Isolated Kannada Character Recognition using Chain Code Features

H. Imran Khan[1], Smitha U. V[2], Suresh Kumar D. S [3]

[1]M.Tech Scholar, Department of ECE, CIT, Tumkur, India

[2]Assistant Professor, Department of ECE, CIT, Tumkur, India

[3]Professor, Department of ECE, CIT, Tumkur, India

Abstract: ***Handwriting character recognition (HCR) is the ability of a computer to receive and interpret handwritten input. Handwritten Character Recognition is one of the active and challenging research areas in the field of Pattern Recognition. Pattern recognition is a process that taking in raw data and making an action based on the category of the pattern. HCR is one of the well-known applications of pattern recognition. Handwriting recognition especially for Indian languages is still in infant stage because not much work has been done it. This paper discuss about an idea to recognize Kannada vowels using chain code features. Kannada is a South Indian language. For any recognition system, an important part is feature extraction. A proper feature extraction method can increase the recognition ratio. In this paper, a chain code based feature extraction method is investigated for developing HCR system. Chain code is working based on 4-neighborhood or 8–neighborhood methods. Chain code is a sequence of code directions of a character and connection to a starting point which is often used in image processing. In this paper, 8–neighborhood method has been implemented which allows generation of eight different codes for each character. These codes have been used as features of the character image, which have been later on used for training and testing for K-Nearest Neighbor (KNN) classifiers. The level of accuracy reached to 100%.***

Keywords: Pattern Recognition, Handwritten Character Recognition, K-Nearest Neighbor, Kannada vowels, Feature extraction, Chain code.

1. Introduction

Pattern recognition is a field of study whose general goal is the classification of objects into a number of categories. The first process of this mechanism is to design a dataset for feature extraction and another data set is for to train the classifier. In next process Chain code feature extraction method are applied on input data set and extract feature from it. This extracted feature applied on K-Nearest Neighbor classifier which was already trained trough input data set is recognized pattern based on match between feature and trained data. Pattern recognition system can be used in so many applications such as face recognition, character recognition, and speech recognition [1].

Character Recognition (CR) has been an active area of research and due to its diverse applicable environment; it continues to be a challenging research topic. A HCR is one of the highly used applications of pattern recognition techniques. Many systems and classification algorithms have been proposed in the past years on character/numeral in various languages like Support Vector Machine for Handwritten Devanagari Numeral Recognition [2], Converting Printed Kannada text Image file to Machine editable format using Database approach [3] and Handwritten Character Recognition using Correlation Coefficient [4].

The proposed system for Kannada character recognition is as shown in Figure 1.
Its different stages are as given below:

- Input: Samples are read to the system through a scanner.
- Preprocessing: Preprocessing converts the image into a form suitable for subsequent processing and feature extraction.
- Segmentation: The most basic step in CR is to segment the input image into individual glyphs. This step separates out text from image and subsequently lines and characters from text.
- Feature extraction: Extraction of features of a character forms a vital part of the recognition process. Feature extraction captures the vital details of a character.
- Classification: During classification, a character is placed in the appropriate class to which it belongs.
- Post Processing: Combining the CR techniques either in parallel or series.

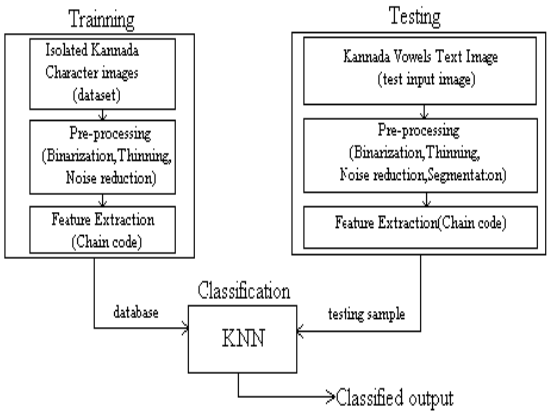

Figure 1: Block diagram of proposed system

In Pre-processing, we are going to perform Binarization, Normalization, Thinning, Noise reduction. Binarization means converting grayscale or color image to binary image.

Normalization, as the size of the character image may be varying to maintain the uniformity we are going to resize the image to 40*40 pixel size. Thinning refers to removing the extra pixels from the body (character) of the image and leaving only one pixel size character. Noise reduction refers to removing of unwanted line segments present in the character image.

In Segmentation process, we consider the processing of entire document containing multiple lines and many characters in each line. Our aim is to recognize characters from the entire document. The handwritten document has to be free from noise, skewness, etc. The lines and words have to be segmented. The characters of any word have to be free from any slant angle so that the characters can be separated for recognition. By this assumption, we try to avoid a more difficult case of cursive writing. Segmentation of unconstrained handwritten text line is difficult because of inter-line distance variability, base-line skew variability, different font size and age of document [5]. The next step is feature extraction and classification which are discussed in the section 2 and section 3.

1.1 Introduction to Kannada Script

Kannada is the official language of the southern Indian state of Karnataka. Kannada is a Dravidian language spoken by about 44 million people in the Indian states of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. The Kannada alphabets were developed from the Kadamba and Calukya scripts, descendents of Brahmi which were used between the 5th and 7th centuries AD. There are 13 Vowels (Swara), 2 Yogavaha and 34 Consonants (Vangana) in modern Kannada script [6]. In this paper we constrain ourselves to recognition of handwritten Kannada vowels. Printed Kannada vowels and their corresponding handwritten vowel samples are shown in Figure 2 and Figure 3, to get an idea about the shape difference between printed and handwritten samples.

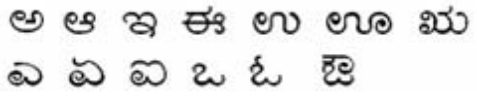

Figure 2: Sample of printed Kannada Vowels

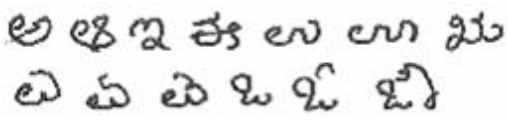

Figure 3: Sample of handwritten Kannada Vowels

2. Feature Extraction

Feature Extraction is the process by which certain features of interest within an image are detected and represented for further processing. Any given image can be decomposed into several features. The term 'feature' refers to similar characteristics. Therefore, the main objective of a feature extraction technique is to accurately retrieve these features.

2.1 Chain code Generation

Chain code is a technique which represents the boundary of an object with a sequence of codes where those codes represent the direction of where is the location of the next pixel. The chain code extraction algorithm of proposed system is shown in Figure 4. Basically chain code is working based on two different manners such as 4- neighborhood method or 8-neighborhood method. In this work, we have implemented 8-neighborhood method for chain code. This method has been implemented as feature for Kannada handwritten characters classification. In order to obtain the chain code, we just focus on the main part (body) of the character image. From the first pixel of the image, we move downwards row by row and consider the first pixel of the body of image which exactly has got one neighbor, as the start point of the chain code. If a character has no initial point, we will consider its chain code as zero. Each pixel of image has received its eight neighbors; to each neighbor we assign one value between 0 and 7. After finding the start point of the chain code neighbors of a pixel and value is assigned to it in a given character image, we move to the next neighbor pixel which also be a part of image. Again in the cases of having two or more neighbor pixels with the above condition use the directional priority. While passing one pixel to its next pixel, we have inserted the number related to that next in the chain code of that image. After receiving the chain code for all segmented characters, we come to know that the chain code for different characters has different size, and size of each chain code depends on the length of the desired character. In more ever, length of the chain codes is usually high; therefore one should convert it into it's normalize form. This chain code as its length will be fixed and limited [7].

Algorithm for generating chain code considering 8-neighborhood is as follows:

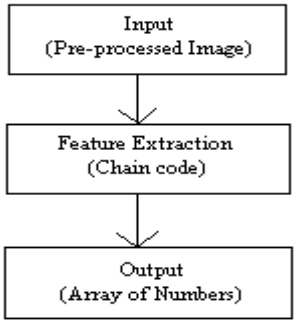

Figure 4: Chain code extraction for proposed system

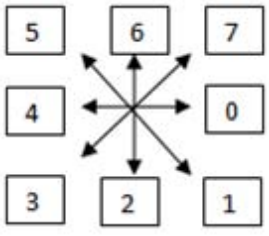

Figure 5: 8-neighborhood for chain code

Step1: Find out starting point which has nonzero values and store it in first
Step2: Initialize 0-7 total eight directions

Step3: Travels all 8 neighbors
Step4: Find first nonzero value
Step5: Add it in to chain code list
Step6: Move to next position
Step7: Check whether we reach to first point or not if not then go to step 3.

2.2 Chain Code Normalization

After applying chain code on segmented image, we came to know that the length of chain code vary for different character. So in this case it is very difficult to keep all the chain code for all the characters. In order to normalize the obtained chain code, we transform it to a two dimensional matrix where in the first row, the value of the chain code, and in the second row, frequency of occurrence of that value are written [7]. This frequency of occurrences of different neighbors can also be considered as histogram of neighboring indexes. The example of chain code generation and normalization is as shown in Figure 6. It should be noted that even though we obtain variable number of neighborhood indices, due to histogram consideration, the number of features for each character becomes fixed which 8 for 8-neighbourhood method is.

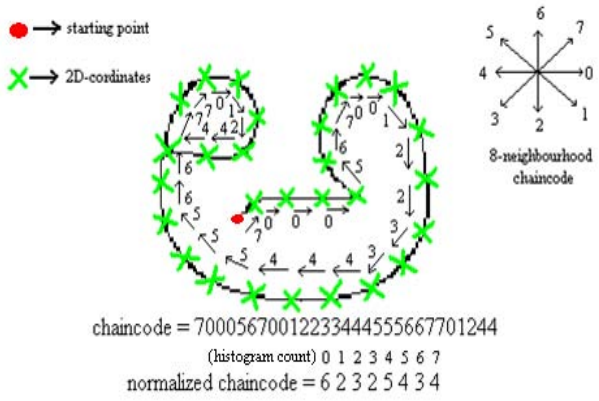

Figure 6: Example of Feature Extraction

3. Classification

K-Nearest Neighbour (K-NN)

K-NN is an uncomplicated classification model that employs lazy learning. It is a supervised learning algorithm by classifying the new instances query based on majority of k-nearest neighbor category. Minimum distance between query instance and each of the training set is calculated to determine the *k*-NN category. Each query instance (test character) will be compared against each of training instance (training character). The *k*-NN prediction of the query instance is determined based on majority voting of the nearest neighbor category. Since query instance (test character) will compare against all training characters, *k*-NN encounters high response time [8]. In this works, for each test character (to be classified), minimum distance from the test character to each of the training characters in the training set is calculated to locate the *k*-NN category of the training data set. A Euclidean Distance measure is used to calculate how close each member of the training set is to the test class that is being examined. Euclidean Distance measuring:

$$d_E(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2}$$

For each test character (to be classified), the training data set is located with *k* closest members (*k*-nearest neighbors). From this *k* nearest neighbor, class labels either repetitions or prolongations are found and class labels of test character are determined by applying majority voting.

4. Experiment and Results

In this work the chain code method as discussed in section 2 was implemented in Mat lab environment. The extracted chain code was normalized and used as features with geometric features for KNN classifier. Implementation Results of KNN & Chain code based character recognition. In the implementation part, we have divided these 1625 features in group of 13 with each group contains 125 features of Kannada vowels respectively. This matrix is used as a feature to train K-NN for training purpose. At the other side for testing purpose, we have taken test image. Pre-processing, segmentation is applied on test image. Same feature matrix is prepared for all the segmented characters and applied these features to K-NN for training and recognition. Here, we have used K-NN classifier for the purpose of recognition. The overall accuracy of 100% was obtained for the test data using KNN.

The chain code method is used to recognize test image shown in Figure 7. Segmentation is done on test image as shown in Figure 8, while the final output of K-NN classifier is shown in Figure 9.

Figure 7: test image

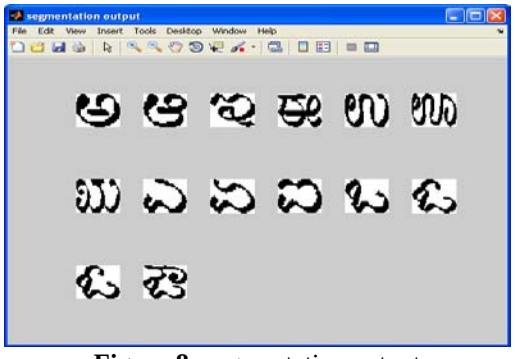

Figure 8: segmentation output

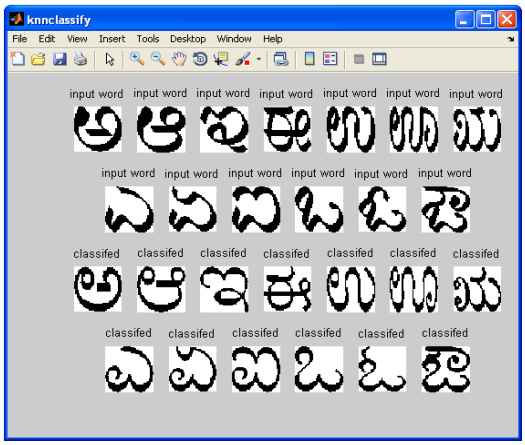

Figure 9: K-NN classifier output

It can be seen that all the Kannada vowels in the test image were recognized properly.

5. Conclusion

A simple and an efficient Kannada handwritten character recognition system using chain code features are investigated. Selection of feature extraction method is most important factor for achieving high recognition ratio. In this work, we have implemented chain code based on 8-neighborhood feature extraction method. With the use of this obtained feature, we have trained the KNN classifier for character recognition. In the investigated work the recognition rate is 100%.

References

[1] Ravi Sheth, N C Chauhan, Mahesh M Goyani, Kinjal
[2] A Mehta, “Handwritten Character Recognition System using Chain code and Correlation Coefficient”, International Conference on Recent Trends in Information Technology and Computer Science (IRCTITCS) pp. 31-36, 2011
[3] Shailedra Kumar and Shrivastava Sanjay S. Gharde, “Support Vector Machine for Handwritten Devanagari Numeral Recognition”, International Journal of Computer Applications (0975 – 8887) Volume 7– No.11, pp. 9-14, October 2010.
[4] B.M. Sagar, Dr. Shobha G, Dr. Ramakanth Kumar P, “Converting Printed Kannada Text Image File To Machine Editable Format Using Database Approach”, International Journal Of Computers Issue 2, Volume 2, pp. 172-175, 2008.
[5] Ravi Sheth, N C Chauhan, Mahesh M Goyani, “Handwritten Character Recognition Systems using Correlation Coefficient”, selected in International Conference V V P Rajkot, 8-9 April 2011.
[6] Yi-Kai Chen and Jhing-Fa Wang, “Segmentation of Single-or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis”, IEEE PAMI vol.22, 1304-1317, 2000.
[7] Sangame S.K, Ramteke R.J, Rajkumar Benne, “Recognition of isolated handwritten Kannada vowels”, Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, pp-52-55, 2009
[8] Ravi Sheth, N C Chauhan, Mahesh M Goyani, Kinjal
[9] A Mehta, “Handwritten Character Recognition System using Chain code and Correlation Coefficient”, International Conference on Recent Trends in Information Technology and Computer Science (IRCTITCS) pp. 31-36, 2011
[10] Lim Sin Chee, Ooi Chia Ai, and Sazali Yaacob "Automatic Detection of prolongation and Repetions using LPCC", in Research and Development (SCOReD), 2009.
[11] http://en.wikipedia.org/wiki/Knearest_neghbors_algorithm

Author Profile

H. Imran Khan, M. Tech final year student of Department of Electronics. He is pursuing his Master’s Degree in Electronics from VTU. He has completed his B.E from VTU.

Mrs. Smitha U.V: Assistant Professor of Department of Electronics and Communication Engineering, CIT Tumkur. She has completed her B.E and M.Tech from VTU.

Dr. Suresh Kumar D.S, Prof. Suresh has completed B.E (IT), M.Tech, MBA and PhD. His area of research is in the field of embedded systems. He authored a few Text books and has published papers in many national/international journals & conferences. He has guided many projects sponsored by KSCST and has recently filed patent of his project on “Health Monitoring System”.