OFFLINE HANDWRITTEN KANNADA WORD RECOGNITION

[1]KRUPASHANKARI.S.SANDYAL, [2]M.S.PATEL

[1,2]Dept. of Information Science and Engineering, Dayanand Sagar College of Engineering, Bangalore-560078, India

Abstract- Recognition of handwritten text has been one of the active and challenging areas of research in the field of image processing and pattern recognition. It has numerous applications which include postal mail application, reading aid for blind and conversion of any handwritten document into electronic form. In this paper we focus on recognition of Karnataka district names in Kannada from a given scanned word image with the help of classifiers. The first step is image acquisition which acquires the scanned image followed by noise filtering, smoothing and resizing of scanned image. Feature Extraction improves recognition rate and misclassification. We use edge detection algorithm to extract sharp edges and features are extracted to train the classifier to classify and recognize the handwritten word.

Index Terms- Handwritten word recognition, feature extraction, Edge Detection algorithm and Classification (key words)

I. INTRODUCTION

As the world moves closer to the concept of the “paperless office,” more and more communication and storage of documents is performed digitally. Documents and files that were once stored physically on paper are now being converted into electronic form in order to facilitate quicker additions, searches, and modifications, as well as to prolong the life of such records. A great portion of business documents and communication, however, still takes place in physical form i.e. handwritten and the fax machine remains a vital tool of communication worldwide. Because of this, there is a great demand for software, which automatically extracts, analyzes, and stores information from physical documents for later retrieval. The overwhelming volume of paper-based data in corporations and offices challenges their ability to manage documents and records. Computers, working faster and more efficiently than human operators, can be used to perform many of the tasks required for efficient document and content management. Computers understand alphanumeric characters as ASCII code typed on a keyboard where each character or letter represents a recognizable code. However, computers cannot distinguish characters and words from scanned images of paper documents. Therefore, where alphanumeric information must be retrieved from scanned images such as commercial or government documents, tax returns, passport applications and credit card applications, characters must first be converted to their ASCII equivalents before they can be recognized as readable text.

Euclidean metric is a popular method to define similarity and index time series, but it is very brittle in computing similarity between time series with different time phases, DTW distance can overcome this problem by searching an optimal match between two given time series. DTW uses a dynamic programming technique to find the minimal distance between two time series, where sequences are warped by stretching or shrinking the time dimension. This method has been efficiently used in speech processing, sign language analysis, online and offline character recognition etc

A. The characteristics of Kannada script

Kannada is the official language of the South Indian state of Karnataka. It has its own script derived from Bramhi script. Kannada script has a base set of 49 characters, comprising 15 vowels and 34 consonants. Further there are distinct symbols that modify the base consonants, called consonant and vowel modifiers [2]. The number of these modifiers is the same as that of the base characters. The characters called aksharas are formed by graphically combining the symbols corresponding to consonants, consonant modifiers (optional) and vowel modifiers using well defined rules of combination.

ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ
ಅಂ ಅಃ

Fig.1: Vowels of Kannada Script

ಕ ಖ ಗ ಘ ಙ

ಚ ಛ ಜ ಝ ಞ

ಟ ಠ ಡ ಢ ಣ

ತ ಥ ದ ಧ ನ

ಪ ಫ ಬ ಭ ಮ

ಯ ರ ಲ ವ ಶ ಷ ಸ ಹ ಳ

Fig-2: Consonants of Kannada Script

A character can be one of the following,
i. A stand alone vowel or a consonant

ii. A consonant modified by a vowel.
iii. A consonant modified by one or more consonants and a vowel

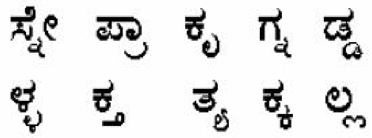


Fig.3: Shows the Conjunct Consonant (Subscript/Vatthu)

B. Motivation

Though so many research papers are available in the literature, still handwritten word recognition is an open problem, especially in the domain of feature design and classification methodology. Researchers have attempted to address this problem from different research perspectives. Most of such attempts are not comparable due to, 1) non availability of standard data sets and, 2) variations in research objectives.

II. PROPOSED METHOD

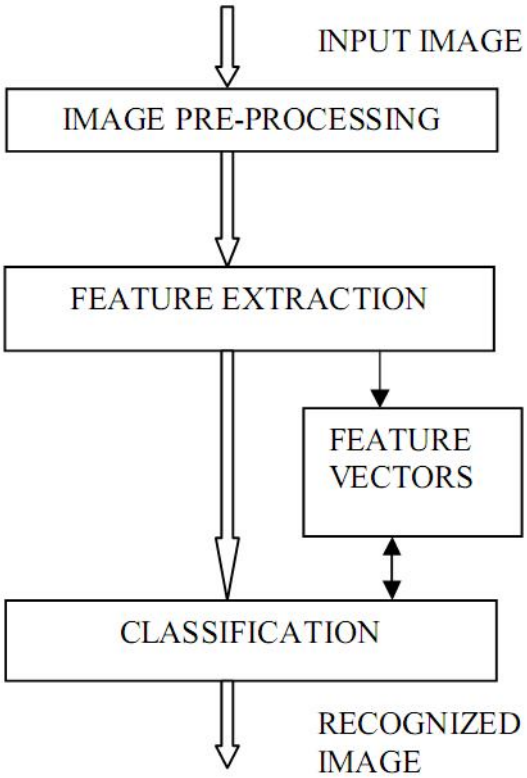


Fig.4: Proposed model

There are 4 Stages in the word recognition process:-
1. Data collection and Pre-processing
2. Feature extraction
3. Classification
4. Recognition

2.1. Typical preprocessing includes the following stages:
2.1.1 Binarization
2.1.2 Noise removal
2.1.3 Skew detection and correction
2.1.4 Thinning

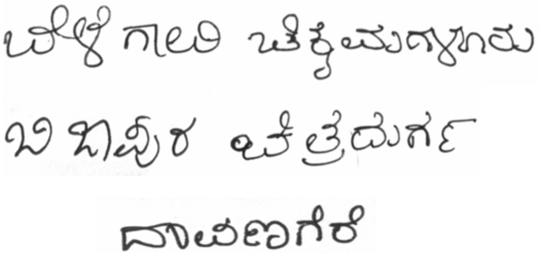


Fig.5: Sample dataset of district names of Karnataka

A total of 1200 Kannada words from around 60 people were obtained and stored as data set. Handwritten Kannada characters usually come in various sizes, shapes and fonts and writers were chosen from schools, colleges and Home. The database was restricted to district names of Karnataka.

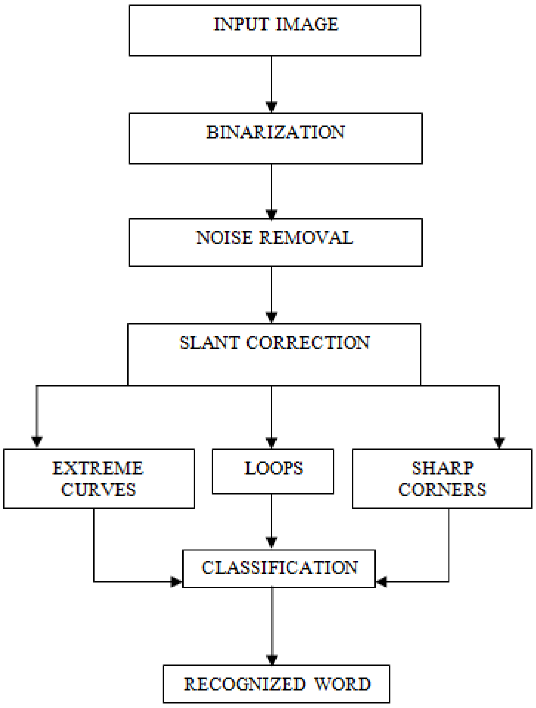


Fig.6: Feature extraction and Classification

2.1.1 Binarization

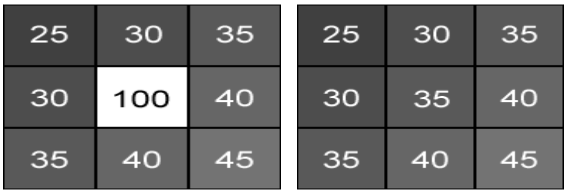
A handwritten document is first scanned and is converted into a gray scale image [1]. Binarization is a technique by which the gray scale images are converted to binary images. The goal is to remove only the background, by setting it to white,and leave the foreground image unchanged. Thus, binarization separates the foreground and background information.

2.1.2 Noise removal

Scanned documents often contain noise that arises due to printer, scanner, print quality, age of the document, etc. Therefore, it is necessary to filter this noise before we process the image.

2.1.2.1 Median filtering

Median filters are particularly effective in the presence of impulse noise, also called 'salt – and – pepper' noise because of its appearance as white and black dots superimposed on an image. For every pixel, a 3x3 neighborhood with the pixel as center is considered [1]. In median filtering, the value of the pixel is replaced by the median of the pixel values in the 3x3 neighborhood.



25	30	35
30	100	40
35	40	45

25	30	35
30	35	40
35	40	45

Fig.7: Median filtering example:
(a)Original image (b)Filtered image

2.1.3 Skew detection and correction

When a document is fed to the scanner either mechanically or by a human operator, a few degrees of tilt is unavoidable. Skew angle is the angle that the lines of text in the digital image make with the horizontal direction [1].

2.1.4 Thinning

The purpose of thinning is to reduce the image components to their essential information so that further analysis and recognition are facilitated. For instance, an alphabet can be handwritten with different pens giving different stroke thicknesses, but the information presented is the same [1]. This enables easier subsequent detection of pertinent features.

Edge detection:-

Canny edge detection algorithm:

Step 1: Read an input image where the sharp edges need to be detected.

Step 2: Call BW=edge(Input_image,'canny',thresh) within Matlab.

Step 3: The above function takes three arguments i.e., 'Input_image' , 'canny' and 'thresh'.

Step 4: Set the 'thresh' value is a two-element vector in which the first element is the low threshold, and the second element is the high threshold. If we specify a value for threshold, this value is used for the high threshold and 0.4* thresh is used for the low threshold. If we do not specify threshold value or if thresh is empty, edge chooses low and high values automatically.

Step 5: If any edge response is above a high threshold, those pixels constitute definite output of the edge detector.

Step 6: Individual weak responses usually correspond to noise, but if these points are connected to any of the pixels with strong responses then there are more chances to be actual edges in the image. Such connected pixels are treated as edge pixels if their response is above a low threshold.

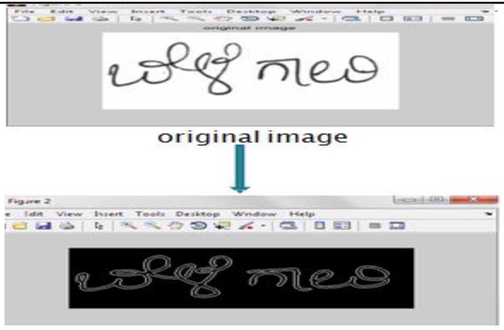


Fig.8: Result of canny edge detection

2. Feature extraction

Major goal of feature extraction is to extract a set of features which maximizes the recognition rate with least amount of elements.

A large number of feature extraction methods are reported in literature; but the methods selected depend on the given application. Since our project is specific to postal application to identify the districts of Karnataka the features we considered are: Corner points, curves and loops.

Steps to detect corner points:

Step 1:Read the input image, it could be gray or binary image.

Step 2:Detect edges using canny edge detector

Step 3:Extract curves from binary edge map

Step 4:Calculate the curvature and find the curvature local maxima as corner points

Step 5:Compare each maxima with its two local minima to remove false corners

Step 6:Finally we get actual corner points.

Freeman's chain code (FCC) is one of the techniques representations based on the boundary extraction which useful for image processing, shape analysis and pattern recognition[7]. Chain code representation gives a boundary of character image where the codes represent the direction of where is the location of the next pixel.

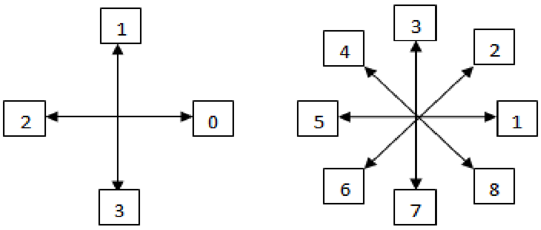


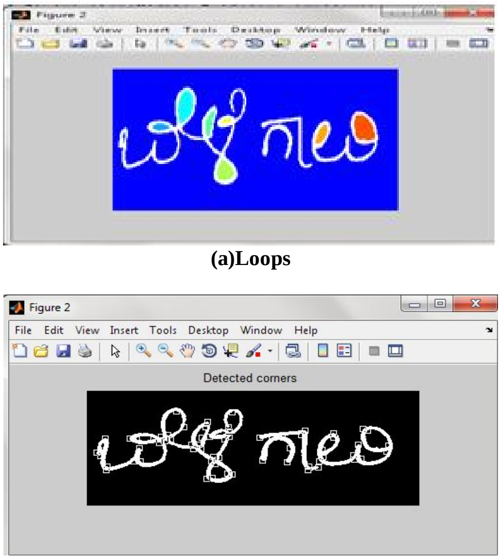
Fig.9: Two Directions of FCC:
(a) 4-Neighborhood (b) 8-Neighborhood.

Freeman's chain code algorithm :

Step 1: Create a binary image from the original input image.

Step 2: Save the image in matrix of 0's for background and 1's for the object. So, while this process is in progress, we compute the area of the object by adding for each iteration.

Step 3: Compute chain code by scanning the image to find the starting pixel of the object. From that pixel, we traverse the boundary and decide directions and save them as an array or list. This step is repeated until we reach the end pixel [5].



(a)Loops

(b)Corner points

Fig.10: Features extracted (a) Loops in a word and (b)Corner points in a word

III. CLASSIFICATION

Classification stage in an HWR(Handwritten Word Recognition) process assigns labels to word images based on the features extracted and the relationships among the features. The principal approaches to decision-theoretic recognition are: minimum-distance classifiers, statistical classifiers and neural networks. There are different types of minimum–distance classifiers like Manhattans, Euclidean distance etc. The Euclidean distance measure is the "standard" distance measure between two vectors in feature space.DTW offers a more flexible way to compensate for the variations like inter character and intra character spacing than linear scaling. Here we use the Euclidean distance and DTW algorithm to recognize the word images based on the features extracted using minimum difference value.

IV. RESULT

The database is created by collecting data from 60 different writers so that recognition engine could be trained with different styles of handwriting. As we were specific to postal mail application we collected writings of all the 30 district names of Karnataka from each writer. So a total of 1200 words dataset was created. The proposed method was trained with a dataset of 250 bitmap images digitized at 300 dpi and the recognition engine is tested for the accuracy. Euclidean distance classifier is used to classify the test images and from experiments done 92% of recognition rate is obtained.

CONCLUSIONS AND FUTURE WORK

The handwritten word is a widely varying recognition target. Relatively high computational power and large storage are required to build a successful HWR system. In this project a method for off-line handwritten word recognition is implemented. This method is based on features identification and distance measure. Comparison of results with other researchers is difficult because of differences in experimental details, the actual handwriting used and the method of data collection.. Euclidean distance function is used to classify the images, which gives the class id of the matched image based on minimum distance value.

The HWR system for Kannada is explored and this method works for connected handwritings with improved recognition rate. The recognition rate can be optimized by using minimum number of features in training process.

REFERENCES

[1] Ahmed Sahlol and Cheng Suen, "A Novel Method for the Recognition of Isolated Handwritten Arabic Characters", Department of Computer Teacher preparation, Damietta University, Damietta, Egypt and Department of Computer Science, Concordia University, Canada, 2013.

[2] Rituraj Kunwar, K.Shashikiran and A.G.Ramakrishnan, "Online Handwritten Kannada Word Recognizer with Unrestricted Vocabulary", Dept of Electrical Engineering, IISc, Bangalore, India,2010.

[3] Venkat Rasagna, Anand Kumar, C.V.Jawahar and R.Manmatha, "Robust Recognition of Documents by Fusing Results of Word Clusters", Center for Visual Information Technology, IIIT, Hyderabad, India,2003.

[4] Toni M. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping", Multi-Media Indexing and Retrieval Group Center for Intelligent Information Retrieval University of Massachusetts Amherst, MA 01003.

[5] Dr. Azzam Talal Sleit and Rahmeh Omar Jabay, "A Chain Code Approach for Recognizing Basic Shapes", Information Technology College University Of Jordan, Amman, Jordan,2006.

[6] Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu and Mita Nasipuri, "Handwritten Word Recognition Using MLP Based Classifier", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013.

[7] Dewi Nasien, Habibollah Haron and Siti Sophiayati Yuhaniz, "The Heuristic Extraction Algorithms for Freeman Chain Code of Handwritten Character", Faculty of Computer Science and Information System (FSKSM) Universiti Teknologi Malaysia Skudai, 81310, Malaysia.

★★★