Horizontal and Vertical Projection techniques for Line & word Segmentation Process in Offline Handwritten Gujarati Text

Ashwin R. Dobariya, Prof. V. R. Rathod

Abstract— This research paper describes two important techniques Horizontal & Vertical Projection for character segmentation process in Offline Handwritten Gujarati Text Recognition process (OHGTR). Segmentation is one of the key steps in Text Recognition process which is one of the factors in recognition accuracy. This research paper mainly focuses on segmentation techniques: Horizontal & Vertical projection. Recognition process is required to perform many preprocessing and post processing steps to recognize the character. If segmentation of scan documentation is properly happen then rest of the process will become easy otherwise it will reflect on final recognition process. It is easy to segment printed character scan document as compare to hand written document because of the same font size and style where in hand written text document is , the content character size is different, even the same document is written by the same person.

Keywords— OHGTR, Pre Processing, Post Processing, Horizontal & Vertical Projection, Offline Hand written Text Recognition (OHGTR)

1. INTRODUCTION

Handwritten Character recognition (HCR) is one of the challenging research areas in image processing. Some other computational areas like artificial intelligence, expert systems and Neural Network have provided an important role in recognition process of handwritten Gujarati characters. The key features in recognition process are speed, efficiency & accuracy to recognize characters. The human can easily identify hand written character, if it is written differently by same person because human is having his / her own intelligence. In case of computer system , it is little bit difficult to identify , hand written character because the same person may write the same document in another way with minor different in character size and style. This research paper is mainly focus on how to segment character from the scan document like line segment, word segment and character segment for the next recognition process. Recognition of handwritten characters process is having some pre-processing steps and post processing steps. Some sub-processes are like scanning a hand written document and create an image file, pre-processing, feature extraction etc. The output of previous step is used as input in next step. Each step is important in recognition process as far as the accuracy and speed is concern.

2. CHARACTER RECOGNTIION PROCESS

Character recognition process is mainly categorized into two types according to the manner in which input is provided to the recognition engine. The following graph describes the classification hierarchy of character recognition. [1]

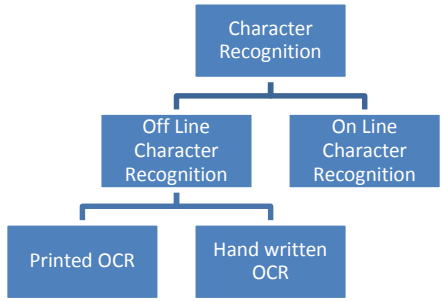

Fig. 1 Classification of Character Recognition

Types of character recognitions are (1) Off-line character Recognition and (2) On-line character recognition. In general recognition process means identifying a character from image document and converts it into machine editable format so same content can be modified and store for long time as soft copy back up. On-line handwritten character recognition process deals with automatic conversion of characters, which are written on a special digitizer, tablet PC etc; where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. In this process, the written character is directly identified & recognized at the time of writing the text. For this type of recognition process, it is required a digitizer type of devices. Off-line handwritten character recognition process deals with a data set, which is obtained from a scanned handwritten document. In this process, text is available on paper and after that paper is process in different phases of recognition process and finally identified & recognized the text. Off line handwritten character recognition process is also classified in further two categories namely (1) Offline Printed Text Recognition and (2) Offline Handwritten Text Recognition. In case of Offline Printed Text Recognition, data is available on paper but all characters are written in equal size and in proper style so it is comparatively easy to recognize as compare to hand written because font size and type of the characters are same in throughout scanning document. It does not require much processing and give comparative good accuracy in recognition. On the other side, Offline Handwritten Text Recognition, data is available on paper but it is written by person and the size of character, style of character are differ from different person and it is little bit difficult to identify and recognize a hand written character because, if a same user write the same paragraph manually again then also character style is different every time. Based on the recognition classification & methods of different languages like Devanagari, Bangla, Tamil, Telugu, Gurumukhi etc, it has been concluded basic steps for GHTR (Gujarati Hand written Text Recognition) process & Model with Artificial Neural Network approach as per follow.[2]

(1) Scanned image of hand written Gujarati Text document
(2) Binarization/ Digitalized of scanned image
(3) Removal of Noise data from scanned image
(4) Skew detection and correction of scanned image,
(5) Divided in Line & character segmentation of image [4]
(6) Feature Extraction Techniques
(7) Recognition Process.

3. HORIZONTAL & VERTICAL PROJECTION FOR SEGMENTATION PROCESS.

Segmentation is very important step for any Optical character Recognition system. The errors or mistake in segmentation process will replicate to next recognition process. An efficient segmentation method is needed to improve the overall recognition rate of any handwritten text. Due to variation in handwritten of persona to persona, the techniques which are used for printed , it will not be directly used in hand written text segmentation process. General steps of hand written text segmentation process are:

(1) Text Line Segmentation
(2) Word Segmentation
(3) Top modifier segmentation
(4) Bottom modifier segmentation
(5) Left modifier segmentation
(6) Right modifier segmentation
(7) Single character segmentation.

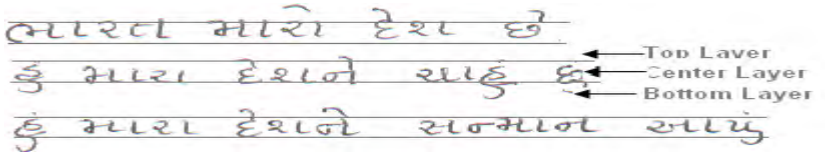

Fig.2 Layer Classification in Text

Text Line Segmentation: In this segmentation process, scanned document is divided into individual lines. Horizontal projection profile is the most commonly used technique to separate lines in the scan document. Lines can be easily separated by white spaces. If the white space is continuously repeat horizontally on the same row, it means there is a separation between two lines. This method works well in printed text as compare to hand written text because in handwritten. Documents lines may not be straight and can have horizontal overlapping between the lines. Some of the lines in handwritten documents are

zigzag i.e. lines form curvature between different lines as shown in Figure 2. After carefully analyzing handwritten documents, it has been observed that, some documents that is written in proper lines without major skew that can be easily separated using horizontal projection method and some of the documents may contain overlapped lines as shown in fig.2 and skewed lines that are difficult to segment.

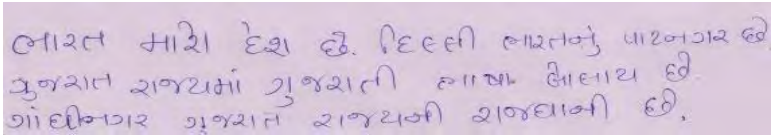

Fig.3 Overlapped Lines in the scan document.

General Algorithm steps for Line Segmentation using Horizontal Projection

1. Scan hand written document and make it skew & slant free.
2. Calculate sum of the white pixel in each row.
3. Find out the rows containing no white pixel. It means black pixels are there.
4. Replace all such rows by 1 or black pixels.
5. Invert the image to make empty rows as 0 and text lines will have original pixels.
6. Mark the Bounding Box for text lines.
7. Copy the pixels in Bounding Box and save in separate file as a line image.

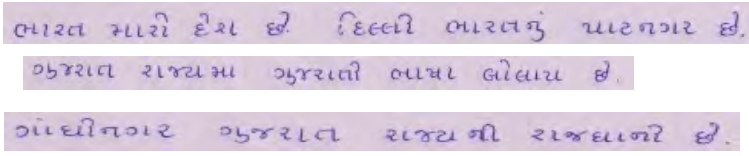

Fig.4 Line Segmentation of scan document.

Using horizontal projection with Histogram also easy to segment line from scan document image. This method is effectively work in printed text document because where all font size and style are same.

Word Segmentation: The word segmentation is performed after text line segmentation process. Once the text lines are segmented from scan document text, words are segmented from lines by line using vertical projection method. Column with zero black pixels is used as delimiter for word or column with all white pixels is the separation between two words.

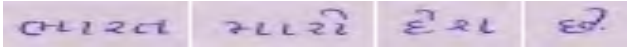

Fig.5 Word Segmentation using Vertical Projection

The following algorithm is used for word segmentation:

Step 1: For each column of the line the number of black pixels is counted and the columns with zero black pixels are used as delimiters for word separation. If there is no black pixel in a particular column means, there is no availability of text pixel.

Step 2: Column position with zero black pixels is identified and stored in a one variable like p1.

Step 3: Word is selected from column 1 to variable p1 position and create a box and separate it and saved in as an image for further recognition process. In this technique document is into vertical strips and vertical projection profile can be obtained by pixel by pixel along with each value on the x axis for each y value. From this gaps between the lines can be observed. After finding the projection values curved can be smoothed by applying some filtering techniques. In this way, horizontal and vertical project are used for segmentation process of scan document image. After segmentation of word the last segmentation processing step is character segmentation from word. In this process, it required to separate different top, bottom, left and right modifiers of the character to finally recognize the character.

4. CONCLUSION

In this research work a simple off-line handwritten Gujarati script recognition process is used. Two segmentation techniques Horizontal & Vertical projection. The main objective of this research paper is to implement a technique in Gujarati Character recognition to improve recognition accuracy. Segmentation of line and characters has been implemented. Here using horizontal projection profile for line segmentation and vertical projection for word segmentation. The future studies are directed towards segmentation of touching characters in a better way to improve the segmentation accuracy.

5. REFERENCES

[1] Handwritten Gurumukhi Character Recognition Using Neural Networks, Naveen Garg, Computer Science And Engineering Epartment, Thapar University, Patiala

[2] Gujarati Script Recognition: A Review By : Mamta Maloo, Dr.K.V.Kale, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011, ISSN (Online): 1694-0814

[3] Handwritten Character Recognition System Using Artificial Neural Networks, Pelin GORGEL Oguzhan O, ZTAS, Journal Of Electrical & Electronics Engineering

[4] Efficiency zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts", S. V. Rajashekararadhya, and P. Vanajaranjan, Journal of Theoretical and Applied Information Technology, JATIT, vol. 4, no. 12, pp. 1171-1181,2008.

[5] "Devanagiri document segmentation using histogram approach",Vigas J Dongre and Vijay H Mankar,International Journal of Computer Science, Engineering and Information Technology,2011

[6] Gujarati Character Identification: A Survey, Mitul Modi, Fedrik Macwan, Ravindra Prajapati, International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering, Vol. 2, Issue 2, February 2014

[7] Segmentation of Offline Malayalam Handwritten Character Recognition, Sangeetha Sasidharan & Anjitha Mary Paul, IJARCSSE, Volume 3, Issue 11, November 2013

[8] "Gujarati handwritten numeral optical character reorganization through neural network", A. Desai, , Pattern Recognition Vol 43 2010 pp. 2582-2589

[9] Off-Line Cursive Handwriting Recognition Using NNs, A. W. Senior, 1994, University of Cambridge, England

[10] A System for Off-Line Oriya Handwritten Character Recognition Using Curvature Feature, Wakabayashi, T. Kimura, Information Technology, (ICIT 2007). 10th International Conference , 2007 IEEE.

[11] Extraction of Characters and Modifiers from Hand written Gujarati Words, Chhaya Patel, Apurva Desai, International Journal of Computer Applications, Volume 73– No.3, July 2013

[12] Kannada , Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network : A Novel Approach, B. V. Dhandra, R.G.Benne, M. Hangarge, Recent Trends in Image Processing and Pattern Recognition, pp. 83-88, 2010

[13] Character Recognition of Gujarati and Devanagari Script : A Review, S. S. Magare,Y. K. Gedam,D. S. Randhave,Prof. R. R. Deshmukh, International Journal of Engineering Research & Technology (IJERT),Vol. 3 Issue 1, January – 2014

[14] Offline Handwritten Devanagari Script Recognition, Ved Prakash Agnihotri, I.J. Information Technology and Computer Science, 2012, 8, 37-42, Published Online July 2012 in MECS

[15] Off-Line Cursive Handwritten Tamil Character Recognition, R. J. Kannan, R. Prabhakar, 2008, Signal Processing, vol. 4, no. 6, pp. 351-360.

6. AUTHORS' PROFILE

Mr. Ashwin R. Dobariya is working as an Assistant. Professor & GTU Coordinator at Faculty of Computer Application, Marwadi Education Foundation Group of Institutions, Rajkot since 2009. He is pursuing his Ph.D in the area of Text Recognition System. He is a member of CSI & having more than 10 years of academic experience. He has been awarded for "Best Outstanding Performer Award" by MEFGI in Feb 2015.

Prof.(Dr.)V.R.Rathod was a Prof.& Head of the Dept. of Computer Science, Bhavnagar University with more than 40 years of academic & research experience. More than 12 research scholars have been awarded Ph.D under his guidance. He has published more than 50 research papers in various National & International journals.