

OFFLINE HANDWRITTEN DEVANAGARI CHARACTER RECOGNITION USING FUSION OF CLASSIFIERS

Prof. Shalaka Deore¹, Snehal G. Zaware², Anagha A. Shelke³, Amina A. Shaikh⁴, Archana B. Dhakne⁵

¹Department of computer Engineering, MESCOE, Pune 411001

²U.G. Student, Department of computer Engineering, MESCOE, Pune 411001

³U.G. Student, Department of computer Engineering, MESCOE, Pune 411001

⁴U.G. Student, Department of computer Engineering, MESCOE, Pune 411001

⁵U.G. Student, Department of computer Engineering, MESCOE, Pune 411001

Abstract - Now a days, there is need for the digitalization of handwritten documents. There is a large scope of research in this area. Continuous improvement in Handwritten Character Recognition (HWCR) techniques milestones in this research area. HWCR system is the software to accept and process the handwritten input images from sources such as documents, photographs. HWCR is the ability to transform them to machine readable and editable format. Devanagari script is used as base for various Indian languages such as Marathi, Sanskrit, Hindi, etc. and foreign languages such as Nepali. The work proposed in our Handwritten Devanagari Characters Recognition System tries to automate recognition of handwritten Devanagari isolated characters by ensembling different classifiers. Ensemble classifier is constructed by using Support Vector Machine (SVM) [1], K-Nearest Neighbor (KNN) [1] and Neural Network (NN) [2] which increases the performance by ensembling classifiers. The proposed system gives better results than individual classifiers.

Keywords- HWCR, SVM, KNN, NN.

I. INTRODUCTION

Handwritten Character Recognition is an important area of research in image processing. It has its applications in banks, post offices, etc. Online handwriting recognition means accepting the input directly on electronic devices such as PDA, or sensors. Offline handwriting recognition means accepts inputs in form of images. These images can be taken from papers, documents, etc. Literature survey carried out for this project showed that character recognition of Indian scripts has a great demand. HWCR is at a very primary stage of research in Indian scripts. The work done for Devanagari scripts can be adopted for different languages such as Marathi, Hindi, etc. The challenge here is create a software system that will recognize handwritten Devanagari characters as these characters can have different font size, font style and shape. There are several character recognition technique such as SVM, NN, Decision tree, etc. for character recognition. Each of these techniques has its own advantages and disadvantages. The drawbacks and loopholes of using individual techniques can be overcome if they are ensembled together. Ensembling the classifiers is an emerging approach which binds together the advantages of different classification techniques. By fusion of classifiers this project aims to improve and maximise the rate of recognition.

II. DEVANAGARI SCRIPT

Devanagari script is the base of many Indian language such as Hindi, Marathi, etc. This script comes under Brahmic family of scripts. Devanagari script consists of 12 vowels, 36 base forms of consonant, 10 numeral characters. Consonant characters, vowel characters and Numeral characters are shown in Figure 1.

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ	अं	अः
---	---	---	---	---	---	---	---	---	---	----	----

a)

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	ठ
ड	ढ	ण	त	थ	द	ध	न	प	फ	ब	भ
म	य	र	ल	व	स	ष	श	ह	क्ष	त्र	ज्ञ

b)

०	१	२	३	४	५	६	७	८	९
---	---	---	---	---	---	---	---	---	---

c)

Figure. 1. Printed samples a) Consonant characters b) vowel characters c) Numeral characters

III. LITERATURE SURVEY

J. Pradeep et. Al [3] proposed an handwritten character recognition system with a new feature extraction technique. The new technique is call diagonal feature extraction and intersection points. They divided the image into zones and extracted features diagonally. The recognition rate found was 99 %.

Anupama Kaushik et.al [4] presented a paper that focuses on how feature extraction is important for classification of handwritten text. They used classifiers such as J48, Naïve Bayes and Sequential Minimal optimization (SMO). They used such attributes which can contribute most to the classification. Also they carried out comparative performance analysis of classifiers used for OCR and pattern recognition.

Sasan Karamizadeh et.al [5] presented a paper that states the advantages and drawback of using the support vector machines. Among multiple supervised machine learning algorithms, SVM is one of them. Linear and non-linear patterns are classified using SVM. Patterns which are easily distinguishable and separable are called linear patterns. In non-linear pattern, it is difficult to classify two clusters. The main drawback of SVM is that there is lack of transparency in results generated.

An ensemble model was presented by Pushpalata Pujari et.al [6] for Odia Handwritten Character Recognition system. The four base classifiers used were Support Vector Machine (SVM) [1], Artificial Neural Network (ANN) [2], C5.0 Decision Tree and Discriminant Analysis (DA). Confidential weighted scheme is used to combine the output of classifiers to obtain ensemble models.

A hybrid KNN-SVM model was presented by Cleber Zanchettin et.al [7] for recognition of character. This sytem is implemented for cursive character. SVM classifier is used to increase the performance rate of KNN . This system showed improved recognition rate for as compared to MLP, KNN [1] and a hybrid MLP-SVM approach .

IV. PROCESSING STEPS OF HWCR

The basic steps of HWCR processing are discussed in this section. It is shown in diagrammatic form in Figure 2. In the proposed system, noise is removed by preprocessing the images. These images are further given as input for feature extraction. The features are then given to the classification phase. According to the classifiers the input is then recognised.

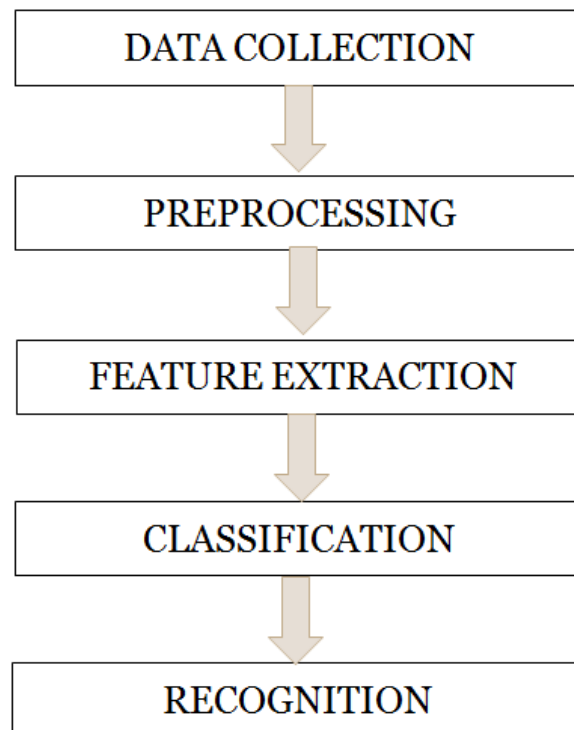


Figure 2 : Processing of HWCR

4.1. Data Collection

Data from various individuals are collected and obtain individual characters. Samples have been taken for each individual character to ensure different orientations and size. For an efficient recognition system, feature extraction and classification are vital players but database characters are equally important. Unconstrained database makes character recognizers more suitable for real life applications, but the basic shape of the characters should not be distorted and they must be recognizable by naked human eyes. The collected dataset is divided into 85% for training and 15% for testing.

4.2. Preprocessing

4.2.1 Binarization

It is a process of converting the RGB image given as input into the gray scale image. This image is then transformed into binary image (0 & 1).

4.2.2 Noise Removal

It is very much essential to remove and filter out the noise before processing the image. The most common way is filter the image as low pass filter and then use it for the later stages.

4.2.3 Skeletonization

This method is also known as thinning. It is a process of reducing the thickness or width of a line. It is as similar as converting the object having many pixels wide to just one single pixel wide. It makes the recognition algorithm more simple as they only have to work on a single pixel wide of a character stroke. It also reduces the memory requirement for storing information about the input character provided. Processing time is also reduced to a greater extent.

4.2.4 Skew Detection & Correction

Skew detection is used to refer to the tilt in an image. A particular degrees of skew is not avoidable. Skew angle is the angle that the lines of the character in the image makes with the horizontal direction.

4.3. Feature Extraction

There are different feature extraction methods that are used for extracting features that can be used in classification stage. Feature extraction is an essential stage as the accuracy of the result depends upon the features extracted from an image and as a result reduces misclassification rate.

4.3.1 Diagonal Features and Intersection Points

Here features from image are extracted diagonally. Also intersection points are found. Taking example of 90 X 60 pixels image. Zones created will be 54. 10 * 10 pixels will be the size of each zone. As a result 19 of such sub features are found for each of the zone. A single feature is extracted by performing the average of 19 sub features for each of the zone. Averaging the zone features using row wise and column wise calculation will result in 9 and 6 features. Hence in all 69 features are obtained per image.

4.3.2 Projection Histogram Features

This technique counts the number of pixels in a specific direction. It has three types which includes: horizontal, vertical and left diagonal. In case of horizontal method, the foreground pixels are counted row wise. In case of vertical method, the pixels are counted column wise. In case of left diagonal method, pixels are counted diagonal wise.

4.3.3 Distance Profile Features

In this method the number of pixels are counted from the boundary of the character image to the outer edge of character. In this technique, profiles of four sides are used. The four sides are left, right, bottom and top.

4.3.4 Zoning

In zoning method N*M zones are created and the densities of pixels present in each of the zone is calculated. The number of pixels are counted for each zone and dividing it by total number of pixels present to calculate density.

4.4. Ensemble Classifier

An ensembling of classifiers is collection of similar and dissimilar classifiers whose outputs are combined to classify the test data. Ensembling improves the overall accuracy and it makes classifier robust. Assume the ensembling of n identical or different classifiers $(f_1, f_2, \dots, f_n)$. If all the classifiers are identical, then result of ensembling is identical to the result of individual classifier. If all the classifiers are different then the correct result will be achieved by majority voting method. Thus fusion of dissimilar classifiers gives better performance than the individual classifiers. So ensembling can be performed by using different classifiers, which are discussed below.

4.4.1 Support Vector Machine

It is a maximum margin classifier. Each data item is plotted in n-dimensional space. Here number of features are represented by n. Number of features are mapped to number of coordinates. An SVM constructs (N-1) dimensional hyper-plane such that there exists maximum separation margin between the two datasets. Hyperplane is used to differentiate the data into two classes. In case of non linear classification, SVM uses kernel function that transforms the given data into higher dimensional space so that it becomes possible to perform separation using lines instead of fitting non-linear curves to the data. Linear SVM classifies dataset by a straight line between different object classes. However, in most of the classification task, data is nonlinearly separable i.e. no straight line exist to correctly classify the objects of dataset. Thus, more complex structures are needed to obtain an optimal separation.

4.4.2 k-Nearest Neighbor

To classify a new object it checks its k-nearest neighbors from the training data and similarity between two patterns can be computed using Euclidean, Minkowski or Mahalabonis Distance. Based upon the value of K, the characters are classified into different classes. For high dimensional data, weighted k-NN can be used.

4.4.3 Artificial Neural Network

Artificial Neural networks is an interconnected group of nodes. These nodes are connected by links and organized into different layers. Input, hidden and output are three layers used in ANN. Patterns are given through the input layer. Input layer is connected to hidden layers. The actual processing is done in these hidden layers via a system of weighted connections. The hidden layer is connected to an output layer. Weights are adjusted to get desired output.

V. CONCLUSION

In this paper, we proposed a method for recognition of offline Devanagari handwritten characters using combination of dissimilar classifiers i.e. SVM, KNN and ANN. In this study, we have used zoning, profile projection and diagonal feature extraction method. Finally, the results from all the classifiers are combined to get the final result using maximum voting method

REFERENCES

- [1] Seema Sharma, Jitendra Agrawal ,Shikha Agarwal, Sanjeev Sharma, "Machine Learning Techniques for Data Mining: A Survey," IEEE 2013
- [2] Guoqiang Peter Zhang, " Neural Networks for Classification: A Survey," IEEE Transactions 2000
- [3] J.Pradeep,E.Srinivasan,S.Himavathi. "Diagonal Feature Extraction Based Handwritten Character System Using Neural Network," International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010
- [4] Anupama Kaushik,Himanshu Gupta,Digvijay Singh Latwal. "mpact of Feature Selection and Engineering in the Classification of Handwritten Text," IEEE 2016
- [5] Sasan Karamizadeh , Shahidan M. Abdullah , Mehran Halimi ,Jafar Shayan and Mohammad javad Rajabi , "Advantage and Drawback of Support Vector Machine Functionality," IEEE 2014
- [6] Pushpalata Pujari and Babita Majhi, "Performance Comparison of Ensemble Models for Recognition of Offline Handwritten Odia Numerals," IEEE 2015
- [7] Cleber Zanchettin, Byron Leite Dantas Bezerra and Washington W. Azevedo. "A KNN-SVM Hybrid Model for Cursive Handwriting Recognition," IEEE 2012