Offline English Handwritten Word Recognizer Using Best Feature Extraction

[1]Anuja Naik, [2]M S Patel

Dept. of I.S.E, Dayananda Sagar College of Engineering,Bangalore,India1
Email: [1]anuja2188@gmail.com, [2]msr_patel@yahoo.com

Abstract— Over several years lot of research has been done in handwritten word recognition but yet need for higher recognition rate still exist. In this paper we have proposed a method that performs preprocessing steps like skew and slant correction. It uses best structural features for feature extraction .Euclidean distance method is applied for classification that produces single matching word having minimum difference value. It is an effort towards automating postal services in Karnataka. We have considered 30 districts in Karnataka state written by 25 different writers

Keywords—Handwritten word recognition, skew correction, slant correction, feature extraction.

I. INTRODUCTION

Due to diverse applications in various fields handwritten word recognition is active research area. It can be dichotomized into on-line and off-line recognition [7]. Further offline recognition can be split into holistic and segmentation based [1]. In holistic approach features extracted from entire word image are considered and thus eliminates problems associated with poor placement of segmentation points [11]. Thus computers can make our lives easier by reducing laborious document transactions.

Major application we have focused in our work is automated postal service. Lot of research is going on in postal automation and many articles are available based on automation of various languages. The postal automation system is implemented with respect to printed and offline handwritten words [3]. Hence there is need to work in that respect considering Karnataka state which has thirty districts and 220 talukas we can automate postal system based on these districts and talukas [13].

The main goal is to extract features from handwritten words like district names and compare it with database .Thus additional splitting of word into characters or sub words is eliminated. The typical recognition pipeline has three different stages as shown in Fig.1

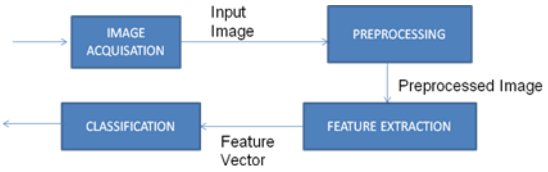

Fig.1. A schematic of recognition system.

This paper contains 5 sections. Section 2 describes pre-processing methods. Section 3 and 4 covers feature extraction and classification. Section 5 includes conclusion.

II. PREPROCESSING

In our work input is a scanned bitmap image of handwritten word. Several preprocessing steps are performed on the input image. Steps include:

A. Skew Detection and Correction

In this, first skew is calculated. To find skew of a word the least black pixel in every column are determined and the set R defined as :

$$R=\{r_i=(x_i,y_i)|\text{lowest pixel in column } x_i\}$$

is populated. Then least squares linear regression is computed to find line of form y=ax+b. Rotation angle is computed as θ=arctan(a).The input image is rotated as per rotation angle to remove skew [10].

B. Slant Detection and Correction

Slant is an angle, clockwise from vertical at which words are drawn. Slant correction is a process that normalises slant of word to vertical. The slant is estimated by finding contour of thresholded image and chain of connected pixels representing edges of stroke. Orientation of those edges close to the vertical is considered as slant. Affine transformation with estimated slant removes slant in word

[12].For each pixel (i,j) in the original image , the new coordinates (x, y) in the slant-corrected image are calculated as follows:

$y = j$
$x = i - (\text{height} - j) * \tan(\theta)$
where θ is the slant angle obtained.

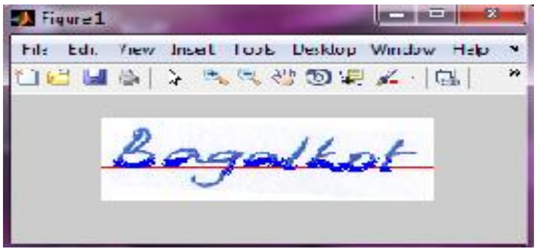

Fig. 2. Skew correction

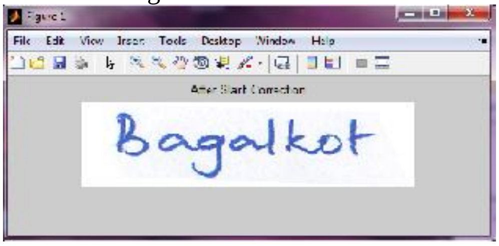

Fig. 3. Slant Correction

C. Baselines Estimation

Upper black pixels and Lower black pixels are used to determine Upper and Lower baselines respectively. Ascender and Descender baselines can be estimated as first y nonzero value and last y nonzero value on vertical histogram of word image. Once baselines are estimated word image is divided as descender part, middle part and ascender part. Remaining part above the ascender baseline and below descender baseline is removed [12].

D.Skeletonization

Input image is first smoothed by convolution with a Gaussian filter to remove noise. Next iterative erosive, thinning algorithm is applied to reduce width of strokes to width of a pixel.

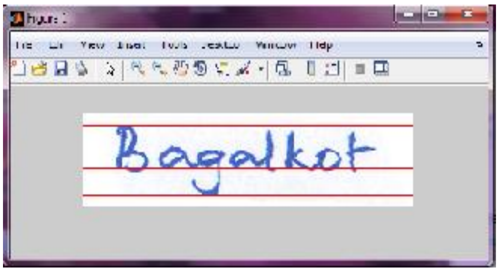

Fig. 4. Baseline Estimation

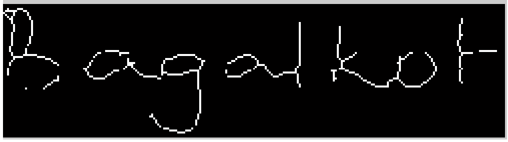

Fig.5. Skeleton of image.

III. FEATURE EXTRACTION

Features extracted from the word are important for classification. This stage plays an vital role as its efficient functioning can lead to better recognition rate. After preprocessing step all the features are extracted from skeleton of an image [4]. In the implementation we have concentrated on structural features. We conducted various experiments and observed that combination of these features gave good results for english handwritten word recognition .Structural features are discriminative and thus error rate is reduced.

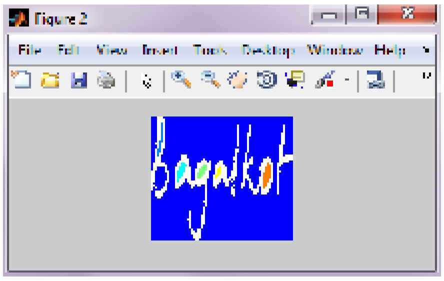

Fig. 6. loops

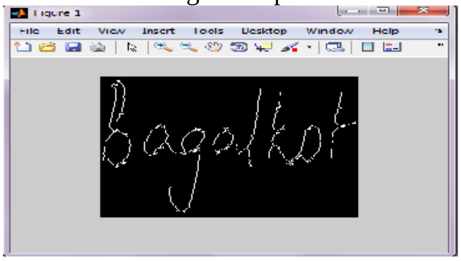

Fig. 7. Junctions

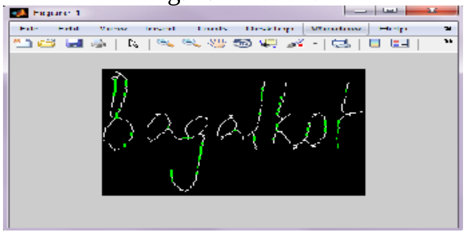

Fig 8:Lines

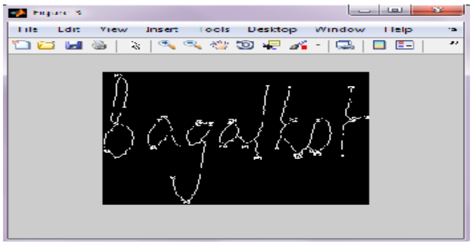

Fig. 9. Endpoint

Features include:

Loops: It represents skeleton inner contours and are found from connected component analysis [8].

Junction points: It is a point where two strokes meet .It can be found as points having two or more neighbours assuming skeleton is of one pixel thickness [5].

Lines: The number of horizontal and vertical lines in preprocessed image [2].

Endpoints: It is a point with only one neighbour indicating end of strokes [5].

For each word in a dataset, features above will be calculated and stored in feature vector to be used for classification.

IV. CLASSIFICATION

Classifier identifies to which set of class test image belongs to based on training data set. In our work Euclidean distance classifier is used. It uses feature vectors of trained and test data set to calculate minimum difference value. It uses Euclidean distance formula and outputs single matching word. We have collected 800 samples of words from 25 different writers of 30 district names of Karnataka. The work is implemented in MATLAB R2012a due to its flexibility of image processing toolbox

V. CONCLUSION

In this paper we have fused various features like loops, lines, junction points and endpoints to get better efficiency. It can be helpful in recognising handwritten responses in forms, postal services and in bankcheck amount. Proposed method is faster than existing methods and more accurate.

REFERENCES

[1] Ankush Acharya, Sandip Rakshit,” Handwritten Word Recognition Using MLP based Classifier : A Holistic Approach,” International Journal of Computer Science Issues, Vol.10, pp.422-427, 2013.

[2] Mamta Garg,Deepika Ahuja"A Novel Approach to Recognize the off-line Handwritten Numerals using MLP and SVM Classifiers," International Journal of Computer Science & Engineering Technology,Vol. 4 No. 07 Jul 2013.

[3] Moncef Charfi, Monji Kherallah, Abdelkarim El Baati, Adel M. Alimi"A New Approach for Arabic Handwritten Postal Addresses Recognition,"proceeding of International Journal of Advanced Computer Science and Applications, Vol. 3, No. 3, 2012

[4] Rajbala Tokas,Aruna Bhadu,” A comparative analysis of feature extraction techniques for handwritten character recognition ,” International Journal of Advanced Technology & Engineering Research (IJATER) ,Vol.2,pp.215-219,2012.

[5] B.B.Saritha,S.Hemanth," An Efficient Hidden Markov Model for Offline Handwritten Numeral Recognition", Proceedings of InterJRI Computer Science and Networking, Vol 1,pp 7-12,2010.

[6] Malik Waqas Sagheer,Chun Lei He,Nicola Nobile,Ching Y.Seun,”Holistic Urdu Handwritten Word Recognition Using Support Vector Machine”, Proceedings of the IEEE,2010.

[7] Rodolfo Luna Perez,Pilar Gome-Gil,” Unconstrained Handwritten Word Recognition Using a Combination of Neural Networks”, Proceedings of the World Congress on Engineering and Computer Science, vol I,2010.

[8] Tal Steinherz, David Doermann,Ehud Rivlin,Nathan Intrator,"Offline Loop Investigation for Handwriting Analysis",Proceedings of IEEE Transactions On Pattern Analysis and Machine Inteigence, vol. 31, no. 2, 2009.

[9] Xiao Chen He,Nelson H. C. Yung “Corner Detector based on Global and Local Curvature Properties,” InOptical Engineering,Vol 47(5),2008

[10] Angélica A. Mascaro and George D. C. Cavalcanti,"Estimating the skew angle of scanned document through background area information ",Proceedings of IEEE Transactions On Computer Graphics and Image Processing, SIBGRAPI '08,2008.

[11] Jose´ Ruiz-Pinales a,*, Rene´ Jaime-Rivas a, Marı´a Jose´ Castro-Bleda,” Holistic Cursive Word Recognition based on Perceptual Features ,” In Pattern Recognition, Vol.28,pp.1600-1609,2007

[12] B.Gatos, I.Pratikakis, A.L.Kesidis, S.J.Perantonis, “Efficient Off-Line Cursive Handwritten Word Recognition,” Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 2006

[13] http://www.karunadu.gov.in/pages/district-list.aspx

◈◈◈