LBG Vector Quantization for Recognition of Handwritten Marathi Barakhadi

Swapnil Shinde Mrs. Vanita Mane

Abstract— Handwritten character recognition has been studied a lot in the past and involves various problems due to many reasons. In this paper, novel method of Handwritten Marathi Barakhadi Character Recognition with Shape and Texture features has been proposed. The Shape features and Texture feature are more unique, so a novel technique based on combination of these is derived and proposed here. For extracting shape features standard gradient operator such as Robert, Prewitt, Sobel, Canny and Laplace are used and vector quantization technique. The gradient mask images of the character images are obtained and then LBG vector quantization algorithm is applied on these gradient images to get the codebooks of various sizes. These obtained codebooks are considered as shape texture feature vectors for handwritten character recognition. In all 45 variations of the character recognition method are proposed using five gradient operators and 9 code book sizes (from 4 to 1024).The database consists of 2100 images which consists of 35 consonants barakhadi written by 5 different people. The crossover point of precision and recall is considered as performance comparison criteria for proposed character recognition technique.

Index Terms—Canny,Edge detection, KEVR, Laplace ,Prewitt, Sobel, Robert, VQ.

1 Introduction

Character recognition is the most widely used area which covers both machine generated and human generated characters for recognition. The research on Character recognition shows that the limitations of the methodology applied is based on two major conditions 1) the data acquisition process(on-line or off-line) and 2) the type of text(machine generated or handwritten) [18].

In general there are five major steps performed in character recognition [18] as

1. pre-processing;
2. segmentation;
3. representation;
4. training and recognition;
5. post processing

On-line and off-line handwritten have different approaches but they share a lot of common problems and solutions [19]. The handwritten character recognition is more complex as it involves hardware and different people have different style of writing. Handwritten character recognition is a technique of a system to receive and interpret handwritten input from sources such as paper, touch screen, images and other sources. Offline handwritten character recognition is method to convert text in an image into letter codes which are usable by machine and various processing applications. Marathi barakhadi involves 36 consonants and 12 vowels. This makes the problem more complex as there will be class for each consonant and separate class for problem domain can be reduced by following two steps as character extraction and character recognition. Character extraction involves scanning the document and using the image to extract the characters present in the document image. Problem arises when we are dealing with connected characters as it recognizes two characters as single one. Character recognition using several different techniques like neural networks, feature extraction. Feature extraction is determining the important properties and using them for recognition of the character. Some of properties used in feature extraction are aspect ratio, number of strokes, average distance from image center, percent of pixels above half point etc.

Optical Character recognition (OCR) is a technology that allows machines to automatically recognize the characters through an optical mechanism [1]. OCR is an instance of off-line character recognition which recognizes fixed shape static character and online character recognition recognizes dynamic motion during writing. The scanned image of handwritten text, characters is converted to machine encoded format with the help of OCR [1]. OCR has its applications in pattern recognition, artificial intelligence, and computer vision. The term OCR can also used to include preprocessing steps such as binarization, skew correction, text block segmentation prior to recognition [2]. The OCR is used for recognition of many languages all over the world such as Hindi, Kannada, Chinese, Japanese, Korean, Bangla, Konkani ,Latin etc. [2], [17]. Many challenges remain even after employing scanning methods, preprocessing techniques, cutting-edge techniques for character recognition [2].

The main challenge in online handwritten character recognition is to distinguish between different strokes used for writing and the variation in the characters that are somewhat similar. Distinguishing between few of the Devanagari characters is time consuming and complex and also may not give exact results. Many models have been proposed for online handwritten character recognition using different approaches and algorithms. Some of the models are structure based models [22], motor models [21], stochastic models [19] and learning based models [19]. Learning based is used widely for pattern recognition and statistical structure based model are used for Chinese character recognition. The structure of character is represented by the joint distribution of the component strokes. Another statistical–structural character modeling is proposed based on the Markov Random Fields (MRF) for Chinese characters [23]. Neural network based models achieve better performance than other models.

2 Literature survey

A lot of research work has been done in recognition of devnagari characters , offline and online are the medium used for the same. The first research work was presented in 1977 and since then many new and advanced techniques have been proposed and implemented. Each technique works for achieving a common goal of recognizing the characters to its maximum possibility. Some of the techniques will be discussed here and a brief overview in form of table will be presented for the same. Recognition mainly depends on the features that are extracted by various methods and which give a lot of information in terms of many factors. The problems related to recognition were the stroke of writing, angle, noise and many other external factors. Some of the features used for recognition were the shape features, texture features , shadow features, aspect ratio, gradient features etc. N Sharma et al.[12]proposed a system where features were extracted from directional chain codes and then they were given to the quadratic classifier for classification. Sushma Shelke et al.[13] designed a multi stage compound character recognition scheme using neural network and Wavelet features. Recognition of Non-Compound characters using combination of MLP and Minimum edit distance was proposed by S. Arora.et al.[14]. S. B. Patil et al.[15] describes a complete system for recognition of isolated handwritten Devnagari characters using Fourier Descriptor and Hidden-Markov model(HMM). The paper by K.Y. Rajput et al.[16] presents a system for recognizing handwritten Devnagari characters by taking handwritten images as input and separate lines , words and then characters step by step, then recognize the character by using artificial neural network approach. Handwritten Devnagari Character Recognition Using Gradient Features by Ashutosh Aggarwal et al.[17] presents a novel method of feature extraction for recognition of single isolated Devnagari Character images. Analysis and study of all the above papers gives a chance to use the other gradient operators to extract the features and combine it with vector quantization. Vector quantization is a codebook generation technique which compresses the feature vectors of fixed size into various codebooks of different sizes.

3 Vector quantization

This is a classical quantization technique used for data compression. It works by dividing large set of points into small groups (vectors) having same number of points closest to them. The density matching property is useful for identifying large and high dimensional data.

VQ has been very popular in variety of research fields such as video based event detection, data compression, image segmentation, face recognition, data hiding etc. This is also called as block quantization or pattern matching quantization that works by encoding values from multidimensional vector space into a finite set of values from discrete sub-space.The multidimensional integration was a problem for VQ but an algorithm was proposed by Linde, Buzo, and Gray based on the training sequence called as LBG which solved the above problem. A VQ designed using this algorithm is referred as LBG-VQ [5]. VQ can be divided into three procedures codebook design procedure, image encoding procedure and image decoding procedure[5]. The LBG VQ design algorithm is an iterative algorithm which requires an initial codebook C. This initial codebook is obtained by the splitting method. In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two code vectors are splitted into four and the process is repeated until the desired number of code vectors is obtained. [6].

Algorithm for LBG

```
Step 1:Divide the image into non overlapping blocks and
       convert each block to vectors thus forming a training
       vector set.
Step 2: initialize i=1;
Step 3:Compute the centroid (code vector) of this training
       vector set.
Step 4:Add and subtract constant error ei i.e. 1 and generate
       two vector v1 and v2.
Step 5:Compute Euclidean distance between all the training
       vectors belonging to this cluster and the vectors v1
       and v2 and split the cluster into two.
Step 6:Compute the centroid (code vector) for clusters obtained
       in the above step 5.
Step 7:increment i by one and repeat step 4 to step 6 for each
       code vector.
Step 8:Repeat the Step 3 to Step 7 till codebook of desired size
       is obtained.
```

4 Edge detection technique

Detection of edge is a necessary preprocessing step in computer vision and image understanding systems[16]. Edge detection is the process of identifying and locating sharp discontinuities in an image [4], [13]. The discontinuities are the abrupt changes in the pixel intensity at the boundaries. The geometry of the operator determines a characteristic direction in which it is most sensitive to edges. Operators can be optimized to look for horizontal, vertical, or diagonal edges [3]. The ways to perform edge detection can be grouped into two categories gradient based and laplacian based. The gradient based detects edges by looking for the maximum and minimum in the first derivative of the image [4] [15].The Laplacian based method searches for the zero crossings in the second order derivative of the image to find the edges [4]. The edge

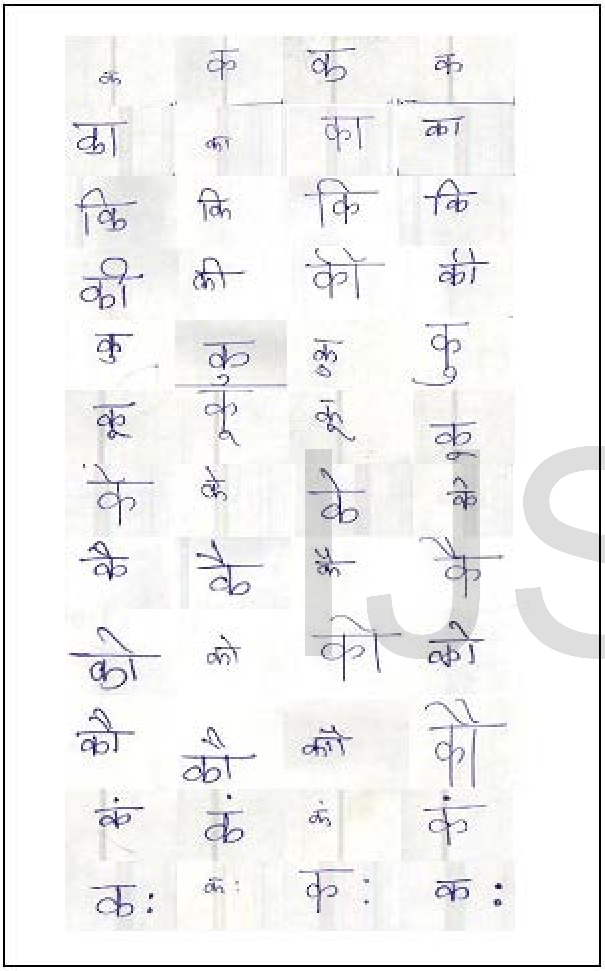
- Swapnil Ramesh Shinde,Currently pursuing ME Computer Science from Mumbai University,India,Email:swapnil.rshinde87@gmail.com
- Vanita Mane, ME Computer Science from Mumbai University,India

detection operators give information about the gradient of the edges. The various gradient operators used for edge detection are Roberts, Prewitt, Sobel, Canny, Laplace, FreiChen, and Kirsch [6].

5 Database generation

The proposed Handwritten Devnagari Character Recognition technique uses various edge detection masks followed by LBG algorithm of Vector Quatization, are implemented on MATLAB 7.10.0 on Intel Core 2 Duo 3GB RAM processor. The results are tested on Handwritten Devnagari Character image database of 2100 images from 5 samples per character with 35 different characters and their barakhadi. Sample database is shown in figure 1.

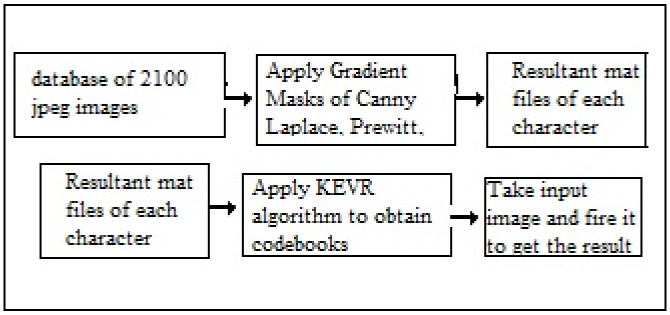
Fig. 1. Sample Handwritten Database



6 Proposed system

The proposed system involves first collecting samples from different persons to generate the database. The database will consist of 35 consonants with their barakhadi written by 5 different people so in all we have a large dataset of 2100 character images. The Gradient operators are then applied over the database to generate mat files containing feature values of each character for each of the operators. These mat files are loaded into KEVR algorithm to generate codebooks of various sizes. There will be 9 codebooks for each operator varying in size from 4 to 1024. In all 45 codebooks will be generated considering we are using 5 operators. The steps for the proposed system shown below.

Fig.2.Proposed System Block Diagram


database of 2100 jpeg images
Apply Gradient Masks of Canny Laplace, Prewitt,
Resultant mat files of each character
Resultant mat files of each character
Apply KEVR algorithm to obtain codebooks
Take input image and fire it to get the result

The feature vectors are stored in the codebooks that are gen erated by applying vector quantization algorithms. These feature vectors are used to compare with the input image when the image is taken for recognition.

7 Conclusion

The vector quantization is a clustering algorithm which involves compression of feature vectors resulting in codebooks which are resultant for recognition.The performance of the algorithm is estimated using two parameters Precision and Recall. This is the first time that vector quantization has been applied on characters for their recognition and will turn a new technology.The crossover point of Precision and Recall acts as a performance measure. For better performance the value of crossover point sholud be high. Codebook sizes 4x12, 8x12, 16x12, 32x12, 64x12, 128x12, 256x12, 512x12, 1024x12 are used. Precission is accuracy while recall is completeness. The average values of precission and recall are calculated and the recognition rate is estimated.

References

[1] "Character recognition" published by AIM, Pittsburgh Optical, 2000.

[2] Suryaprakash Kompalli · Srirangaraj Setlur, Venu Govindaraju,"Devanagari OCR using a recognition driven segmentation framework and stochastic language models", Springer, 2009.

[3] Djemel Ziou and Salvatore Tabbone, Report on "Edge detection Techniques-An overview", University of Canada.

[4] Raman Mani and Dr. Himanshu Aggarwal "Study and comparison of various Image edge detection techniques", International journal of Image Processing (IJIP), Volume (3): issue (1).

[5] Ms. Asmita A.Bardekar, Mr. P.A.Tijare,"Implementation of LBG algorithm for image compression",IJCTT Volume 2 Issue2,2011

[6] Dr H.B.Kekre,Dr Sudeep D. Thepade, Shrikant Sanas, Sowmya Iyer, Jhuma Garg" Shape Content Based Image Retrieval using LBG Vector Quantization" International Journal of Computer Science and Information Security. (IJCSIS)Vol. 9 No. 12 DEC 2011.

[7] A.Amali Asha S.P. Victor A. Lourdusamy "Performance of Ant System over other Convolution Masks in Extracting Edge", IJCA, 2011.

[8] Mamta Juneja, Parvinder Singh Sandhu ,"Performance evaluation of edge detection techniques for images in spatial domain".IJCTE, 2009.

[9] Lijun Ding, Ardeshir Goshtasb,"On the Canny edge detector" Pattern Recognition Society, published in Elsevier, 2000.

[10] Indra Kanta Maitra, Sanjay Nag, Samir K. Bandyopadhyay ,"A Novel Edge Detection Algorithm for Digital Mammogram",IJICTR,2012

[11] Chen Yu, Indiana University "Canny edge detection and Hough Transform".2010.

[12] Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier N. Sharma, U. Pal, F. Kimura, and S. Pal, Springer, 2006.

[13] Sushama Shelke, Shaila Apte " A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features ",International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 1, March 2011.

[14] Sandhya Arora, D. Bhattacharjee, Mita Nasipuri, "Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance", IJCSS.

[15] Sandeep B. Patil, G.R. Sinha and Kavita Thakur3, "Isolated Handwritten Devnagri Character Recognition using Fourier Descriptor and HMM ",IJPAST, 2012.

[16] K. Y. Rajput and Sangeeta Mishra,"Recognition and Editing of Devnagari Handwriting Using Neural Network", SPIT-IEEE Colloquium and International Conference, 2012.

[17] Ashutosh Aggarwal, Rajneesh Rani, RenuDhir , " Handwritten Devnagari Character Recognition using Gradient features" , IJARCSEE , Vol 2,Issue 5, May 2012.

[18] Prachi Mukherji, Priti Rege, "Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition", Journal of Pattern Recognition,2009.

[19] Nafiz Arica and Fatos T. Yarman-Vural "An Overview of Character Recognition Focused on Off-Line Handwriting", IEEE transactions, May 2001.

[20] H. Swethalakshmi1, Anitha Jayaraman, V. Srinivasa Chakravarthy, C. Chandra Sekhar "Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines", IIT Madras.

[21] In-Jung Kim and Jin-Hyung Kim "Statistical Character Structure Modeling and Its Application to Handwritten Chinese Character Recognition", IEEE transaction, Nov 2003.

[22] Lambert R.B. Schomaker & Hans-Leo Teulings "A Handwriting Recognition System Based on Properties of the Human Motor System", Nijmegen institute of cognition research and information Technology, Netherlands.

[23] Kan fai Chan and Dit yan yeung "Elastic Structural matching for recognizing on-line handwritten alpha numeric characters.", March 1998.

[24] H. B. Kekre, Tanuja K. Sarode, "New Clustering algorithm for vector quantization using rotation of error vector", International Journal of computer and Information Security, Vol .7,No 3,2010.