additional processing. Those skills grants apparatus to text is identify by itself. It's such that grouping of eye and brain of person. An eye could easily inspect the letter from the pictures however in fact the processes of brain and understand too that text that extracted study by human vision. In progress of system of computer of OCR, some issues may take place. Initial: there is extremely small visible dissimilarity between a number of digits and letters for computers to recognize. For example, it may be complicated to PC to

distinguish among letter "o" and digit "0" and. Next: this may be extremely complicated to take out text that is entrenched in especially dark environment or written on different characters or graphics. In 1955, the earliest marketing method was set up at the reader's digest, that OCR is beneficial to enter marketing data into a PC and later than that OCR system becomes useful in computerizing the physical office papers. OCR have numerous applications, includes: License plate recognition [1],[23]and [25]. Image text taking out from original view images [10], text is extracted from file or paper that is scanned [17] etc. The approach discussed in [2] is to amend the text reacquire from camera clicked images. Thomas Desolaters et al. (2012) offered an OCR method that is used for identified letters written by hand and digital text changed by characters.

Near millions of speakers of language of Arabic arrive from the Algeria, Egypt, Comoros Djibouti, Sudan , Chad ,Bahrain, Qatar, Oman, Ethiopia, Sudan, Morocco ,Mauritania, Somalia ,Saudi Arabia, Syria , Tunisia, Turkey , Western Sahara , the West Bank, the United Arab Emirates, the republic of Yemen Arab. Historically, Arabic have holding a vital position in the East of Middle since it is conventional teachings language (i.e., the other literature).

To start with, ease of understanding of material of Arabic (magazine, books, documents, journals etc.) in digital category, particularly electronically, significantly enhances the observability of in and out-of-print content that does not exist apart from in few shelves of library. Secondary, the skill to across the world access ancient and present objects (in Arabic) may be one of the sources for bridges making among the west and world of Arabic-speaking. A digitization attempt purposeful on text of Arabic can straight advantage from fast and correct OCR methods.

Sheba alphabet is the alphabets of south Arabian. The earlier familiar dedication under date of alphabet from regarding 500 BC. It's not known origin, although it developed the single theory from the alphabet of Byblos. The alphabet of Sabaean is consideration to have developed into the script of Ethiopic and the Arabic text is in cursive nature is the core barrier for some OCR system of Arabic. Mainly, lettering is linked to every other in is called "base line" to create sub words. A few letterings may join to everyone over the base line. Researchers encompass rewarded particular interest to this difficulty and established a lot of algorithms to segment Arabic text into characters [5],[6].

In order to described above OCR analysis process for Sheba character and proposed solution in an efficient way, this paper is organized as follows. Language character of this Arabic material with their alphabets and representation is discussed in section 2. Section 3 in brief illustrates historical view of OCR System and section 4 provides the processing structure of OCR system with various approaches and method used at different stages of OCR. And after discussing several methodologies in section 5, proposed methodology for recognition of Sheba character is presented in section 6 and evaluation matrices and results of experiments are analysed in sec-6 among on the entire conclusion of this proposed research summarized in sec-7.

LANGUAGE BACKGROUND

Sabaean, could be called as Sabaic or Shabaic, it was an Old language of South Arabian that was used in Yemen between c. (1000-600 BC). By the Sabaeans it is 1000 BC and the 6th century AD. That is adapted as language written by several many persons of earliest Yemen, with the, Ghaymānites, Ṣirwāḥites, Humlanites and Radmānites, Ḥashidites, Ḥimyarites [15],[9]. The language of Sabaean appertain to the South Arabian subdivision of the Semitic group of the Afroasiatic language family [7]. It is illustrious via the various persons of the ancient South Arabian cluster with usage of h to spot the third personality, and as a prefix of causative; the else languages whole usage s1 in these cases; therefore, Sabaean is known as h-language & the else s-languages [16]-[18].

Sabaean was done in the alphabet of South Arabian, and such that Arabic and Hebrew marked merely consonants, the simply warning of vowels being among matres lectionis. some decades the merely wording exposed were writing in the official Masnad script (Sabaean ms3nd), but in 1973 papers in an additional minuscule and cursive script be found, dating reverse to the moment shared of the BC of 1st century; some letter have so far been in print [8].

Sheba language is the oldest language in Middle East and found in many of countries as Yemen, Saudis Arabia and some of places in Africa (Ethiopian and Djibouti). It has 29 characters and written as of right to left as shown in fig. The words in Sheba language written as separated characters and the words separated by special character see fig (2).

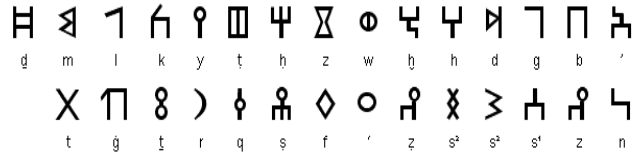

Figure 1. Sheba Character

Figure 2. Sheba word

Figure 3. Sheba Text

In the starting of 8th century the South Arabic alphabet used in Yemen, Ethiopia, Eritrea and Djibouti in all three places, afterwards developed into the Ge'ez character. The language of Ge'ez though is no far deliberate to be a successor of Sabaean, Ancient South Arabian [13],[14]; and there is linguistic proof which languages of Semitic be in utilize and used Eritrea and Ethiopia as for spoken as early as 2000 BC [12].

Sabaean is prove in a few 1040 dedicatory writing, 850 building writing, 200 authorized wording and 1300 short graffiti (contain individual names) [19]. No literary texts of any length have yet been fetch to light. This shortage of resources and the restricted forms of the inscriptions has made it complicated to acquire an entire image of Sabaean grammar. Hundreds of written in a cursive script (known as Zabur) cut into sticks of wooden which has discovered and period of Middle Sabaean; these symbolize inscription & authorized papers and like consist of a lot broader range of forms of grammatical.

HISTORICAL VIEW OF OCR

The source of recognition of character was discovered in 1870 while Carey made-up the scanner of retina—system of transmission of an image (J. Mantas, 1986). Afterwards, as on year 1890, Nipkow made-up in order scan device, that is a main crack via mutually for current world TV and reading machines. Though, character recognition was firstly well thought-out as a help to the visually handicapped and the before time on victorious trial were prepared by the Tyurin the Russian scientist in1900.

A short explanation of the past of OCR as following. In Germany Gustav Tauscher acquire a on OCR patent in year 1929, pursued by Handel that acquire a USA patent on OCR in year 1933. A US patent was also approved by Tauschek on his system in year in 1935. Mechanical machine named Tauschek's machine which is used a photo detector and templates. In 1949 RCA engineers as working over initial primitive computer-type OCR to assist sightless persons for the Administration of US Veterans, but alternatively of changing the in-print letters to words of machine, its machine changes to language of machine and after that letters be spoke. It proves far too costly and was not follows later than testing [3],[4].

Kurzweil Products in year 1978 of Computer start trade a marketing edition of the OCR set of instructions of computer. The first customer was LexisNexis and purchased the set of instructions to upload legal paper & documents onto its nascent databases which is online. RWA and Reader's Digest in 1965 work together to make an OCR Document reader draft to convert into digital the successive records on Reader's Digest coupons come back from commercial ads. After few years, Xerox purchased Kurzweil's company that had significance in additional advertising paper-to-computer conversion of text. Kurzweil Computer goods develop into a supplementary of Xerox recognized as Scan soft, now Nuance Communications.

Due to flexibility, toughness and effectiveness, the marketing OCR order may be separated into four generation. The initial generation method may be categorized by the inhibited letter form that the OCRs understand like equipment came into view in the earliest of 1960s. The initial broadly marketized OCR of generation was the IBM 1418 that was intended to read a particular IBM font [5] ,[6]. The acknowledgment way was logical pattern matching where the positional connection was completely utilized. The after that generation is considered by the acknowledgment ability of a set of usual devices printed lettering and characters hand-printed too. In previous years, the capacity was limited to number merely. Like equipment come into view mid of 1960s. Now generation, IBM1287 was the first and well-known OCR system, which was display at the 1965 New York world fair [20]. Considering as configuration of hardware, the method was a hybrid one, combining digital and analog technique. The very basic automatic letter-sorting machine for postal code numbers of Toshiba was grown too throughout present era. The procedure was depending on the structural study way. The third generation may be distinguishing by the OCR of

destitute print quality alphabet, and hand printed alphabets for a huge group character set. Commercial OCR order with such capacity came into view approximately throughout the decade 1975 to 1985 [20]-[22]. The 4th generation may be described by the OCR of difficult intermixing of documents along graphics, wording, table and symbol of mathematics, no restricted handwritten data, colour document, less-quality noisy documents such as fax and photocopy, etc. Several sections of task on difficult documents given better outcome. Even though numerous sections of task on unrestricted character that is handwritten are existed in the literature, the identification efficiency almost not goes above 85%. Extremely small learning on published of colour documents and problem exploration is progressing. Also, study on noisy document is in development [12]. Between commercial goods, in the market postal address readers are existing. In the US, regarding 70% of the sorted printed by hand without human intervention [6]. Reading assist for the sightless is present too. An included OCR with speech result method for the sightless has been advertising for English language by Xerox–Kurzweil [16].

At current, additional complicated optical readers are existed for Arabic, Japanese, Chinese and Roman text [18]. These people who read can procedure documents that have been type typeset, written, or printed different kinds of printers. As may identify letters among dissimilar fonts and sizes and unlike formats as well as graphics and intermixed text. Among the overview of narrow ranges canners, measuring 3 to 6 in broad, too scanning columnar is currently feasible. Through scanners an optical reader may identify columns number or page division or lists of mail. Few are capable of along checking of spell software, and words or letters [9],[17]

TECHNICAL PROCESS OF OCR

In common, the procedure of OCR may be separated into following steps:

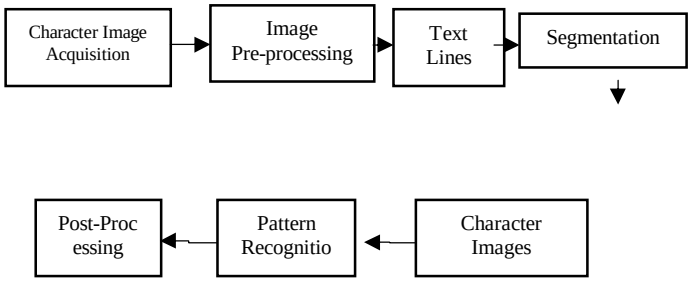

Figure 4. Flow Diagram of OCR

As showing in Fig. 4, OCR is a bit by bit procedure containing of acquisition of image, improvement, segmentation of image, drawing out of characteristic, pattern recognition & post-processing. Dissimilar Fig. 1, a few automatic image analysis methods can select to omit few steps, while few might contain its application particular steps. All steps specified before is a difficult mission of one step due to success or failure significantly influences the achievement or breakdown of the coming step.

Image Pre-processing& Enhancement

The obtained images via samples might be of low quality because of the dissimilarity in staining and lighting circumstances and might be incorrect oriented. In an image, inappropriate staining and lighting circumstances go ahead to messy and unnecessary object. The quality of image is also ruined with unnecessary indication that known as noise. So that is predictable that the images that are obtained should be of high-quality to create the proposed result. Steps and Means that are usually used to improve an image quality in OCR may be cluster like explained next segment.

Skew Correction

Image acquired from the earlier stage might not be accurately oriented; It might be associated at any angle. So we necessitate carrying out skew modification to create certain that the image forwarded to following stages is accurately oriented (A.Amin, 1998).

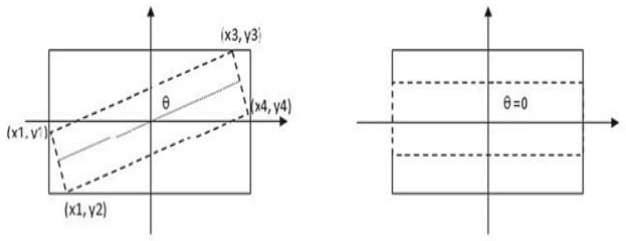

Figure 4. a. Skewed Rectangle b. Rectangle with zero Skew

Starting this point move ahead now utilized the text edge pixels as a document representative. So that, pixel on four corners of a document are used, that shows a min. and max. Column, min. & max. row. Those pixels of parallel rectangular corner are used to compute the region of axes with use the area of rectangle as in mathematical equation,

$$Area\ of\ Rectangle = Height * Width \quad \text{Eq.1}$$

Now height is evaluated by accomplishment the dissimilarity of min. and max. Column & unchanged for width, that is evaluated by obtaining a dissimilarity of max. & min. row on which way the rotated rectangle is considered by circling edge pixels at ±1. If an region of a rectangle by increasing pixels on edge by 1^0 of an area of rectangle is fewer than the space of the original rectangle, so that the revolution way is +ve and if an space of a rectangle by pixels' edge decrementing by 1^0 and less than the area of the original rectangle then way of rotation is -ve. Firstly, skew degree is fixed to zero while original area of rectangle is evaluated. Then space of rectangle is checked by turning the image in a hoop, although rotating rectangle skew angle is increasing or decreasing of rotation direction consequently. The approach of Bisection is used to find the angle that provides the minimum area. To confirm the evaluate θ way the correct skew angle; it is distinguished with the original rotated angle.

Binarization

Its way changing a Binary image by Coloured image (comprising with black and white). Generally, in practice, this change of Coloured image to Binary image is completed by generating a midway greyscale image. It could be accomplished using distinct methods as

1. **Otsu's Method**

This technique is an international thresholding technique so as to changing gray scale into bi-level image. This method split the pixels following categories: foreground and background. It selects an optimal threshold which categories the image into two dissimilar classes. Values of threshold are selected like inside class discrepancy is decreased & maximized the among class variance . The weighted inside variance of classes is specified as given in below equation

$$\sigma^2 p(t) = p_1(t)\sigma_1^2 + p_1(t)\sigma_1^2(t)$$

Eq.2

Otsu method provides its most excellent presentation for merely clear bi-modal pattern that contained by images. Although, documents that ruined usually can't contain clear-cut pattern. Further this and can't carry out fine for images among rough shadow.

2. **Niblack's Method**

It is a local thresholding means. In these techniques, a dissimilar of all pixels is evaluated by threshold value. It utilizes local statistics of the image, like variance, range to compute the threshold. In Niblack technique a window of rectangle is slided above the image of gray scale to estimation pixels threshold. It uses the local statistics indicate and window's standard deviation to estimation the threshold. T (i, j) is calculated as described in Eq. 3

$$T(i,j) = \mu + k * \sigma$$ Eq.3

In Eq. μ shows the indicate of the window and σ signifies the window standard deviation. k value's is not a variable and it describes the quality & size of binarization. So, this technique is reliant upon image the local characteristics, it obtained pretentious by the image that contain black region, for images which is inefficient too with noise in the background.

3. **Sauvola Method**

This technique is the enhancement in previous technique. Standard deviation uses by Its local variance technique. Threshold is evaluated as given Eq. 4

$$T(i,j) = \mu * k\left[1 + k\frac{\sigma}{R} - 1\right]$$

Eq.4

In Eq. 4 μ = mean and σ = window's standard deviation. k and R are 0.5 and 128 suggested values. The image quality if influenced by the size of window and value of k but R not affected too much. The section for those documents that contain texture is light, documents that are stained and uneven illumination. Although, this technique later than its application thins the text.

4. **Bernsen**

This technique employs the contrast of image. The threshold is approximate maximum and minimum intensity values in the window that the average. The Eq. 5 evaluates the window's local contrast.

$$C(i,j) = I_{max} - I_{min}$$ Eq.5

With the comparison of threshold value by local contrast the pixels are divided as background or foreground, if the local contrast is originated to be lower than the threshold, so on is considered pixel would order as background. Complex background contained by images in that Bernsen technique doesn't perform well.

5. Local Maxima and Minima

It utilizes contrast which depends upon the local min and max. A factor of normalization is established that will compensate the influences of background image variation. Image contrast is evaluated as given in Eq. 6

$$C(i,j) = \frac{I_{max} - I_{min}}{I_{max} - I_{min} + \varepsilon}$$ Eq.6

The contrast image helps out to found High contrast image pixels. Then local thresholding is performed with threshold value evaluated from the found high contrast

image pixels. This technique is not appropriate for the bright background among bright text.

6. Adaptive Contrast

An image gradient and consolidation contrast. This technique discards the over-normalization issue of Local maxima minima technique. Contrast map is binarized and build up. Canny edge map builds up too and united among the first resultant of step. This discovers the actual text stroke edges. By using local thresholding, the text is drawing out. It is approximate calculation via standard & mean difference of the discovered text stroke pixels inside a window. The disadvantage about specific technique is the canny edge detector take out false boundaries too.

7. Global-to-Local Approach

This following by local thresholding is accomplished to image binarization. Pre-processing is used by Gaussian filter. Using canny edge detector, the Edge map is constructed.

Noise Removal

Noise (small dots or foreground components) might initiate simply over image though scanning it throughout Acquisition of Image because of less clarity camera, Shadow on image etc. This noise must be removed so that the image will be clean and uniform.

Thinning and Skeletonization

Different images have words in it with different width of strokes. This variability is very high in the case of handwritten words. So, by using Skeletonization techniques, we can make all strokes to have uniform width (Maybe 1 pixel wide or few pixels wide)

Figure 5. (a) Original Image. (b) Converted to Greyscale. (c) Binarized image. (d) Thinning Skeletonization (e) Noise Removed (A.Amin, 1998)

Segmentation

It is the mainly decisive action for application of any image processing. This procedure of recognize region of image that the application being created. Segmentation manages by means of describing these ROIs via nearby region of text. Even however all the before-explained ROIs is being segmented in OCR, a mainstream of the works have determined over means of labelling and nuclei of cell segmentation. In this part demonstrates the non-similar techniques utilized to segment the ROIs from images of medical. Segmentation of image is accomplished in the below series.

Line level Segmentation

The most important purpose of LLS (A.A.Shide et al., 2013) is to conclude the lines co-ordinates in image that could separate the lines from image. To separate lines of text, the HPP of the text document image is discovered. The horizontal projection profile (HPP) is a Histogram of a number of ON pixels with each row of image. When we plotted the projection profiles, in the plot we may observe valleys and peaks. White space among the text lines is utilized to segment the text lines. Fig 5 illustrates resultant de-skewed text document along with its horizontal projection. The profile of projection has value with height of zero among the lines of lines. At this point segment of line is ended. Where numerous steps in the line segmentation technique those are logically explaining here.

Step-1: Run length smearing

Step-2: Recursive procedure to get middle lines for segmentation.

Step-3: Finding candidate line

Step-4: Resolving the problems of overlapping and touching component

Word Level Segmentation

The gapping among the words is utilized that is for segmentation of word. In English script usually, gapping among the words is larger than the gapping among the letters that contain in a word. The gapping among the words is discovering by holding the VPP of a line of test that is input. VPP is the addition of ON pixels with all image columns. A sample input text line and its VPP is demonstrated. From the Profile it is obvious that the width of the zero-valued valleys is further among the words in the line as analysed breadth of zero-valued valleys that occur among word from characters. This data is utilized to count & separate words from the line input text.

Character Level Segmentation

That dependent on the context where OCR will be utilized.

- If system of OCR is being apply on text that is printed, in that letters surrounded by word are not connected among the others, Characters get segmented too in the earlier action itself due to text that is printed keep a significant (yet small) consistent gapping among characters. There is no requirement of accomplishing Character level Segmenting (CLS).
- As system of OCR is apply on text which is written by hand (Cursive Handwriting), characters within words are connected, as in this step as execute CLS to accurately segment characters.

So, currently the task of CLS is to split out single characters (other symbols, digits and alphabets) from the words that are split out as of earlier case.

Methods that segment a characteristic demonstration of the image (G.Richard et al.)

- Hidden Markov Model
- Non- Markov Model
- Mixed Approach

Feature Extraction

Usually, the characteristics that have been extracted from the OCR images consist of Gary level, binary and vector images. The below sections describe the different characteristics and characteristic extraction process that have been employed for study OCR Image. General characteristic extraction technique for OCR Image

- Zoning
- Projection Histogram
- Distance Profile
- Background Directional Distribution

Pattern Recognition

In an object set, pattern classified as it is a physical object or a shared property or may be an abstract notion or a of a objects set. According to logic that may be represented by an n dimensional feature vector signified by $\mathbf{x} = (x1, x2,...,xn)^T$where each xi represented an individual characteristic. Recognition defines as having seen or perceived. It is the procedure of classifying and passing on a label to an unidentified object. categorization & clustering are two pattern recognition tasks which are utilized in the recognition of medical analysis of image.

Classifiers may also be generally categorized into parametric & non-parametric methods. In Parametric scheme rely on the supposition that the functional form of class-conditional distributions of the provided characteristics is well-known while nonparametric scheme builds minimal assumptions regarding the form of the distributions. Classifiers of Parametric who analyse the parameters like variance and mean by assuming a data distribution for that provided.

The classifiers are below which are utilized for categorization of OCR.

- Bayesian Classifier
- Support Vector Machine (SVM)
- Linear Discriminant Analysis (LDA)
- K- nearest neighbour classifier
- Multiple Instance Learning (MIL)
- Neural Network Classifiers

Post-Processing

The text achieved by systems of OCR often undergo from less correctness due to images irregularity, low scans or merely the nature of letters arranged in the form of word. For example, reading “lwo” instead of “two”, “ia” in-stead of “is”, “m” instead of “rn”, to name a some. These mistaken characters acutely readability and hamper the quality of a changed document. Rectifying and classifying such mistaken characters in each OCR-processed document physically is a boring task because of the data sheer volume. Accordingly, a method is necessary to identify this kind of OCR faults and rectify them in order to enforce standards of purity and archived data quality [18].

RELATED WORK

While approximately a numerous millions of people globally [3], in numerous unlike languages, for writing use Arabic letters (alongside Urdu and Arabic that are mainly famous), Arabic character recognition not explored as systematically as Chinese, Latin or Japanese. The work that is firstly published on Arabic OTR could trace back to 1975 and was by Nazif and author bulid a system which recognise the printed Arabic characters in his master’s thesis depending on extracting strokes and defined as radicals [8] and used radicals and their positions. He used correlation among the image of character and the templates of the radicals. A segmentation stage was included to segment the cursive text. After later so many Years, Shimura, Badi [11], [16] researched on printed Arabic characters and Amin [20] on written by hand Arabic characters. In the 1980sresearchwork on AOTR improved noticeably, a trend i.e. progressing year 1990s.The merely analysis on AOTR, of which the authors are conscious, [20]-[21]

Those surveys, though, have a restricted figure of references on AOTR.

In [14], author proposed a hybrid Arabic character recognition system depend on Moment Invariants employing classifier of an ANN. The phase of characteristic drawing out utilizes a set of moment invariants descriptors those are invariants under transformation. The genuine categorization is through using a network of multilayer perception network among learning of backpropagation. As a pre-processing step, a three-level line, word, and sub-word is used to segmentation of Arabic words. A. Elgammal et al. (2001) proposed a structural segmentation of graph-based approach depend on the relation of topological among the baseline and the line adjacency graph (LAG) expression of the text. We call "scripts "when the text is segmented to sub-character units. An approach of structure analysis is utilized for recognition of these units. A distinct classifier is utilized to recognize diacritic and dots signs. The last recognition of character is accomplished with the help regular grammar that tells the idea letters are composed from scripts.

M. Sarfraz [3] proposed a non-manual recognition of text that is printed on Arabic utilizing ANN. The important characteristics of the system are text's pre-processing, text segmentation to individual characters, characteristic drawing out utilizing moment invariant method and recognition using RBF Network. In [2], author proposed A new system to recognise written by hand Arabic characters depend on neural networks. Haar wavelet features were used to train the neural networks and variation in handwriting approach has presented a actual challenge to the system. Furthermore, few of the samples of character were badly written to the extent that and may not predictable by the human eye. The samples of character in the corpus have been categories into two classes according to the level of their readability. While the system shows a satisfactory recognition rate for samples belonging to the first class, it badly recognised samples from the second class.

in [2] author presented Template Matching System with Interpolation (TMI) technique to recognize the oldest language in Southern of Arabian Peninsula which named Sabic characters (Almusnad).

PROPOSED METHODOLOGY

We proposed an OCR system that depends on neural network CNN and for feature extraction we used Complete LBP (CLBP) with object wise strategy which is less computation complexity, preventing the noise and its related effects with improved segmentation based on character image. And before feature extraction we pre-processed image through all level of image processing skew correction, binarization, and noise removal, thinning and skeleton to correct orientation and remove noise in query image. And after pre-processing we use horizontal and vertical projection for segmentation of processed image.

Any non-manual recognition characteristic matching and drawing out method is comprised of particular and necessary phases. All of these phases is created because of its committed operation and likely result generation. According to this, the projected automatic OCR system includes fixed phases that may be brief as below phases:

i. Pre-processing
ii. Segmentation
iii. Extraction of Feature
iv. Pattern Recognition/classification

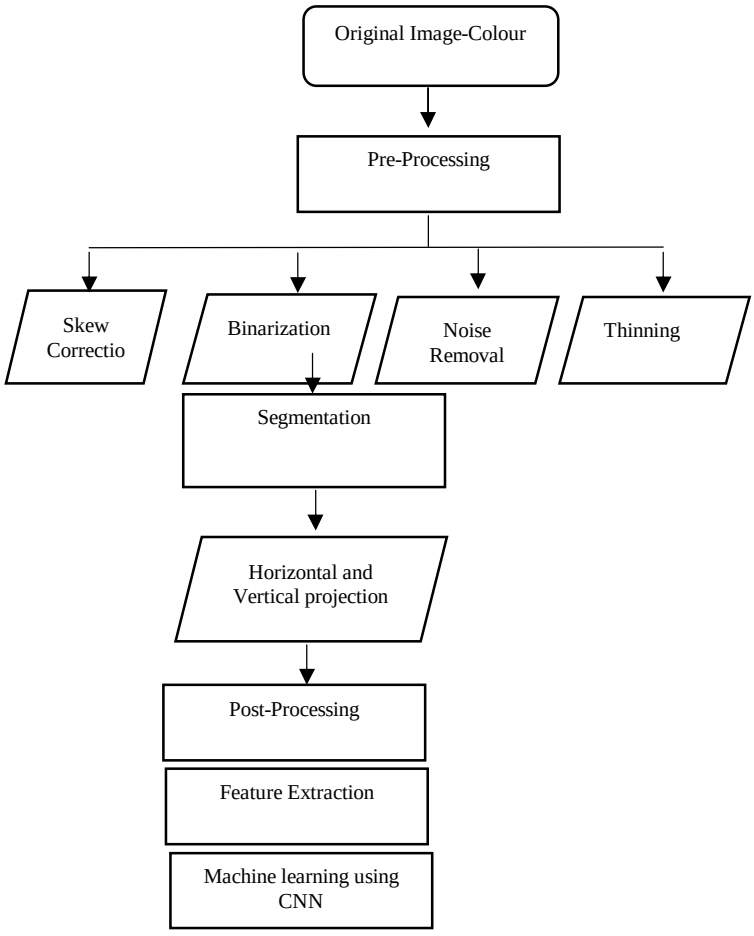

Figure 6. Procedural steps of projected framework

Pre-Processing

A word images acquired and pre-processed from the input Sheba language text before segmentation. In this phase first we correct skew orientation through Hough transform of the input image. Hough transform is considered to identify the lines, with the help of

parametric demonstration of a line. Hough parametric space is explaining in (a, b) form in the system of Cartesian coordinate. Here, a and b are line intercepts that is lying through point (x1, y1). It is 2-D space. Transform is

$$b = ax_1 + y_1 \qquad \text{Eq.7}$$

Hough transform in the Cartesian system is given in Fig. 2. given that dynamic range of parameters a & b is [−∞, ∞], but range of a is split in two parts: -1≤a≤1 and -∞≤a< -1 and 1<a≤ -∞, that's provided in below equations:

$$b = \begin{cases} -ax_1 + y_1 & -1 \le a \le 1 \\ -ay_1 + x_1 & otherwise \end{cases} \qquad \text{Eq.8}$$

The parametric space b=f(a).

Figure 7. (a,b) images of skewed document from Algerian newspapers; (c) image of skewed document from Arabic book.;(d, e, f) images of skew-corrected document of the images (a, b, c) correspondingly

After skew orientation, Binarization of image using Adaptive contrast is processed due to its performance on non-uniform lighting condition and remove over normalization issue of local maxima and minima than apply noise removal and thinning approach as discussed above.

Segmentation

It is not anything but divide the complete image into sub-images for advance processing, Histogram Projection technique, later than changing the colourful image contain black & white pixels only and that contain by binary image that means white colour pixel specify existence of foreground pixel & black colour shows nonexistence of this. As in technique, we calculate the number of foreground pixels alongside image all rows and the resulting histogram array size, equivalent to number of rows of image (image height). Analytical the resulting Horizontally proposed array as a histogram (graph plotting).

And Vertical Histogram Projection, calculate the number of foreground pixels with every image column and the resulting array is equivalent to number of image column (Width of image). Analytical the follow-on Vertically Projected array as a histogram (graph Plotting).

Feature Extraction

A quick colour segment technique was used to take out the characteristics of interest, the shape, area, namely, and texture of the character. LBP [66] is a easy but strong local texture descriptor depend on thresholding the neighbourhood of every pixel in opposition to the centre value and in view of the output as a binary code. The primary LBP values in a (P, R) neighbourhood are calculating as bellows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad \text{Eq.9}$$

The intensity value g_C correlate to the intensity value of the local neighbourhood's centre pixel, & g_P (=0,..., P-1) correlate to the grey values of the P pixels that is equally-spaced on a circle along radius R ($R > 0$) which create a neighbour set of circular symmetric.

But, within this proposed methodology we used Complete Local Binary Pattern CLBP over LBP. Textural characteristics extracted from already defined and training mitoses are the mainly important finished LBP elements, i.e. the sign & magnitude of Completed Local Binary Patterns (CLBP). The advantage and preference of these characteristics come back to their uncomplicated computation and its particular statistical modalities too. The Completed local binary patterns could calculate for distinct neighbouring radius (*R*) & pixels no. (*P*). In statement, the LBP sign & magnitude of a provided pixel in a computerized image in a neighbourhood among persistent the value of *R* and *P* is calculated shown below: Sign of CLBP,

$$LBP\_S_{R,P} = \sum_{i=0}^{P-1} t(g_i - g_c, 0)2^i \qquad \text{Eq.10}$$

Magnitude of CLBP,

$$LBP\_M_{R,P} = \sum_{i=0}^{P-1} t(g_i, g_c)2^i \qquad \text{Eq.11}$$

$t(x, c)=1$ if $x \geq c$, and $t(x, c)=0$ in other way. This value observing that c in above equations is merely a thresholding parameter utilized in the explanation of the piecewise and multivariable function $t(c, x)$.

The flexibility increases additionally to algorithm of CLBP in that is block-based, a new object-wise CLBP approach is employed in the characteristic taking out. The drawing out of object-wise characteristics contain below advantages above the pixel or block-based characteristic taking out:

- Low computational complexity because of low pixels consideration for characteristics extraction.
- Avoid from like the noise and its associated impact from being spread and obviously decreasing the sensitivity of noise.
- Giving segmentation improvement depends on biological units that are planned for eminent or acknowledged.

Table.2 Algorithm for Object-wise CLBP Feature Extraction Process

```
For i → n
//n is no segmented objects
        Evaluate the number of pixel in object (i) and place them in Os_region & their coordinates in (X, Y)
        // Os is all extracted objects
        // (X, Y) coordinate vectors of foreground pixels belonging to Os
        for j → Os_area
                Calculate the sign LBP_S_R,P and Mean LBP_M_R,P for every pixel of the Os within its boundaries because of associated (X(j), Y(j)) and P, R parameters
                //P is no of neighbouring pixels
                //R is neighbourhood radius
                if Os (i)=mitotic object then
                 index ← 1
                 //index is mitosis and non-mitosis label
                else
                 index ← 2
                end
end
build CLBP (i, index) for every object
```

Pattern Recognition

Categorization, a supervised knowledge technique, goal to allocate a label to not known patterns. It contains two phases:

a. Phase of training
b. Phase of testing

In first phase develop the classifier and the testing stage categories the not known patterns provided to the classifier. The patterns are utilizing for testing and training a classifier shall as unique. Over the phase of training, with training data the classifier is issued. The data of training carry patterns set among labels of their labels of class. The algorithm of classifier gain knowledge of as of the training facts and develops the model of classifier. Throughout the phase of testing, the classifier is provided along patterns whose label of class is to be decided with the help classifier. The different characteristics those are taking out from the images that utilized to categorize the unidentified images. Technically, a function *f is a* classifier that maps input characteristic vectors $\mathbf{x} \in \mathbf{X}$ to output class $y \in (1,...,C)$,where $\mathbf{X}$ is the characteristic space and C be the classes set to that the information can belonging.

.

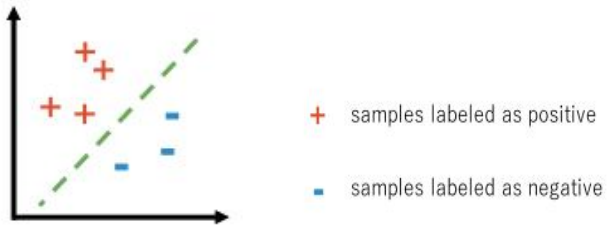

Figure 9. Supervised Learning

CNN neural network for prediction and detection system is used over SVM and other conventional classifier in this projected framework due to computationally efficient nature of machine learning and effective prediction with image processing without human intervention and following are the reason to use CNN:

- Unlike other classifier and method CNN can produce any number of outputs while SVMs have only one output.
- An n-ary classifier with a neural network CNN can be trained in one go while n-ary classifier with SVMs trains each of them one by one.
- SVMs are isolated systems whereas, CNN are one whole.
- CNN are parametric models while SVMs are non-parametric models

Performance Evaluation and Result

The proposed method implemented successfully using MATLAB simulation with image processing tool to evaluate and validate empirical results. In this work we tested different sheba character images as a test data over the above methodology.

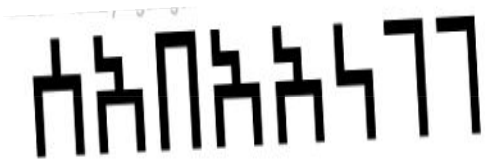

Figure 8. Input Sheba Character image data

Firstly, test image is pre-processed from colour to gray image and skew orientation as discussed above.

Figure 9. a. Skewed Sheba Character b. Sheba Character with zero Skew

Figure 10 a. Coloured Character b. Gray Character

After that we applied segmentation process using Horizontal and Verica projection as discussed in this methodology, the process starts by counting the intensity of white and black colour through histogram evaluation.

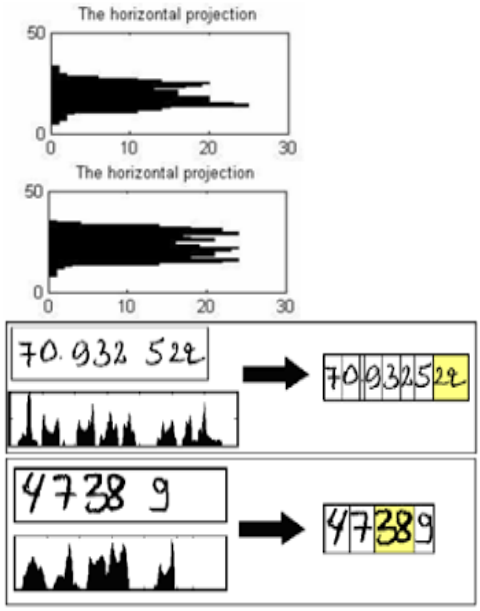

Fig. 11. a. Horizontal Projection, b. Vertical Projection

For the performance of this project model based on machine learning CNN algorithm we evaluate these two parameters

Accuracy: Evaluate the algorithm's performance in an explainable way which is done by Accuracy metric. To resolute afterwards the model standards and is measured in the form of a percentage. When comparing with true data find that how accurate models's augury is compared to the true data.

Loss Function: Optimization a machine learning algorithm which is by loss function. The loss is evaluated on training, validation and its understanding is depend on well the model is working in such two sets. That is addition of errors made for every illustration in training or validation sets. Loss value means how badly or well a model operates following every iteration of optimization.

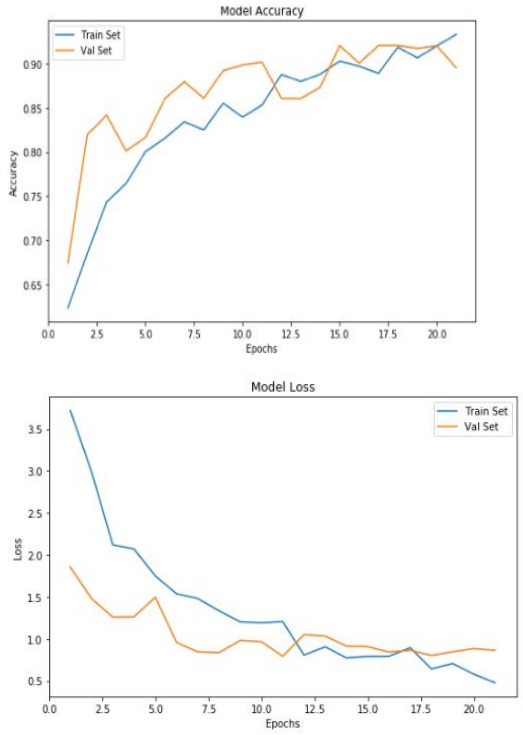

Fig.12 (a) Model Accuracy (b) Model Loss

CONCLUSION

An automated OCR system for Sheba character using Hough transform for skew correction, Adaptive contrast for binarization in pre-processing phase and CLBP based feature extraction due to its statistical modalities and simple computation with CNN neural network for pattern recognition is proposed this contain the concept of OCR processing techniques with their stages, necessity and applied research area on Character recognition is also analysed. By using this method Sheba character is

segmented, classified and pattern is predicted and analysed.

In future, the research study will be extended into unimproved approach for character recognition with high accuracy among the speedy upgradation in data innovation and information which is interactive, the utilization of highly developed solution is exposing every stage. For prediction CNN used in this method is a conventional method, in future we can use multi-classifier to overcome the limitation of this conventional approach and other method like support vector machine.

REFERENCES

1. Anagnostopoulos, C. N. E., Anagnostopoulos, I. E., Loumos, V., & Kayafas, E. **A license plate-recognition algorithm for intelligent transportation system applications.** *IEEE Transactions on Intelligent transportation systems*, *7*(3), 377-392. **.** (2006)
2. Asiri, A. M., & Khorsheed, M. **S**. **Automatic Processing of Handwritten Arabic Forms using** Neural Networks. In *IEC (Prague)* (pp. 313-317). **.**(2005, August).
3. Bassil, Y., & Alwani, M**.**. **Ocr post-processing error correction algorithm using google online spelling suggestion**. *arXiv preprint arXiv:1204.0191*. (2012)
4. Beg, A., Ahmed, F., & Campbell, PHybrid OCR **Techniques for Cursive Script Languages-A Review and Applications**. In *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks* (pp. 101-105). IEEE. **.** (2010, July).
5. Beg, A., Ahmed, F., & Campbell, P**. Hybrid OCR Techniques for Cursive Script Languages-A Review and Applications.** In *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks* (pp. 101-105). IEEE. (2010, July).
6. Biswas, B., Bhattacharya, U., & Chaudhuri, B. B. **A global-to-local approach to binarization of degraded document images**. In *2014 22nd International Conference on Pattern Recognition* (pp. 3008-3013). IEEE. (2014, August)
7. Camastra, F., & Vinciarelli, **A. Estimating the intrinsic dimension of data with a fractal-based method**. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(10), 1404-1407. (2002)
8. Deselaers, T., Gass, T., Heigold, G., & Ney, H.. **Latent log-linear models for handwritten digit** classification. *IEEE transactions on pattern analysis and machine intelligence*, *34*(6), 1105-1117. (2011)
9. Alwayle, I. M. G. Sabic Characters Recognition using Template Matching with Interpolation Technique Hadhramout University Journal vol. 10 issue 1 June 2013
10. Kumar, M., Jindal, M. K., & Sharma, R. K.. **Review on OCR for handwritten Indian scripts character recognition. In *International** Conference on Digital Image Processing and Information Technology* (pp. 268-276). Springer, Berlin, Heidelberg. (2011, September)
11. Mandal, S., Das, S., Agarwal, A., & Chanda, B. **Binarization of degraded handwritten documents based on morphological contrast intensification.** In *2015 Third International Conference on Image Information Processing (ICIIP)* (pp. 73-78). IEEE.. (2015, December)
12. Patel, C., Shah, D., & Patel, A. **Automatic number plate recognition system (anpr): A survey**. *International Journal of Computer Applications*, *69*(9). (2013).
13. Plamondon, R., & Srihari**, S. N Online and off-line handwriting recognition: a comprehensive survey.** *IEEE Transactions on pattern analysis and machine intelligence*, *22*(1), 63-84. **.** (2000)
14. Roy, A., & Ghoshal, D. P**. Number Plate Recognition for use in different countries using an improved segmentation**. In *2011 2nd National Conference on Emerging Trends and Applications in Computer Science* (pp. 1-5). IEEE. (2011, March)
15. **Sabaean Alphabet, available at** https://omniglot.com/writing/sabaean.htm
16. Sauvola, J., & Pietikäinen, M. **Adaptive document image binarization. *Pattern recognition***, *33*(2), 225-236. (2000).
17. Shinde, A. A., & Chougule, D. G. **Text pre-processing and text segmentation for OCR**. *International Journal of Computer Science Engineering and Technology*, *2*(1), 810-812.. (2012)
18. Siddiqua, S., Naveena, C., & Manvi, S. S **Recognition of Kannada Characters in Scene Images using Neural Networks**. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (pp. 146-150). IEEE. **.** (2019, November).
19. Simanic, S. ***South Arabia in Transition from the Ancient to the Islamic Era*** (Doctoral dissertation, uniwien). (2017).

20. Su, B., Lu, S., & Tan, C. L. **Binarization of historical document images using the local maximum and minimum.** In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems* (pp. 159-166). (2010, June).

21. Su, B., Lu, S., & Tan, C. L. (2012**). Robust document image binarization technique for degraded document images.** *IEEE transactions on image processing*, *22*(4), 1408-1417.

22. Gill, T.K**. Document Image Binarization Techniques- A Review**. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(5). (2014)

23. Wen, Y., Lu, Y., Yan, J., Zhou, Z., von Deneen, K. M., & Shi, P. **An algorithm for license plate recognition applied to intelligent transportation** system. *IEEE Transactions on intelligent transportation systems*, *12*(3), 830-845. (2011).

24. Woodard, R. D. (2004). ***The Cambridge encyclopedia of the worlds ancient languages*.** Cambridge univ. press.

25. Zheng, L., He, X., Samali, B., & Yang, L. T. **An algorithm for accuracy enhancement of license plate recognition**. *Journal of computer and system sciences*, *79*(2), 245-255. (2013)