Review of Automatic Handwritten Kannada Character Recognition Technique

[1]sharana Basappa, [2] Dr. Jitendra Sheetlani

[1]Department of Computer Science
SSSUTMS, Sehore

Abstract— Handwriting recognition has been one of the dynamic and testing research zones in the field of example recognition. It has various applications which incorporate, perusing help for dazzle, bank checks and transformation of any manually written record into auxiliary content frame [1]. As there are no adequate number of deals with Indian dialect character recognition particularly Kannada content among 15 noteworthy contents in India [2]. A manually written kannada character is resized into 60x40 pixel. The resized character is utilized for preparing the neural network. Once the preparation procedure is finished a similar character is given as contribution to the neural network with various arrangement of neurons in concealed layer and their recognition precision rate for various kannada characters has been ascertained and thought about. The outcomes demonstrate that the proposed framework yields great recognition exactness rates practically identical to that of other manually written character recognition frameworks.

Keywords: Kannada character, Pattern recognition,Image Processing

I. INTRODUCTION

The programmed preparing of these structures comprises of gathering the data put away in the structures and changing over it into an electronic arrangement (machine-comprehensible). The recognition of manually written characters is hugely imperative to the advance of the robotization procedure. This is delegated offline and online detection. Programmed frames preparing (AFP) utilizes the offline detection technique since it incorporates the programmed transformation of content into a picture into letter codes that can be utilized as a part of word handling and PC preparing applications [1]. AFP incorporates a total sweep of a frame with a scanner. The checked picture experiences a few preprocessing, character segmentation, and handwriting recognition operations. India is a multilingual and multi-content nation that incorporates eighteen authority dialects, including Kannada. A few works have been done to recognize Kannada's transcribed characters. The major preprocessing ventures of AFP incorporates edge detection, morphological operations to make it reasonable for segmentation. Segmentation isolates the picture content reports into lines, words and the characters. Thungamani M and RamakanthKumar P [2] examine two segmentation systems, for example, classical approach and all encompassing approach. In classical approach, the information picture is divided into sub pictures. In all encompassing approach, the characters are perceived without analyzation. Mamatha H.R and Srikanatamurthy K [3], proposed a segmentation conspire utilizing projection profiles. Morphological operations are utilized to evacuate the clamor. After this content lines are removed utilizing even projection profile, words and characters are separated utilizing vertical projection profiles.

Kannada script has large number of character set. This may reduce the recognition accuracy and increase the computational cost. To avoid this problem an algorithm has been proposed to reduce symbol set [4], where the vowel modifiers (kagunitha) and consonant modifiers (vattakshara) which are not connected to base characters are considered as separate classes. Devanagari script has similar characteristics as Kannada script like vowel modifiers, consonant conjuncts etc,. The recognition of this script consists of three phases: segmentation, decomposition, i.e,. decomposing a composite character into base part and modifier parts; and recognition [5]. Only small subsets of compound characters (upper and lower signs) are considered for the recognition. Many Arabic letters also share common primary shapes, which differs only in the number of dots and the dots above or below the primary shape. A survey on offline recognition of Arabic handwriting recognition is presented in [6]. Different segmentation, feature extraction and recognition engines used for OCR are also discussed. An overview of character recognition methodologies with respect to the offline character recognition systems such as pre-processing, segmentation, representation, recognition and post-processing methods are presented in [7]. Shape based features such as Fourier descriptors and chain codes are used for the recognition of handwritten Kannada characters (vowels and numerals) are discussed in [8]. Support Vector Machine (SVM) is used for recognition purpose and an accuracy of 95% is obtained. A brief survey on offline recognition of Devanagari script is presented in [9]. Performance of different feature extraction techniques using different classifiers is tabulated. Gradient and PCA based features with PCA, SVM and Neural Network classifiers are found to have better recognition accuracy. Development of two databases for two popular Indian scripts Devanagari and Bangla for numeral recognition is presented in [10]. This uses a multistage cascade recognition scheme using wavelet-based multi-resolution representations and multi layer perceptron (MLP) classifiers to achieve higher recogniti accuracy. This is then used to the recognition of mixed numerals for three Indian scripts such as Devanagari, Bangla and English. Unconstrained handwritten recognition of Kannada characters using very large dataset of 200 samples using ridgelet transform is discussed in [11]. To reduce the dimension of feature vector PCA is used. It is found that ridgelet features offered promising result than PCA. A zone based method for the recognition of handwritten Kannada vowels and consonants is presented in [12]. Character image is divided into 64 non-overlapped zones and from each zone crack codes are computed. SVM is used as classifier and an accuracy of 87.24% is achieved. Literature records few papers on Kannada character

recognition. Choice of methods for feature extraction is important for achieving efficient character recognition for large classes. In this paper, Kannada handwriting recognition for automatic form processing is considered. PCA and HoG are used for feature extraction. Performances of features are compared for 57 classes.

II. REVIEW OF LITERATURE

Today, many researchers have been done to recognize Kannada characters. But the problem of interchanging data between human beings and computing machines is a challenging one. Most of the research was focused on recognition of off-line handwritten characters for Devanagari and Bangla scripts. It is observed from the literature survey that there is a lot of demand on Indian scripts character recognition and an excellent review has been done on the OCR for Indian languages [5]. A Detailed Study and Analysis of OCR Research on South Indian Scripts is presented in [6]. Rajput and Mali [7] have proposed an efficient method for recognition of isolated Devanagari handwritten numerals based on Fourier descriptors. In [8] zone centroid is computed and the image is further divided in to n equal zones. Average distance from the zone centroid to the each pixel present in the zone is computed. This procedure is repeated for all the zones present in the numeral image. Finally n such features are extracted for classification and recognition. F-ratio Based Weighted Feature Extraction for Similar Shape Character Recognition for different scripts like Arabic/Persian, Devnagari English, Bangla , Oriya, Tamil, Kannada, Telugu etc is presented in [9]. An efficient and novel method for recognition of machine printed and handwritten Kannada numerals using Crack codes and Fourier Descriptors is reported in [10]. Kannada along with other Indian language scripts shares a large number of structural features. Kannada has 49 basic characters, which are classified into three categories: swaras (vowels), vyanjans (consonants) and yogavaahas (part vowel, part consonants). The scripts also include 10 different Kannada numerals of the decimal number system.

Table I

Maximum possible combinations of Kannada characters

Character Type	V (Vowel)	C (Consonant)	CV (Kagunita)	CCV (t.Æ Ú)	CCCV (vÀì÷ å)	N	Total
Possible combinations	15	34	510	17340	589560	10	589585

III. PROPOSED METHODOLOGY

Data Collection

To the best of our knowledge standard dataset for written by hand and printed Kannada numerals isn't accessible till today. Thusly, dataset of absolutely unconstrained written by hand Kannada numerals 0 to 9 is made by gathering the manually written archives from authors having a place with various callings. The skew in the archives has not been considered. An example picture of checked record is appeared in Fig 1. The individual numerals were separated physically from the checked archives and named. The marked numerals are then pre-handled.

Kannada does not have a standard proving ground of character pictures for OCR execution assessment. To make manually written Kannada characters text style and size free is a test in the research since speculation will be more troublesome. The testing some portion of Kannada written by hand character recognition is the qualification between the comparative molded parts. A little variety between two characters or numerals prompts recognition many-sided quality and level of recognition exactness. The style of composing characters is profoundly unique and they come in different sizes and shapes. Same numeral may take diverse shapes and on the other hand at least two distinct numerals of a content may take comparative shape. A few cases of the comparative molded numerals are as appeared in figure 2.

Figure 1. A sample sheet of Kannada Handwritten numerals 0 to 9

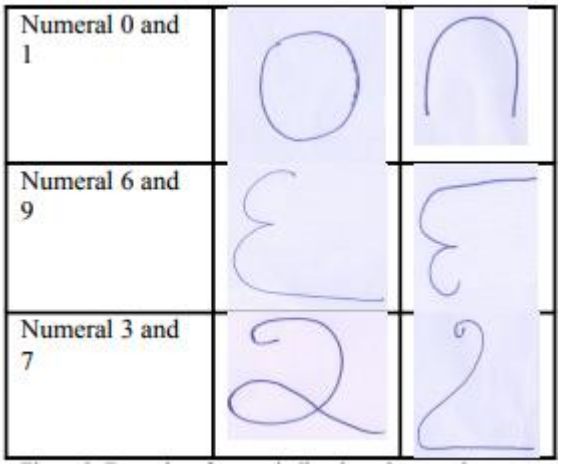

Figure 2: Examples of some similar shaped numerals

In the literature it is observed that most of the character recognition methods focuses on extracting either statistical features such as zoning, moments etc. or the structural features based on the geometry of the character. In this paper we have proposed a method, which attempts to combine both the statistical and structural features. Here we divide the entire image into 9 equal sized zones as indicated in the figure 3

In a binary image, whenever a pixel value changes from 0 to 1 or 1 to 0 it indicates the information about the edge. This information is very significant as it denotes the geometry of the character and helps in identifying the character [13]. In order to capture this information, we have used Run Length Count (RLC) technique. In the proposed method, for every zone, we find the Run Length count in horizontal and vertical direction. A total of 18 features will be extracted for each characters and this will serve as feature vector. The above method is summarized in algorithm

Algorithm
Begin
Input: a set of pre-processed sample images
Output: feature vector for the numerals
Method:
Step1: Binaries the image using a threshold value.
The threshold value for an image is fixed using the Otsu's method.
Step 2: The image is resized to 72 * 72 pixels
Step 3: Divide the image into 9 blocks as shown in the figure 4
Step 4: For each block the horizontal and vertical run length is found. Figure 5 illustrates the procedure to find the horizontal run length for a block. Similar approach can be adapted to find the vertical run length. Horizontal run length is 10 in this case Step 5: A feature vector of length 18 i.e. one horizontal and one vertical run length for each block is obtained for each of the numerals.
End

Zone 1`	Zone 2	Zone 3
Zone 4	Zone 5	Zone 6
Zone 7	Zone 8	Zone 9

Figure 3 Image Zoning

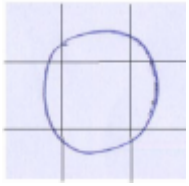

Figure 4: Image Zoning of numeral 0

0	1	1	1	0	0
0	0	0	1	1	0
1	1	1	0	0	0
0	1	1	0	0	0
1	1	1	0	0	0
1	1	0	0	0	0
0	0	0	0	1	0

Figure 5: Illustration of horizontal run length count

Over the last decade, ensemble technique is widely used in many different applications. There are different types of ensemble methods. One such type is classification fusion method. In this method, many classifiers are trained on a same feature space. The results of these classifiers are combined to obtain a more accurate classification.

The Neural Network Classifier is an efficient technique to use when the classification problem has pattern classes that display a reasonably limited degree of variability. It considers each input pattern given to it and classifies it to a certain class by calculating the distance between the input pattern and the training patterns. It takes into account only prototypes to the input pattern during classification. The decision is generally based on the majority of class values of the k nearest neighbors. In the NN classification, we compute the distance between features of the test sample and the feature of every training sample. Linear classifier is a stastical classifier, which makes a classification decision based on the value of the linear combination of the features. A linear classifier is often used in situations where the speed of classification is an issue, since it is often the fastest classifier [12]. Linear classifiers often work very well when the number of dimensions in feature vector is large.

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right), \quad (1)$$

Where wj is weight vector, learned from a set of labeled training samples. Xj is the feature vector of testing sample. f is a simple function that maps the value to the respective classes based on a certain threshold.
The classifier fusion method is summarised in the following algorithm

Algorithm
Begin
Input: a set of training samples and test sample
Output: the class label for which the test sample belongs Method:
Step1: Extract the directional chain code and run length code features for the training sample and test sample using previously discussed
Step 2: Apply the features obtained for the training samples to train the NN and Linear classifiers
Step 3: Apply the features obtained from the test sample to each of the classifier. Let the prediction of the classifiers be p1,p2,p3 and p4
Step 4: Predict the class of the test sample as Class = Majority of { p1, p2, p3, p4 }
End

Experimental results

For the experimentation, we have thought about a database with almost 600 examples. Experimentation was led utilizing directional chain code highlights and run length tally. For arrangement, we have utilized straight classifier and K-closest neighbor classifier. 80% of the information is utilized for the preparation of the classifiers. From the experimentation, we noticed that the general numeral recognition precision of the directional highlights is superior to anything the run length check highlights. However the run length tally highlight utilizes just 18 includes as contrasted and the 72 highlights of the directional highlights. For the directional highlights both the classifiers yielded same normal exactness. The exactness of the KNN and Linear classifier shifted for the run length check include. We processed exactness of the individual numerals and watched that for the directional chain code include, the least precision was acquired for the numeral 3 in the directional element strategy. The regular misclassification of numeral 3 was with numeral 7, which is fundamentally the same as fit as a fiddle. However the technique could accomplish great recognition rate for the other comparative molded numerals like 0 and 1, 6 and 9. Correspondingly for the RLC highlights, the least precision was gotten for the numeral 9.The basic misclassification of numeral 9 was with numeral 2 and 7 despite the fact that they are not of comparative shape. This might be because of the flat strokes exhibit in those numerals.

Programmed Form Processing framework includes Image obtaining utilizing scanner, pre-preparing of filtered shape, just written by hand character extraction, manually written character segmentation, character recognition and capacity as appeared in The format of the Birth testament is made with all the required information fields. The candidate is then taught to fill the shape in Kannada dialect with all the base characters in the upper box and conjuncts in the lower box. The filled shape is then examined utilizing flatbed filtering.

IV. Conclusion

Flow research isn't specifically worry to the characters, yet in addition words and expresses, and even the total records. From different investigations we have seen that choice of pertinent component extraction and characterization procedure assumes an essential part in execution of character recognition rate. This audit builds up an entire framework that believers examined pictures of manually written characters to content archives. This material fills in as a guide and refresh for perusers working in the Character Recognition territory. The edge segmentation for kannada characters with probabilistic neural network can give high proficiency than other existing Methods. The proposed approach is effective in ordering distinctive sizes of Characters and textual styles. In Future the approach can be connected to continuous recognition of Characters.

REFERENCES

[1] Afef Kacem, Asma Saidani, Abdel Belaid, "A system for an automatic reading of student information sheets", International Conference on Document Analysis and Recognition, pp 1265-1269, 2011.
[2] M. Thungamani and P. Ramakanth Kumar, "A Survey Methods and Strategies in Handwritten Kannada Character Segmentation", International Journal of Science Research, Vol 1, Issue 1, June 2012.
[3] H.R Mamatha and and K. Srikantamurthy, "Morphological operations and projection profile based segmentation of handwritten Kannada document", International Journal of Applied Information Systems (IJAIS), Vol 4, No.5, October 2012.
[4] Nethravathi B, Archana C.P, Shashikiran K, A.G Ramakrishnan and VIjay Kumar,
"Creation of huge annotated database for Tamil and Kannada OHR", 12th IEEE International Conference on Frontiers in Handwriting Recognition, Nov 2010, Pages 415-420.
[5] R M. K. Sinha and H. N. Mahabala, "Machine Recognition of Devanagai Script", IEEE Transactions on Systems, Man and Cybernetics, Vol. Smc 9, No 8, August 1979.
[6] Liana M. Lorigo and Venu Govindaraju, "Offline Arabic handwriting recognition: A survey", IEEE Transactions on Pattern Analysis and machine Intelligence, Vol 28, No.5, May 2006.
[7] Vengatesan K., and S. Selvarajan"Improved T-Cluster based scheme for combination gene scale expression data" International Conference on Radar, Communication and Computing (ICRCC), pp. 131-136. IEEE (2012).
[8] Kalaivanan M., and K. Vengatesan." Recommendation system based on statistical analysis of ranking from user. International Conference on Information Communication and Embedded Systems (ICICES), pp.479-484, IEEE, (2013).
[9] K. Vengatesan, S. Selvarajan: The performance Analysis of Microarray Data using Occurrence Clustering. International Journal of Mathematical Science and Engineering, Vol.3 (2) .pp 69-75 (2014).
[10] K Vengatesan, V Karuppuchamy, S Pragadeeswaran, A Selvaraj," FAST Clustering Algorithm for Maximizing the Feature Selection in High Dimensional Data", Volume – 4, Issue-2, International Journal of Mathematical Sciences and Engineering (IJMSE), December 2015