HANDWRITTEN GUJARATI CHARACTER RECOGNITION BASED ON DISCRETE COSINE TRANSFORM

[1]MADHUREE ARDESHANA, [2]ANKIT K. SHARMA, [3]DIPAK M. ADHYARU, [4]TANISH H. ZAVERI

[1,2,3,4]Institute of Technology, Nirma University, Ahmedabad, Gujarat, India
E-mail: [1]ardeshana.g27@gmail.com, [2]ankit.sharma@nirmauni.ac.in, [3]dipak.adhyaru@nirmauni.ac.in, [4]ztanish@nirmauni.ac.in

Abstract— Gujarati script is used by more than 600 million people around the world. Handwritten character recognition of Gujarati script is important area of research due to its wider applications. The work done for handwritten Gujarati character recognition is negligible in literature. Here we have suggest a method based on Discrete Cosine Transform (DCT) as feature extraction method and Naïve Bayes (NB) classifier for classification of characters. For experiment database used here is around 22000 samples. The result obtained from above setup gives good accuracy.

Keywords—Handwritten Gujarati characters, Discrete cosine transform, Naïve Bayes classifier

I. INTRODUCTION

Optical Character Recognition (OCR) system can convert scanned document into text format for storage, editing, transmission, indexing and coordinating into different applications. Advancement of OCR innovations of Indian content has more challenges than western content due to the perplexing character set and also presence of joint characters and modifiers [1]. As Gujarat is one of the eminent state of India, Gujarati is a well-known and culturally rich dialect. Therefore improvement in effective OCR of Gujarati dialect will contribute more. Gujarati script is obtained from antique Devanagari script and it has much likeness with other north Indian dialect, basically Hindi. The significant distinction between other north Indian dialect and Gujarati is the nonappearance of a head line going through every characters framing the word [1]. Because of its unconventional characteristics Gujarati should to be treated by different approach from other Indo-Aryan dialects. Gujarati Character Recognitions offers more difficulties like most other Indian scripts relative to the western languages due to these reasons: (a) presence of joint characters, (b) presence of similar looking characters, (c) huge and complex character set by combination of characters and modifiers [2]. Set of Gujarati characters in handwritten form is shown in fig. 1.1 which includes basic consonants and vowels of Gujarati language.

For Gujarati character recognition very less significant work is available in the literature. These are also restricted to finite number of images for recognition [3]. For researchers Handwritten character recognition has turned into an extremely fascinating topic since most recent couple of decades as it is very challenging area due to variations in writing style.We have performed the experiments for large set of character images and provided a significant outcome for handwritten Gujarati character recognition. This paper depicts the outcomes acquired by using the Naïve Bayes classifier. The Discrete Cosine Transform features are used here. We believe that the work presented here will be useful for the further development of OCR system of Gujarati script having high accuracy.

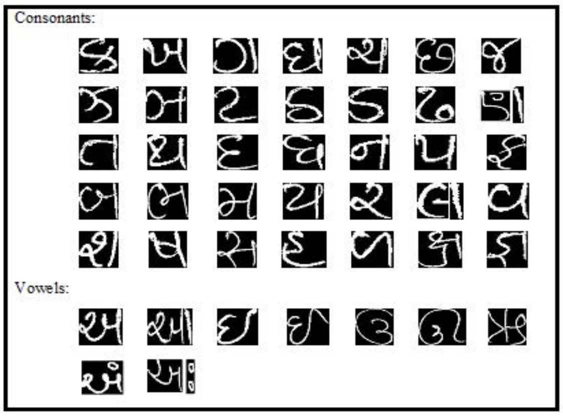

Fig. 1 Set of handwritten Gujarati characters

This paper describes all important stages of OCR system. Figure 2 shows the general architecture of OCR model. For any OCR system, feature extraction and classification stages are important. The rest of the paper is organized as below. Section II and Section III discuss about preprocessing and segmentation stage respectively. Dataset preparation and information about it is discussed in section IV. Feature extraction and experiment result are discussed in section V and VI respectively.

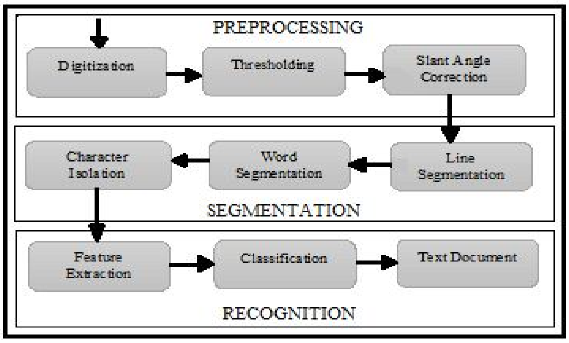

Fig. 2 OCR architecture

II. PREPROCESSING

The handwritten characters in crude form will be subjected to various preprocessing stages to make it usable regardless how it is gained. The preprocessing stage aims to produce the data relevant for the recognition stage. The principle goals of pre-processing are:

(1) Binarization:-
For any grayscale captured image every pixel varies between 0 to 255 value. With binarization process greyscale value is thresholded and it is being converted into black('1')for foreground or white('0') for background pixel. The method used for binarization of any image is called thresholding in which gray scale image is converted into binary by a finite threshold value. Thresholding is divided into two groups which are local and global thresholding. Adaptive or local thresholding utilizes different threshold values for every pixel based on local area information. Whereas in global thresholding , for entire image one threshold value is utilized which is estimated by intensity histogram. Local thresholding is generally useful for images having different level of intensities, for example, images from satellite cameras. Whreaeas global thresholding would be sufficient for simple images such as handwriting on white background [10].

(2) Noise Reduction:-
Any scanned document may have noises for example, holes in lines, disconnected segments etc. for improving quality of the document noise reduction process is performed.[11] There are more than hundreds of techniques available for noise reduction these are mainly classified into three groups: filtering, noise modeling and morphological operations. Filters can be intended for sharpening, smoothing, contrast adjustment etc. Different morphological operations can be intended for thinning the characters, extracting boundaries, smoothing the contour etc [10].

(3) Normalization:-
Through normalization process images with random size is converted into standard size. Here, normalization in size maintain inter class invariation of characters. A couple of methods are available for size normalization for example, bi-cubic interpolation, bilinear etc [12].

(4) Skew Correction:-
This process is carried out for aligning the scanned document with coordinate system of the scanner. Generally correlation, projection profiles etc methods are used for skew detection [11].

(5) Slant Removal:-
Writing style varies from person to person therefore the slant of handwritten characters will be different. Slant angle between a vertical direction and longest stroke in a word is one of the quantifiable factor of various handwriting styles. Slant removal is used for normalization of each characters in a standard form [11].

III. SEGMENTATION

Segmentation is one of the most imperative part of preprocessing step. It permits the classifier to extract information from every characters. For handwritten character recognition this process is very crucial as characters have a tendency to be joined with one another, distorted or overlapped. For complex and joint strings advanced methods must be utilized. Segmentation is carried out for breaking a single text line from scanned documents, single word from single lines, and single character from the single word. Segmentation can be classified majorly in two classes [11]:

a) External segmentation, which is performed for isolation of paragraphs, single lines or words.
b) Internal segmentation, which is performed for isolation of single characters.

There are many techniques available for segmentation of individual characters which are based on projection profiles, connected component labeling or white space and pitch [12].

IV. DATASET

Remarkable research works are available on printed Gujarati script. However, a very few work is available on handwritten Gujarati script. For research work on handwritten Gujarati characters needs proper standard or benchmark databases. So our research work also incorporates to develop a large anddelegate sample databases for handwritten Gujarati script. Database was gathered by filling an application form by distinctive gatherings of the students of different age. Application forms were gathered and isolated handwritten characters were extracted from the details mentioned in application forms by students. All these application forms were scanned through HP Scanjet 3600 multipage scanner at 300 dpi resolutions. For handwritten Gujarati characters more than 22000 character samples were extracted from these forms. These individual Gujarati character samples were saved in JPG format. Extraction of isolated characters from the scanned documents manually takes lot of time so for saving time by software isolation of characters is done. Segmentation of line and word and characters are done. Thus these database in not generated in laboratory environments. The database is equally distributed among different classes for correlation purposes. Our database comprises of 22000 samples and total 44 classes. This Database of each class is also divided in training part and testing part in ratio of

5:1. So The entire arrangement of available data of each individual class have been part into a training set a test set. The samples are stored in grayscale images. So that the researchers can explore different techniques of preprocessing, thresholding etc.

V. FEATURE EXTRACTION

For increasing the recognition rate of classifier, unique features are computed from each individual characters. Feature extraction method transforms the input data into the set of features called feature vector which is a reduced representation of the input data [8]. Feature extraction is very crucial stage. Features are extracted after preprocessing and segmentation stage. Here we have used Discrete Cosine Transform for extracting features from input data. The DCT is a linear transformation which expresses a sequence of finitely many data points which is a sum of cosine functions that oscillates at different frequencies which preserve the most useful features. DCT transforms an n-dimensional vector into the set of n coefficients. This transformation is spatial domain to the frequency domain. It has only real transform domain coefficients and includes only positive frequencies In an image, the lower frequencies stores most of energy so by transforming an image into frequency domain, we are able to reduction in data that describe the image with sufficient quality.For an image, two-dimensional(2-D) DCT need to be used as the input data are two- dimensional. The 2-D DCT can be derived from one-dimensional (1-D) DCT because each dimension of an image can be handled separately. For that A one-dimensional DCT is computed by the length of the columns and then along with the rows or vice versa which means operation is applied 1D DCT horizontally to rows and then apply 1D DCT vertically to resultant horizontal DCT.With DCT, each character image is represented as one vector. DCT can also convert the energy of an image into a few coefficients. By applying DCT on the character image with size *16x16*, 256 DCT coefficients of the image are obtained. The number of Extensive experiment was performed on the whole database of 22,000 character images.

VI. EXPERIMENT RESULTS

The features obtained here is DCT and it is utilized by the classifier. Naïve Bayes classifier is used here for classification. The database used for experimentation is too large from any other approaches done before for the handwritten Gujarati character recognition.There are 500 images for each characters and experiment is performed on 22000 samples. Overall accuracy for proposed scheme is found 78.05% from Naïve Bayes classifier. From the result of the NB classifier, we can conclude that even though classifier used here is simple it delivers good recognition rate with DCT features.

CONCLUSIONS

In this paper we have implemented a NB classifier for Gujarati Handwritten character recognition system. The features used for recognition is DCT which is easy to obtain. Here we have obtained good recognition rate of 78.05 % for large dataset.We hope that this work will also be useful for the research work of optical character recognition for other Indian scripts.

ACKNOWLEDGMENTS

Authors acknowledge the Institute of Technology, Nirma University for providing the research facilities.

REFERENCES

[1] MukeshGoswami and SumanMitra, P. Maji et al," Structural Feature Based Classification of Printed Gujarati Characters " , In Springer-Verlag Berlin Heidelberg, LNCS 8251, 2013.

[2] MandarChaudhary, GitamShikkenawis, Suman K. Mitra, MukeshGoswami, "Similar looking Gujarati printed character recognition using Locality Preserving Projection and Artificial Neural Networks" ,In third International Conference on Emerging Applications of Information Technology (EAIT), IEEE , 978-1-4673-1827-3/12 2012.

[3] JigneshDholakia, ArchitYajnik, AtulNegi," Wavelet Feature Based Confusion Character Sets for Gujarati Script",In International Conference on Computational Intelligence and Multimedia Applications, IEEE ,0-7695-3050-8/07, 2007.

[4] Desai AA , "Gujarati handwritten numeral optical character recognition through neural network". Pattern Recognition, 2582–2589, 2010.

[5] Patel CN, Desai AA , "Segmentation of text lines into words for Gujarati handwritten text. In: Proceedings of international conference on signal and image processing", In IEEE Xplore,2010

[6] Patel CN, Desai AA, "Zone identification for Gujarati handwritten words. In: Proceedings of international conference on emerging applications of information technology", In IEEE Xplore, EAIT 2011

[7] Patel CN, Desai AA, "Gujarati handwritten character recognition using hybrid method based on binary tree-classifier and k-nearest neighbour" , In J Eng Res Technol , 2013.

[8] Lipi Shah, Ripal Patel, Shreyal Patel, Jay Maniar," Handwritten Character Recognition using RadialHistogram",In*International Journal of Research in Advent Technology, E-ISSN: 2321-9637,2014.*

[9] Hetal R. Thaker, C. K. Kumbharana," Structural Feature Extraction to recognize some of the Offline Isolated Handwritten Gujarati Characters using Decision Tree Classifier",In International Journal of Computer Applications, 2014.

[10] AzizahSuliman, Mohd. NasirSulaiman, Mohamed Othman, RahmitaWirza ," Chain Coding and Pre Processing Stages of Handwritten Character Image File".

[11] Gaurav Y. Tawde , Mrs. Jayashree M. Kundargi , "An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting".

[12] Dr. Jangala. SasiKiran, N. VijayaKumar , N. SashiPrabha, M. Kavya," A Literature Survey on Digital Image Processing Techniques in Character Recognition of Indian Languages".

[13] IsraaHadiAli , "New Method for Image Features Extracting Based on Enhanced Chain Code-",In ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY,2013.

[14] Satish Kumar, “Performance Comparison of Features on Devanagari Hand-printed Dataset”,In International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[15] Ian H. Witten and Eibe Frank, Data Mining, “Practical Machine Learning Tools and Techniques”,In Morgan Kaufmann Publication, 2005.

[16] Jiawei Han and MichelineKamber,” Data Mining: Concepts and Techniques”,In Morgan Kaufmann Publication, 2001.

[17] Mr. Mukesh M. Goswami, Mr. Harshad B. Prajapati, Mr. Vipul K. Dabhi,” Classification of Printed Gujarati Characters using SOM based K-Nearest Neighbor Classifier”,In International Conference on Image Information Processing (ICIIP) Proceedings of the 2011 International Conference on Image Information Processing (ICIIP), 978-1-61284-861-7/11, 2011.

★★★