iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** VI **Month of publication:** June 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Offline Handwritten Sanskrit Character Recognition

Sujata S. Magare[#1], Ratnadeep R. Deshmukh[*2]

[#] *Department of CS-IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (M.S.), India*

Abstract— In this paper, we describe the procedure of developing dataset for offline handwritten Sanskrit character recognition. There is no standard dataset available for handwritten characters. Researcher has to develop own character dataset. This paper will provide a way for researcher to develop a dataset for offline handwritten Sanskrit character recognition. This paper describes use of Hough Transform and Euclidean Distance classifier for recognition. This paper describes basics of dataset, challenges associated with it and its processing.

Keywords— Offline character recognition, Binarization, Euclidean Distance, Hough Transform.

I. INTRODUCTION

Optical character recognition is the process of recognizing optically scanned characters. Character recognition has two types: Offline and Online .One of the challenging problem in pattern recognition is Offline character recognition. Offline character recognition takes scanned image of required document paper. Offline character recognition can be done in two ways: Handwritten and Printed.

Handwritten character recognition is abbreviated as HCR; handwritten characters have number of variations as different people have different writing styles. HCR can recognize offline character and online characters. Offline HCR takes input from scanned image of paper document and Online HCR takes input from digital pen [1]. There are many handwritten historical documents exist in electronic form, HCR is used to recognize such documents.

Researcher has to develop own character dataset collected from minimum 10-15 people, because there is no standard dataset available for handwritten characters. Different people have different writing styles, which includes variations in dataset.

A. Sanskrit

Sanskrit is the classical language of Indian and the liturgical language of Hinduism, Buddhism, and Jainism. It is also one of the 22 official languages of India. The name Sanskrit means "refined", "consecrated" and "sanctified". It has always been regarded as the 'high' language and used mainly for religious and scientific discourse. Sanskrit has 36 consonants and the 16 vowels.

The oldest known text in Sanskrit, the *Rigveda*, a collection of over a thousand Hindu hymns, composed during the 2nd millennium BC.

Today Sanskrit is used mainly in Hindu religious rituals as a ceremonial language for hymns and mantras. A modern form of Sanskrit is one of the 17 official home languages in India.

Since the late 19th century, Sanskrit has been written mostly with the Devanagari alphabet. However it has also been written with all the other alphabets of India, except Gurumukhi and Tamil, and with other alphabets such as Thai and Tibetan. The Grantha, Sharda and Siddham alphabets are used only for Sanskrit.

II. DATASET DEVELOPMENT

Offline Handwritten character recognition takes scanned image of required document paper. For this purpose we have taken a blank paper. Small blocks created on that blank paper. We have also added Name, Age and Sub-code field to the document. Although these fields are optional, but it is good to keep information about user which will help us to recognize handwritten styles with different age groups. After this block creation document paper is ready to take data from user. We have taken dataset from 26 people. We have given two paper

ISSN: 2321-9653

sheets for each person to take data from them. Each paper sheet contains 16 characters and person has to write each character at three times so that variation can be included in that.

Fig. 1 Scanned Document

Now paper documents are ready to Digitize. We have scanned each paper document at 300dpi and stored that document in JPEG file format for the better resolution of image and store it in separate folder. Then we have cropped each word from the scanned document and saved cropped image in BMP file format.

B. Image Cropping

For Image cropping firstly we have used Paint Tool. While working with Paint tool we have noticed that it is quite time consuming as each time image need to cropped then need to copy and then paste in new file. Finally need to save in BMP format. To save our time we have used Screen Snaper Tool. Screen Snaper is very easy and fast tool for image cropping. Screen Snaper is a gadget that will add a powerful, flexible and intuitive screen-capture utility that allows us to capture anything from the screen.

Fig. 2 Cropped Image

Screen Snaper can capture full screen images and windows or objects in the screen. Screen Snaper can also capture regions from the screen. It Copies image to clipboard automatically and automatically save captured images. Screen Snaper Saves images in 7 popular formats:

BMP/EMF/GIF/JPG/PNG/TIFF/WMF.

III. PREPROCESSING

Preprocessing technique is used to do improvement of image data that enhances some image features required for processing and suppresses unwanted noise and distortion from image data and aims to correct degradation in an image [1].

A. Binarization

Binarization is the process of converting grayscale image in to binary (Black and White) image, so that image data will only contain 0 and 1. Binarization technique is usually used for separating foreground from background using required level of thresholding.

B. Noise Removal

Digital image consist of variety of noises. These noises are required to be removed from an image for better processing. Morphological operation, Median filter and Weiner filter is used to remove noise from an image. Median filter reduces blurring of edges.

C. Thinning and Filling

Smoothing implies both Filling and Thinning. Thinning reduces width of character while Filling eliminates gap, small breaks and holes in digitized character.

D. Normalization

To obtain characters of uniform size, rotation and slant Normalization is applied on image. To improve the accuracy of character recognition Normalization reduces shape variation.

E. Skew Detection and Correction

During the digitization of document page it is often that image is not align correctly or it may be happen by human while writing document. To make in correctly align Skew detection and correction technique is used.

Skew detection technique can be classified in to groups: Analysis of Projection profile, Hough transform, clustering, connected component and correlation between line techniques.

IV. SEGMENTATION

Segmentation of an image is the process of subdividing image into number of parts. Segmentation takes the form as Paragraph Segmentation, Line Segmentation, Word Segmentation and Character Segmentation.

Line wise segmentation can use a horizontal projection profile based techniques. Character wise segmentation divides words into characters [2]. Chain code histogram can be used for each segment. Horizontal projection file method is used for segmentation.

We have used Hough Transform for line detection with Prewitt edge detection for vertical edges in the character. It is a discrete differentiation operator, computing an approximation of the gradient of the image intensity function [3]. The Prewitt operator is based on convolving the image with a small, separable, and integer valued filter in horizontal and vertical directions and is therefore relatively inexpensive in terms of computations.

V. FEATURE EXTRACTION

Feature extraction technique is aims to extract the essential and important features & characteristic of the given image and remove redundancy from image data. For feature extraction we have taken structural features. We have found presence of Vertical bar, its position in the character, Presence of Enclosed regions and End point detection. Position of a bar implies that it is at Middle or End. If bar position is less than middle column then it is a Middle Bar and if it is equal or greater than middle column then it is an End Bar.

Enclosed region are detected by tracing the region boundary of binary image. Connected Component method can also be used to detect enclosed regions.

Character that used as feature for segmentation generates FCC. To solve the problem of finding several branches and revisiting the same node, Heuristic can be used to generate FCC correctly to represent character. In Pattern recognition this is one of the difficult stages to implement. Selection of right feature extraction technique leads to achieving high performance for recognition.

Feature extraction technique is divided into three groups : Distribution of points, Transformation & series expansion and Structural analysis. Structural analysis extract the feature which represents geometric and topological structure of character [4]. Structural analysis gives feature with high tolerance of noise and style variation. Commonly used features are intersection between lines and loops.

VI. CLASSIFICATION

Classification aims to classify features according to its properties. Training and testing is done at the classification phase. Number of classifier can be used to train the character. We have used Euclidean Distance Classifier.

We have calculated Euclidean distance between the input image and training images. After calculation, we have considered those classes whose Euclidean distance is minimum. Accordingly, testing has been done and Output is displayed.

In paper [5], they have used K-NN method at classification stage. Although K-NN method is straight forward and powerful it consumes a lot of time, proposed accelerating the GAT correlation method by reformulating its computational model and adopting efficient lookup tables to reduce the computational cost of matching in *k*-NN classification.

Neural network is one of the well known classifier used for character recognition system. Neural network advantage of their adaptive nature. Feed forward NN and Back propagation NN is used for character recognition[6].

VII. RESULT AND DISCUSSION

We have collected a database of 32 compound characters from 26 individuals of different groups. Resulting into 78 samples for each character. There are 2496 total characters into the database. These samples are preprocessed and applied for segmentation. At preprocessing stage Morphological opening operation fill the gap between characters and removes unwanted strokes from character image. And then it is thinned. At the segmentation stage Vertical lines are detected using Hough transform and Prewitt Edge detection.

Position of a vertical bar and presence of enclosed region is detected at the Feature extraction. After extracting features, image is then normalized to 32x32 images. This image is then proceeding for a Classification. Minimum Distance classifier used to classify s. We have calculated Euclidean distance between input image and training images. Class, whose distance is minimum, is qualified for the recognition. Based upon these testing has been made and Corresponding Recognition result is displayed.

Proposed system is designed using MATLAB R2013a. Proposed system provides 90% recognition rate.

ACKNOWLEDGMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. Authors are thankful to the Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India, for providing necessary facilities for carrying out research work.

REFERENCES

[1] S.S.Magare, R.R.Deshmukh, Y.K.Gedam, D.S.Randhave, "Character Recognition of Gujarati and Devanagari Script: A Review", *IJERT*, Vol.3, Issue 1, pp. 3279-3282, Jan. 2014.

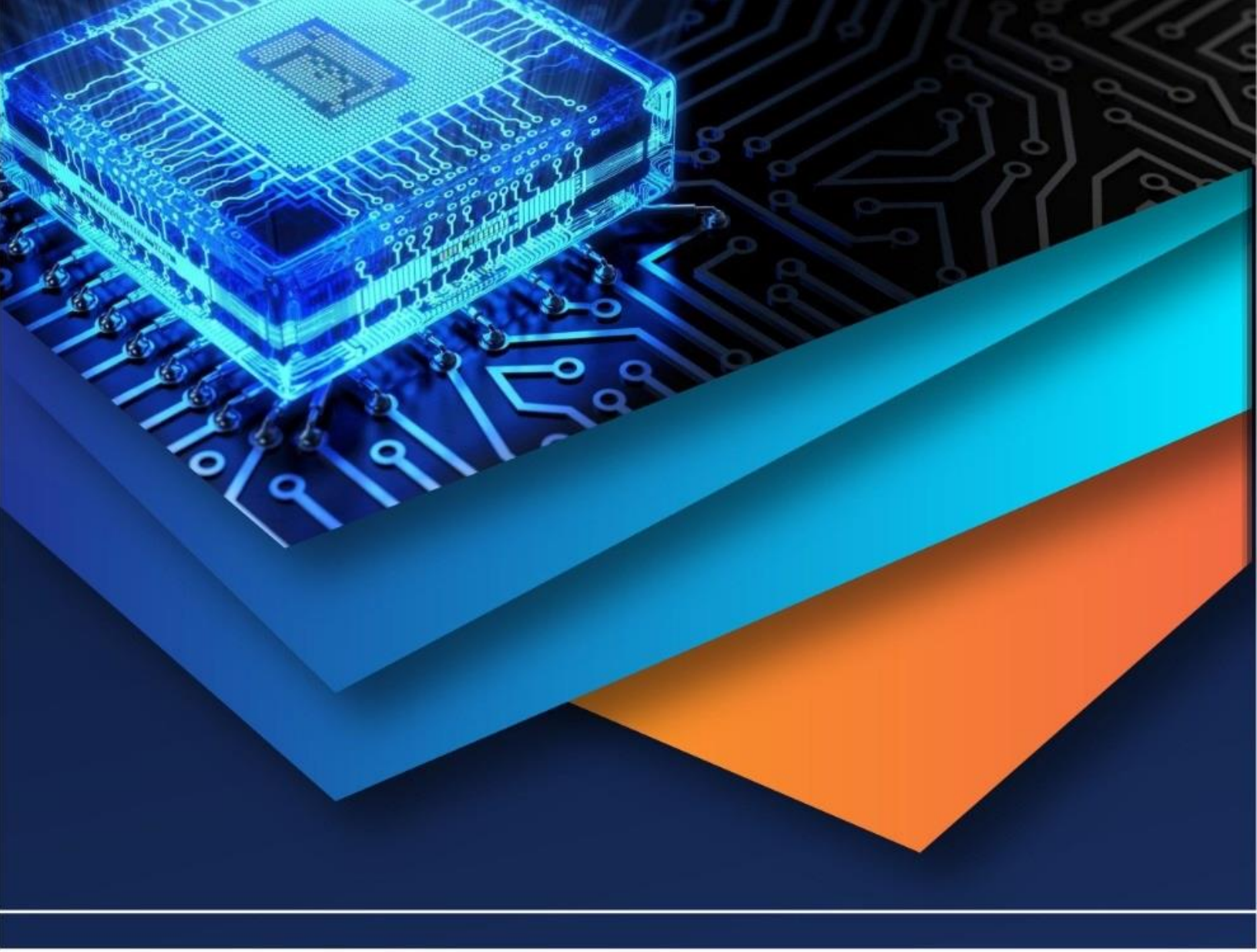
[2] Veena Bansal and R.M.K. Sinha, "Segmentation of touching and fused Devanagari character", Pattern Recognition, Vol. 35, Issue 4, pp. 875-893, 2002.

[3] Namita Dwivedi **,** Kamal Srivastava and Neelam Arya, "Sanskrit Word Recognition Using Prewitt's Operator And Support Vector Classification", IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology(ICECCN 2013)", pp. 265-269,2013.

[4] Olivind Due Tier, Anil K Jain, Torfin Tax, "Feature Extraction Method For Character Recognition: A Survey", Pattern Recognition Vol. 29, No. 4, pp. 641-662, 1996.

[5] Toru Wakaharaa, YukihikoYamashitab, "*k*-NN classification of handwritten characters via accelerated GAT correlation", ELSEVIER, Pattern Recognition**,** Vol. 47, Issue 3, pp. 994–1001**,** March 2014.

[6] S. Arora, D.Bhattacharjee, M. Nasipuri, D. K. Basu & M. Kundu, "Recognition of Non-Compound Handwritten Devanagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue 1.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY