International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May - 2021 ISSN: 2582-3930

A Survey on Handwritten Text Recognition

Balasubramanyam Kosuri [1], Prof.Mahendra Kumar B[2]

[1] *Balasubramanyam Kosuri, PG scholar, MCA, Dayananda Sagar College of Engineering*

[2] *Prof.Mahendra Kumar B, Asst.Professor, MCA Dept., Dayananda Sagar College of Engineering*

ABSTRACT:

Scientists have found out many technologies to find hand written character into text. Handwriting recognition is the way of software to read handwriting as an actual text. To process handwriting to text, this is the best software that can be used to overcome the many difficulties that are present in recognition of words. This paper mainly focuses on basic difficulties that are faced in recognition of the handwritten character with types of pattern recognition techniques. Based on the study taken for this problem, say that segmentation of the character is poor and prevents high popularity precision of profligate handwriting. Handwriting character recognition has acquired a lot of consciousness in the field of pattern recognition and machine learning due to its uses in various fields. Handwritten Character Recognition has zone to apply. Various practices have been proposed for character identification in handwriting recognition system. Even though, sufficient research and papers describe the approaches for converting text content from a paper document into machine readable format. In coming days, character recognition system might act as a key factor to create a paperless habitat by digitizing and processing existing paper documentations. This paper show cases a detailed practice in the field of Handwritten Character Recognition.

Keywords :

Handwritten Character Recognition , slant removal

Pattern Recognition, kohonen neural network, kannada, segementation

1. INTRODUCTION

Handwritten Character Recognition is a method of transforming handwritten text into machine readable format. There are mainly three stages in pattern recognition: observation, pattern segmentation and pattern division. Recognition of character has become very engaging topic in pattern recognition for the researchers during last few years. In regularity, handwritten recognition has been classified in to two types as on-line and off-line recognition . Offline handwriting recognition contains the automatic changing of text into an image into character codes which are usable within various softwares and text-processing applications. The data gained by this process is regarded as a static representation of handwriting. But, in the on-line pattern recognition, the two dimensional synchronization of successive points are show cased as a function of time and the order of strokes made by the author are also obtainable. Offline pattern recognition is comparably more challenging due to structure of characters, great differentiation of character symbols, types of handwritting styles and document condition. Several

recognition correctness, Character recognition is a rudimentary, but most difficult in the field of pattern recognition with huge number of functional applications.

It has been a tough field of research since the early days of computer's due to it being a natural way of communication between computers and human beings. More accurately handwritten recognition is the process of identifying and recognizing characters from the input image and converts it into ASCII or other similar machine language format. The process by which a computer system can identify characters and other symbols written by hand in natural handwriting is known as handwriting recognition system.

Handwriting recognition is divided into offline handwriting recognition and online handwriting recognition. If handwriting is examining and then understanding by the system, it is called offline handwriting recognition. In case, the handwriting is imagined while writing by means of touch pad using stylus pen, it is called online handwriting recognition. From the classifier means Handwritten recognition systems are divided into two main groups i.e. segmentation free (global) and segmentation based (analytic).

The segmentation also known as the holistic approach to identify the character without splitting it into sub parts or characters. Each word is considered as a set of global features, e.g. ambiance, like, cupids, etc. Whereas segmentation based approach each character/literature is divided into sub parts either uniform or non-uniform and sub divided parts are considered free of dependency. Handwritten character processing systems are realm and application particulars, like it is impossible to generate a generic system which can process all types of handwritten languages . Lots of work has been done on European languages and Islamic (Urdu),Egyptian languages. Whereas lcal languages like Telugu, Kannada, Bangla, Tamil, Malyalam etc. are very less concentrated due to limited access.

The Handwritten Character Recognition (HCR) is a vast area of research in Soft Computing, artificial intelligence (AI), pattern recognition (PR) and computer vision.

HCR is a basic process of handwritten text or digital images of printed so that they can be electronically imported, stored and browsed more efficiently and correctly. While study of ages and enhancement in this category, machines are still not even near to human's analyzing capabilities.

The objective of an HCR process is to recognize a manuscript (same as humans) in a difficult document. Various research attempts have been point of attention on innovative methods and schemas that would deplete the processing moment while offering loads of recognition precision .

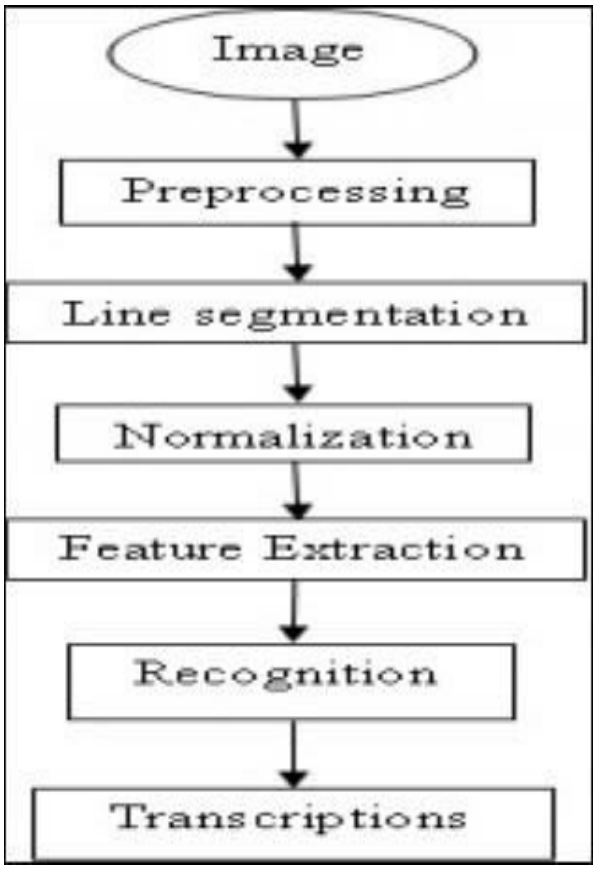


Image Acquisition Digital Image is generally taken as input. The most basic of these devices is the electronic tablet. These devices use a pen that is digitalized in nature. Input images can be taken by using other methods such as scanners, photographs or by directly inputting in the system by using a stylus. Pre-processing is the basic phase of character recognition and it's important for good recognition results. The main motto of pre-processing steps is to normalization of strokes and removing variations that would by mistake complicate recognition and decrease the recognition rate. These variations or discretions include the different sizes of text, missing things during pen movement, jitter available in text, left or right bending in the handwriting and unevenly distribution of points from neighboring places.

Pre-processing includes five steps:

- size normalization
- centering
- interpolating missing points
- smoothing
- slant correction
- Re-sampling of points.

Segmentation is carried by separation of the individual letters of an image. Generally document is analyzed in a hierarchical way. At first level each line is segmented by row histogram. Next step, words are extracted by column histogram and finally letters are pulled from words. The main point of feature extraction phase is to extract the pattern which is most pertinent for division. Feature extraction processes like Principle Component, Linear Discriminate s , Chain Code , Scale Invariant, Feature Extraction, zoning, Gradient based features, Histogram can be applied to obtain the features of individual characters.

Classification by input image is presented to HCR system; its habits are extracted and given as an result to trained classifier like artificial neural network machine. Classifiers differentiate the input with stored pattern and find out the most matching class for input. Post-processing is the procedure of correcting misclassified output by giving linguistic knowledge.

Post-processing is method of the output from shape recognition. Language information can make the accuracy obtained by shape recognition. Some shape recognizers yield a single string of characters, while others yield a number of alternatives for each character.

2. LITERATURE SURVEY

An early notable practice in the area of pattern recognition research was done by Grimsdale in 1959. The origin of a great deal of research work in the early sixties was related on an approach known as analysis-by-synthesis which was suggested by Eden in 1968. The great implication of Eden's work was that he basically proved that all handwritten characters are done by a finite number of schematic features, a point that was included in previous works.

This notion was further used in all methods in structural approaches of pattern recognition. K. Gaurav, this issue deals with the various pre-processing techniques involved in the pattern recognition with types of images ranging from a simple handwritten form based document and others containing colored and complex background. In this, different techniques like skew detection and correction, image enhancement techniques of contrast stretching, binarization, noise

However, even after applying all the said techniques it may not possible to gain the full accuracy. Here, we discuss hybrid Hidden Markov Model (HMM) model is proposed for recognizing unconstrained offline handwritten texts. The integral part of the optical model has been dealt with Markov chains, and a Multilayer Perception is used to determine the occuring probabilities. Different methods are applied to remove slope and slant from handwritten text and to reduce the size of text images with superior learning methods. The key features of this system is to develop a machine having high accuracy in preprocessing and pattern recognition, a modified quadratic classifier scheme to identify the offline handwritten numerals of six popular Indian scripts is suggested. Multilayer perceptron is used for recognizing Handwritten English letters. The needs are obtained from Boundary tracing and their Fourier Descriptors. The character is fetched by analyzing its shape and comparing its features that differentiate each character. Also an analysis has been taken out to determine the number of hidden layer nodes to obtain high deliverance of the back propagation network. A recognition accuracy of 92% has been reported for Handwritten English characters with least training time period.

3.PROBLEM IN HCR

Handwritten Character recognition (HCR) is a method of machine simulation of human interpretation. It is a skill set of obtaining, cleaning, understanding and segmenting pattern from an image. This procedure converts the image beneath into a revisable design pattern. process of handwritten characters usually is available in numerous fields like signature and courtesy amount of bank cheques , information filled in tax notations, reading zip codes on letters etc. Various techniques for extraction of features of letters are helpful ways as it focuses on local letters of the characters. Thus it assists to gain good knowledge of alphabets therefore updating the procedure of recognizing pattern. Task of Handwritten character recognition (HCR) is high in creating a paperless characters by converting older handwritten articles into electronic collection.

HCR gives to the growth of automation process and thus improving the communication between human and machine. It is not an easy job to build a software program to achieve its precision for identifying the handwritten English alphabets just as still though humans create errors to identify correctly. Handwritten characters differ based upon the writer.So there is always a necessity to enlarge a proficient handwritten recognition system. HCR has some potential uses which long for the requirement for developing such schemes in an advanced manner. It facilitates to decrease the need to save the data .

4. OBJECTIVES

The main objectives of are:

- Binarization
- Noise reduction
- Skew correction
- Slant removal

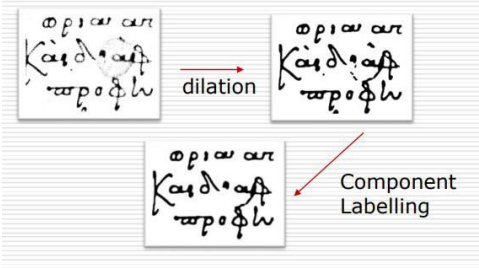
4.1 Binarization

Document image binarization refers to the changing of a gray-scale image into a binary image. Two types of thresholding.

Global, holds one threshold value for the entire document image which is based on an estimation of the background level from the intensity histogram image. global threshold parameter is chosen by using Otsu's method.

Adaptive, uses different values for each pixel to the local area information

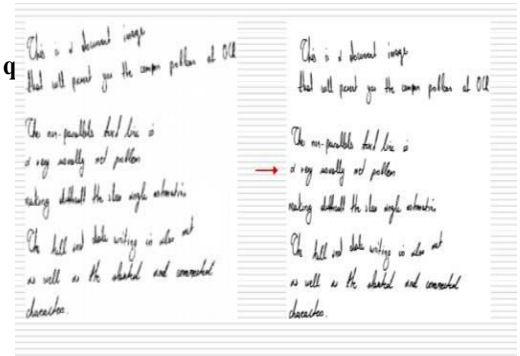
4.2 Noise Reduction.



Noise reduction improves the clarity of the document. Three main approaches Filtering, Morphological Operations, Component Labelling.

4.3 Skew Correction

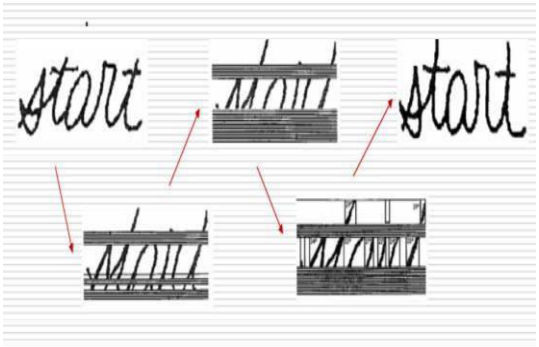
Skew Correction methods are used to align the paper document with the preceding system of the scanner. Main techniques for skew detection include correlation, projection profiles, Hough transform.



q

4.4 Slant Removal

The slant varies from user to user. Slant removal methods are used to normalize the all characters to a standard format.



4.5 Segmentation

Text Line Detection ,Word Extraction vertical projections, connected component analysis.

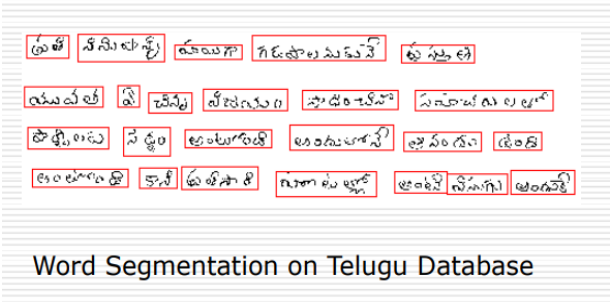
4.5.1 Explicit Segmentation

Here one tries to identify the smallest possible segments that may be smaller than each unique letter, but surely cannot be segmented further. Further in the recognition process these primitive segments are assembled into letters based on input from the character reorganization machine.

.

4.5.2 Implicit Segmentation

In implicit the words are recognized entirely without segmenting them into letters. This is most effective and applicable only when the set of possible words is small and known in advance.



Word Segmentation on Telugu Database

4.6 Feature Extraction

In feature extraction each character is presented as a feature vector, which becomes its identity. The major aim is to extract a set of possibilities, which maximizes the recognition with the least amount of elements to be searched. Obtaining these qualities is a difficult task.

Feature extraction methods are based on these types:

Statistical

Structural

Global transformations

4.7 Statistical Features

Represents an image by statistical distribution of points taking care of style variations in pattern. The major features used for character representation are:

- Zoning
- Projections and profiles
- Crossings and distances

5. CHALLENGES

For best and high accuracy pattern recognition, HCR techniques expect high quality or high resolution images with some basic structural properties like different text and background. The way images are generated is an important and determining factor in the identification, since this often affects the quality of images.

Usually HCR with images scanned by scanners gives high results and good performance. In contrast, images produced by cameras usually bad input as scanned images to be used for HCR due to the environmental factors. Numerous errors might occur.

The style in Islamic,Egyptian writing is extremely different from other languages as English which includes a letter after another as many similar other scripts as Mandarin. These scripts are plausibly followed in both printed script as well as documents of handwritten recognition.

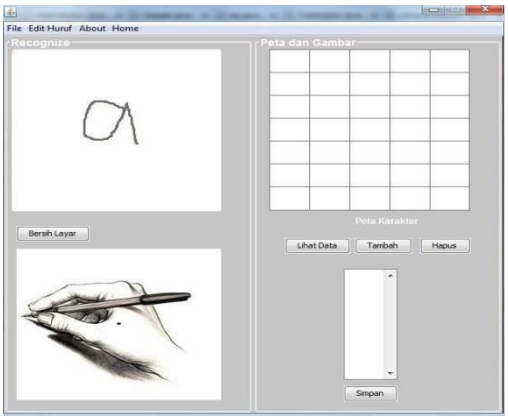
Words in Urdu script follow somewhat blank style in writing where before the previous character is finished. The most pain staking problem in the words is the gap between the letters inside one word. Text is often split into words, sub-words. A number of these characters are not connected to the next character. The recognition system accuracy may be affected by the multiple.

6. HCR USING KOHNEN NEURAL NETWORK

Kohonen Neural Network method is an unsupervised learning process of studying the distribution of a set of pattern without any sufficient class of information. The general idea of this process is understood from how human brain takes input images/patterns that have been decoded through eyes, and then human brain is able to reveal the output back. Therefore, the application of this model is widely used in handwritten pattern recognition.

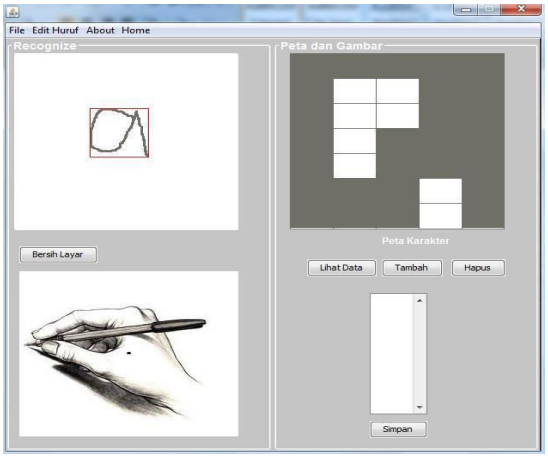
6.1 Drawing Characters

Step one of HCR process is to draw characters on canvas. When the mouse is released from the canvas, there will be a straight line at the top, right, left, and bottom of picture later the image is reflected.



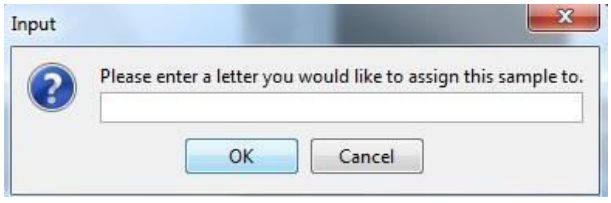
6.2 Mapping Image

When mouse is cancelled, then the mapping will be performed into canvas map.



6.3Saving Character Data

This can be performed by pressing plus sign button on keyboard, and there will be a character input formed.



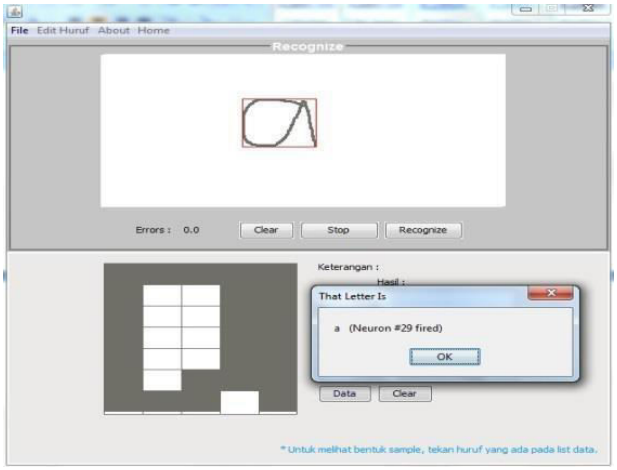
6.4 Comparing Pixels

Pixel differentiation is the process of the image in handwritten character recognition input resulting on canvas with the data that is saved in backup.

```
 :00000000000000000000000000000000
A:00110001100111001010111111100110001
B:11111100011000111111110011000110111
C:11111100001000010000100001100001111
D:11111100011000110000100011000111111
E:11111100001000011111100001000011111
F:11111100001000011110100001000010000
G:01110110001000010111100011000111111
H:10001100001000111001111111000110001
I:11111001000010000100001000010000111
J:11111001000010000100101001010011100
K:10001100111111101101010010100111100 11
L:10000100001000010000100001000011111
M:11111101111011110100101001010110101
N:11111110111000110001100001000110001
O:11111100011000110000100011000111111
P:11111100011001111110100001000010000
Q:01111110011000110001100111101101111
R:11111100011000111011111101001110001
S:01111100011000011110000100001111111
T:11111001000010000100001000010000100
U:10001100001000110001100011001111110
V:10001100011101101011011100111000110
W:10101101011010110101101011011111111
X:10011110100111001100111001011010010
Y:10001110110111001100010000100001000
Z:11111000110011001100110001000011111
a:11111100111001110011100111011111101
b:10000100001000010000111111100111111
```

6.5 Character Recognition

Pixel comparison process is done using in order to be able to recognize characters that are written in hand.



7. METHODOLOGY

The performance of HCR models on multilayer artificial neural networking. For computer terminology, the most basic types are recurrent neural networks (RNN) or more precisely long short-term memory (LSTM), and convolution neural networks (CNN).

Machine Learning focused on using gradient-based learning techniques using multi-module, a precursor to some of the interval end-to-end modern deep learning. The next major development in producing HCR accuracies was the use of a Hidden Markov Model. This approach uses letters as a state, which then allows for the context of the character to be available for when determining the next hidden letter. This leads to higher results compared to both feature extraction techniques and the Naive Bayes approach

.

The main drawback is still the manual extraction, which requires prior knowledge of the language needed and was not particularly good to the diversity and complexity.

One of the most efficient papers for the task of handwritten text recognition is Scanning, Attend, and Reading: End to-End Handwritten Paragraph Recognition with MDLSTM Attention language. This approach is to take an LSTM layer for each scanning direction and decode the raw image data to a feature. The model would then use attention to get certain feature maps over others.

7.1 Where it has been used

7.1.1 Healthcare and pharmaceuticals

Patient prescription digitization is a major hectic task in healthcare industry. The company Roche handles millions of petabytes of medical PDFs on daily basis. Another area where handwritten pattern recognition has key impact is patient enrollment and form digitization.

7.1.2 Insurance

An insurance company receives more than 10 million documents a day and a delay in processing the claiming amount by the customer can impact the company terribly.

7.1.3 Banking

People write cheques on a daily basis and cheques play a vital role in most non-cash fund transactions. In many developing countries, the general cheque processing pattern requires a bank employee to read and manually

Enter the details present on the cheque and also verify the credentials. Handwriting pattern recognition system can save cost effectively and hours of human labor.

7.1.4 Online Libraries

Large amounts of historical data are being digitized by uploading the image scans for granting to the entire world. But this method is not very help full until the text in the images can be identified later can be indexed, queried and searched.

7.1.5 Algorithms used for Kannada Handwritten Character Recognition

We proposed an HCR system to recognize kannada characters in a poetry . A CNN model called AlexNet is used for classification in the system. We recorded an accuracy of 92% for kannada language.

We propose this system using a CNN for classification using a char74K dataset.The system recorded an accuracy of 98% .

A Hidden Markov Model is used and SIFT methods are used .

For classification an HMM model Baum-welch algorithm is applied, and the method was able to result in accuracy of 66%.

8. ADVANTAGES

Handwriting recognition software allows everyone to translate all those signature and notes into electronic characters in a text document form. The advantage of this electronic storage is that this data only requires very less physical space than the storage of the physical data.

- Cost Reduction
- Superior Data Security
- 100% Text-searchable Documents
- Makes Documents Editable
- Disaster Recovery
- High Accuracy
- Increased Storage Space

8.1 How does it work?

The handwriting to be recognized is digitalized through scanners or cameras. later the image of the document is segmented into lines, words, and individual characters. Finally, each character is recognized using HCR techniques. Errors are checked using lexicons or spelling checkers.

8.2 How is it helpful?

Alternatively, the movements of the pen tip may be sensed "on line". A handwriting recognition system handles the formatting of texts, performing correction of segmentation into characters, and finds the most suitable words.

9.FUTURE WORK

Creating new hierarchical classification schemas based on the rules after the clear examination of the corresponding confusion matrix.

Exploiting latest and new features to improve the current performance of the system.

10. CONCLUSION

The paper discusses in detail all advances and advantages in the area of handwritten pattern recognition. The most accurate solution provided in this area directly or indirectly depends upon the quality, quantity as well as the nature of the material to be read by the user. Various techniques have been described in this paper for patterns in handwriting recognition system.

From the study done so far, it is observed that the selection of the objects as well as the feature extraction techniques need to be proper in order to attain good results in recognizing the character. Studies in the paper reveals that there is still scope of enhancing the algorithms as well as enhancing the rate of recognition of pattern in characters.

REFERENCES

[1] Iwata S, Osama W, Wakaba , Kimura F (2016) Recognition and transition framing of Urdhu news for video retrieval trends. 23rd international conference on character recognition (ICCR) 2016.

[2] K. Gaurav and Bhatia P. K., “Techniques for Offline Handwritten pattern Recognition”, 2nd International ." Document Analysis (IDA), 2017 14th IAPR International Conference on. IEEE, 2018.

[3] Puigcerver, Joan."Multidimensional Recursion Layering Necessary for Handwritten Pattern Recognition.

[4]. Pavlidis T. Algorithms for pattern and image processing in hand written pattern recognition. New Jersey: Murray Hill; 1972.

[5] Rehman A, Mohamad D, Sulong G. Implicit vs explicit based script slan removal and recognition: a performance comparison on benchmark database system. Computing Mathematics. 2019.

[6]. Asworo, Ed, "Comparision Between Kohonen Neural Network and Learning Vector Quantization process on Real Time Hand pattern Recognition System", Institute of Technology Surabaya 2014.

[7]. Suresha M, Ali AAA. segmentation of handwritten text along with touching of line. 2018.

[8]. Rajashekararadhya, S. V., and P. VanajaRanjan. "Neural networking referencing on handwritten pattern recognition of Kannada scripts." TENCON 2008-2008 IEEE Region 10 Conference. IEEE, 2009.