ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Special Issue
National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

KNN classification of Kannada Characters using Hu’s Seven Variants and Zernike Moment Features

[1] Duddela Sai Prashanth, [2] C N Panini [3]Bharath Bhushan
[1][2] Research Scholar SCSVMV University
[3] Sahyadri College, Mangalore

Abstract:- **Identifying the text is one of the promising field of research in the domain of computer vision and pattern recognition. This paper copes with identity of kannada text. Eliminating noise and extracting the textual content from the scanned or captured picture is first step. Segmenting the lines and characters is the second step which is essential. Noise removal and extracting the textual content can be done by way of the usage of any noise filter and foreground subtraction algorithm. Otsu set of rules facilitates to gain the task of foreground extraction. Horizontal and Vertical Profiling is a method of extracting lines and words from the image document. Extracting the knowledge from the dataset the use of Hu’s Seven variations and Zernike Moments features helps to come over many problem. After training method knowledge is being generated through the usage of the above mentioned methods. KNN classifier is used to understand the unknown characters through the quest approach through calculating the capabilities.**

Keywords:-- **Computer Vision; Character Identificatoin; OCR Techniques;**

I. INTRODUCTION

India is multilingual country with 22 reliable languages and greater than 1600 languages in lifestyles, kannada is one of the professional languages and extensively used in the state of Karnataka. Identification of the text written via human is one the promising studies owing its great place of applications concerned. Complexity starts off evolved from extracting textual content out of the image that scanned or captured. Segmentation of the text from the image includes two steps horizontal profiling and vertical profiling which leads to separate lines and words inside the image. Preprocessing of the image are the default steps involved in any image processing technique to put off the noise and to split foreground and background. On this research, figuring out hand written kannada textual content through extracting feature of the textual content written and developing knowledge database for the textual content written.

The character level segmentation is done after removal noise through salt pepper filter and for the foreground and background separation, a traditional and efficient method which became followed with the aid of many researchers is Otsu algorithm. From the foreground, horizontal profiling for the line segmentation and vertical profiling for word or character segmentation. Hu’s Seven Variants and Zernike Moments are used to extract the features and developing knowledge database for the input images.

II. DATA COLLECTION AND PRE-PROCESSING

Data Collection

The valuation of the method that is proposed is only possible with the appropriate dataset. A dataset of our own is created for this research because of the unavailability of any standard dataset of kannada language. A dataset from 20 students of college with the age group of 21 to 25 years are considered. All the alphabets of kannada language are written on a paper and scanned. All the training and testing of the data is being done with the help of this data set. Variation in writing is the obvious thing that happens which helps to generate a strong knowledge for test.

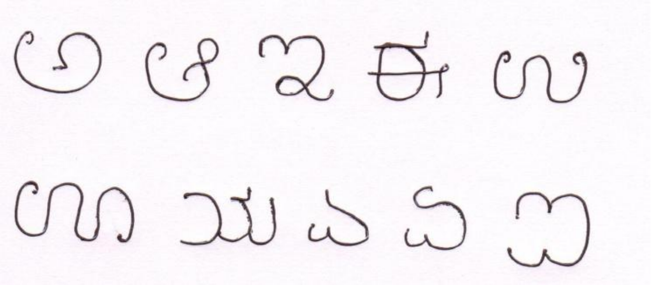


Fig. 1 Sample Data Set

Pre-Processing

For any image to get better performance noise should be removed from the image. For the removal of noise from the image, morphological operations are used. For extracting the foreground from the background Ostu Algorithm is used which converts the image into binary format. Black pixels in the image having 0's and white pixels with 1's are separated henceforth, all the text is extracted from the scanned image.

III. METHODOLOGY

This part of the articles presents the proposed algorithm for identification of hand written kannada scripts. The algorithm for identification of hand written kannada scripts can be divided into different sections like training and testing stages with the subsections like preprocessing, feature extraction and representation.

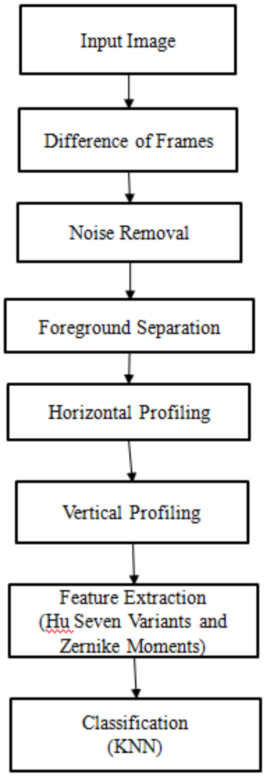


Fig. 2 Flow Chart of Proposed Methodology

Input:

This subsection presents the details of the input considered for training the algorithm. Once the collection of hand written document are collected, they are subjected for horizontal and vertical profiling. These horizontal and vertical profiling is adapted for segmentation of lines and words intern individual characters respectively. This is process is presented in the figure 3.

Once the character level segmentation carried out, features are extracted. In this paper we are considering the well known feature extraction algorithms like Hu's Seven and Moments and Zernike moments.

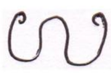
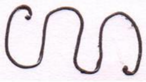
Hu's Seven Variants:

In 1962 Hu presented seven nonlinear functions which are designed with translation, scaling and invariant features. These seven invariant moments of Hu is made an huge impact in the pattern recognition domain and researchers started considering these invariant features for various applications. These seven invariant features are calculated for each word which is segmented through horizontal and vertical profiling methods. These seven variants are formulated as follows.

$V_1 = (n_{20} + n_{02}),$
$V_2 = (n_{20} - n_{02})^2 + 4n^2_{11},$
$V_3 = (n_{30} - 3n_{12})^2 + (3n_{21} - n_{03})^2,$
$V_4 = (n_{30} + n_{12})^2 + (n_{21} + n_{03})^2,$
$V_5 = (n_{30} - 3n_{12})\,(n_{30} + n_{12})\,[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2] +$
$(3n_{21} - n_{03})(n_{21} + n_{03})[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2],$
$V_6 = (n_{20} - n_{02})[(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2] + 4n_{11}(n_{30} + n_{12})(n_{21} + n_{03}),$
$V_7 = (3n_{21} - n_{03})(n_{30} + n_{12})[(n_{30} + n_{12})^2 - 3(n_{21} + n_{03})^2]$ - $(n_{30} + 3n_{12})(n_{21} + n_{03})[3(n_{30} + n_{12})^2 - (n_{21} + n_{03})^2].$

Following table present some of the kannada numerals with the Hu's seven segments. Once of the major reasons for selecting Hu's variants rather than centralised moments, since they do not comprise a complete set of image descriptors.

Table 1 : Following Table Present Some of The Kannada Numerals With The Hu's Seven Segments

	0.1775 0.0002 0.0000 0.0000 0.0000 0.0000 0.0000		0.1901 0.0003 0.0000 0.0000 -0.0000 0.0000 0.0000
	0.1846 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000		0.1882 0.0026 0.0000 0.0000 0.0000 -0.0000 -0.0000
	0.1858 0.0008 0.0000 0.0000 -0.0000 0.0000 0.0000		0.1954 0.0056 0.0000 0.0000 -0.0000 0.0000 -0.0000

Zernike Moments:

Moments based features are statistical measurements used to preserve pixel distribution around center of gravity for identification of character shape information. These techniques are developed to preserve both global and geometric information about the input character. The 2D Zernike Moments of order n with repetition m of an image f(x,y) is defined as

$$Z_m = \frac{n+1}{\pi} \sum_{x=0}^{x=M-1} \sum_{y=0}^{y=N-1} f(\alpha,\beta) V'_{pq} \quad (x,y), \quad \alpha \leq 1$$

Where,

(α,β) = polar coordinates

V'_{pq} = complex conjugate

(α,β) where $\alpha = \sqrt{x^2+y^2}$ And $\beta = \arctan(y/x)$

V'_{pq} is a complex polynomial defined inside a unit circle with the formula

$V'_{pq}((\alpha,\beta) = R_{pq}\ (\alpha)^{jm\,\beta}$

Following figure presents some of the kannada numerals with the zernike moments.

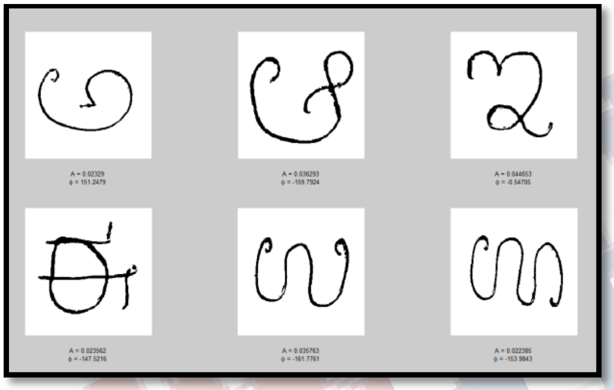


Fig. 2 Flow Chart of Proposed Methodology

Classification stage:

This section of the proposed model addresses the issue of the identification of Kannada character. Identification of kannada characters will be achieved by adapting k-nearest neighbor classifier. Once the features are extracted, knowledge base will be constructed by Hu's variants and Zernike moments. Once the training is process is completed, query sample will be given as input to identify the unknown character to known redefined character. The same feature extraction algorithm is applied and extracted features are compared with the knowledge base by considering Euclidean distance as a proximity measure. The overall procedure is as shown in figure 4.

IV. SIMULATIONS AND RESULTS

Any proposed method need to be evaluated for the effectiveness for which it is developed. To check the efficiency of the proposed model, a simple hand written dataset is constructed. The most well known evaluation metrics like precision, recall and f-measures are considered for the evaluation of the proposed model. The following table presents the details of the experiments carried out on the dataset.

Dataset: Any algorithm need to be verified for correctness. Here also the proposed model is checked for correctness by considering the dataset generated by 20 college students aged between 22 years to 25 years.

Table II Result Of Character Identifcation

Datasets	40% : 60%			60% : 40%		
	Num of Training	Num of Testing	f Measure	Num of Training	Num of Testing	f Measure
Dataset 1	40	60	0.8540	60	40	0.8962

For the purpose of evaluation of results f-measure which is the harmonic mean of precision and recall are calculated for each set of experiments using the equations (2),(3),and (4) respectively.

f-measure = 2PR / P+R . . . (2)

P(Precession) = a / (a + c) . . . (3)

R(Recall) = a/(a+d) . . . (4)

Where

a,b,c and d respectively denote the number of correct positives, false negatives, false positives and correct negatives.

IV. CONCLUSION

The main focus is to classify the Kannada characters using KNN where features are extracted using Hu's Seven Variants and Zernike Moment The performances of the proposed method is measured and displayed in the table II. The major focus of this method is converting the comparing operation into the search method. Huge knowledge database is generated from the dataset. Instead of comparing the image, by extracting the features from the new character and search the features value from database makes identification faster. The drawback of the proposed methodology is knowledge generated during training is pretty high. In future dimensionality techniques can be used to come over the problem.

REFERENCES

1) Dixit, Sunanda, and Suresh Hosahalli Narayan. "Segmentation of Kannada Handwritten Text Line through Computation of Variance." International Journal of Computer Science and Information Security 12.2 (2014): 56

2) Panyam, Narahari Sastry, and RamaKrishnan Krishnan. "Modeling of palm leaf character

recognition system using transform based techniques."Pattern Recognition Letters 84 (2016): 29-34.

3) Pal, U., and B. B. Chaudhuri. "Indian script character recognition: a survey."pattern Recognition 37.9 (2004): 1887-1899.

4) Dhandra, B. V., and M. B. Vijayalaxmi. "A Novel Approach to Text Dependent Writer Identification of Kannada Handwriting." Procedia Computer Science 49 (2015): 33-41.

5) Huang, Zhihu, and Jinsong Leng. "Analysis of Hu's moment invariants on image scaling and rotation." Computer Engineering and Technology (ICCET), 2010 2nd International Conference on. Vol. 7. IEEE, 2010.

6) Karthik, S., and Srikanta Murthy. "Deep Belief Network Based Approach to Recognize Handwritten Kannada Characters Using Histogram of Oriented Gradients and Raw Pixel Features." International Journal of Applied Engineering Research 11.5 (2016): 3553-3557.