PEHCHAAN: HINDI HANDWRITTEN CHARACTER RECOGNITION SYSTEM BASED ON SVM

SONIKA DOGRA
Dapartment of CSE, Lovely Professional University
Jalandhar (INDIA)
Sonika997@gmail.com

CHANDRA PRAKASH
Assistant Professor , Dapartment of CSE, Lovely Professional University
Jalandhar (INDIA)
Cpiiitm2010@gmail.com

Abstract— **Optical Character Recognition is a technique by which you can automatically recognize the characters with an optical mechanism. OCR technology allows you the recognition of printed or handwritten text documents. Main aim of this research is to prepare a recognition system which can be used for the recognition of offline handwritten Hindi characters. For this proposed system Support Vector Machine is used as classifier and Diagonal feature extraction approach is used to extract features.**

Keywords- Handwritten Character Recognition, OCR, Feature Extraction, SVM.

I. INTRODUCTION

It is easy to recognize typed characters but handwritten characters cannot be recognized with 100% accuracy by computer machine. So, Handwritten character recognition is still a difficult task. Handwritten character recognition is further divided into two domains i.e Offline handwritten character recognition and Online handwritten character recognition

A. *Offline handwritten character recognition:*

In case of offline character recognition, the typed/handwritten characters are scanned and then converted into binary or gray scale image. Then feature extraction and recognition process is carried over the binary image. Offline character recognition is a more challenging and difficult task as there is no timing information about character strokes is available. Therefore offline character recognition is considered as a more challenging task then its online counterpart.

B. *Online handwritten character recognition:*

Online handwritten character recognition is also known as real time recognition of characters. In this case writing and recognition are done simultaneously. User will write character on any sensory area where sensors will pick up the pen movements and then on the basis of those pen movements characters are recognized. Online character recognition is much easier than offline character recognition, because timing information is available there.

In proposed system, SVM will be used as classifier and diagonal feature extraction approach is used for the extraction of features of handwritten characters.

II. SUPPORT VECTOR MACHINE

SVMs were originally proposed by Boser, Guyon and Vapnik in 1992 and gained increasing popularity in late 1990s.SVMs are currently among the best performers for a number of classification tasks. Currently, SVM is widely used in object detection & recognition, text recognition, biometrics, speech recognition, etc. SVM is based on binary classification, means at a time it can classify two class groups.

[1]Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. Most classification tasks, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). [2] In all the experiments, the results have shown that at character level, SVM recognition rates are significantly better due to structural risk minimization implemented by maximizing margin of separation in the decision function.

III. DEVANAGRI SCRIPT AND DATA COLLECTION

Devanagri is composed of two two Sanskrit words, “deva” and “nagri”. Deva means God and nagri means city. Hindi is written using Devanagari script. This script is used to write many other languages, such as Nepali, Marathi etc. Devanagari consists of 11 vowels and 33 consonants. There are no capital letters. Devanagari is written from left to right. [3] Hindi is an Indo-Aryan language which is written using devanagri script.

Table 1 vowels in Hindi

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ

Table 2 Consonants in Hindi

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	श
ष	स	ह		

IV. PROPOSED SYSTEM

Proposed recognition system is consisting of several phases. They are shown with following diagram:

A. Digitization

Digitization is the process in which paper document is converted into electronic form. For this, handwritten documents are scanned thus an image is produced. This image is then fed to the next pre-processing phase.

B. Pre-processing

Pre-processing is the initial stage of character recognition. In this phase, the character image is normalized into a fixed size. The pre-processing is a series of operations performed on the scanned input image. It essentially enhances the image rendering it suitable for segmentation.[4] Pre-processing phase enhances the image rendering. The various tasks performed on the image in pre-processing stage are binarization process that converts a gray scale image into a binary image and many more.

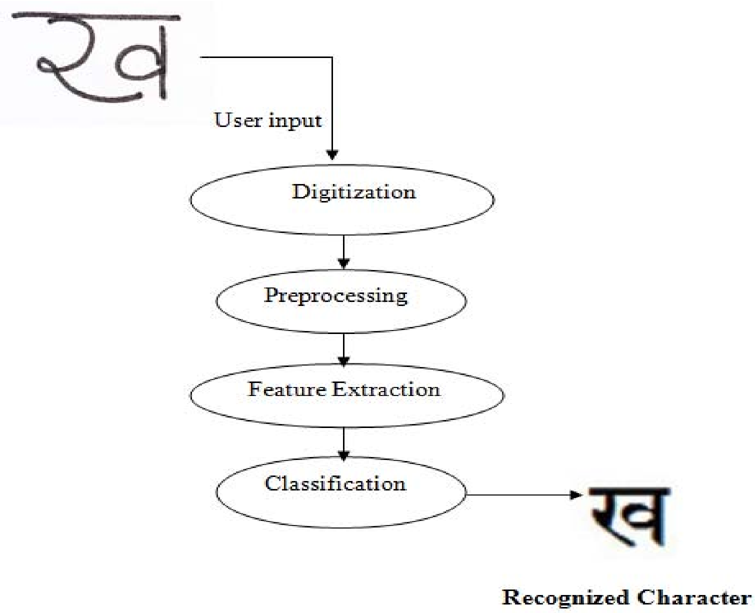
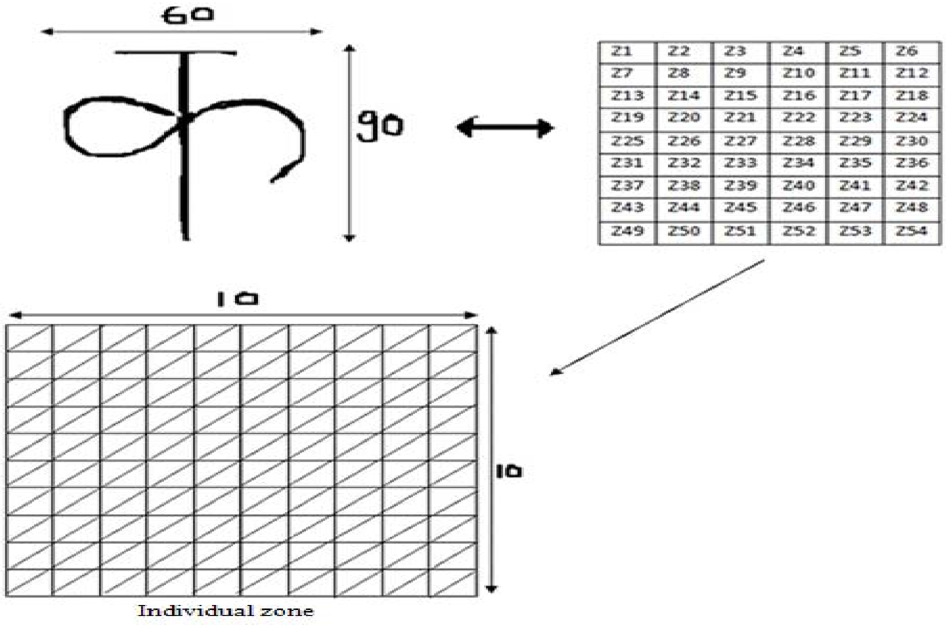


Figure 1 Proposed system

C. Feature extraction

The feature extraction stage analyzes a handwritten character image and selects a set of features that can be used for the classification of input character. In proposed system diagonal feature extraction approach is selected for extracting the features from input image. Every character image of size 90x 60 pixels is divided into 54 equal zones, each of size 10x10 pixels. The features are extracted from each zone pixels by moving along the diagonals of its respective 10X10 pixels. Each zone has19 diagonal lines and the foreground pixels present along each diagonal line are summed to get a single sub feature and thus 19 sub-features are obtained from the each zone. These 19 sub-features values are averaged to form a single feature value and placed in the corresponding zone. This procedure is sequentially repeated for the all the zones. Finally, 54 features are extracted for each character. [4]



Z1	Z2	Z3	Z4	Z5	Z6
Z7	Z8	Z9	Z10	Z11	Z12
Z13	Z14	Z15	Z16	Z17	Z18
Z19	Z20	Z21	Z22	Z23	Z24
Z25	Z26	Z27	Z28	Z29	Z30
Z31	Z32	Z33	Z34	Z35	Z36
Z37	Z38	Z39	Z40	Z41	Z42
Z43	Z44	Z45	Z46	Z47	Z48
Z49	Z50	Z51	Z52	Z53	Z54

Figure 2 Diagonal Feature Extraction Approach

D. Classification

Classification phase is the decision making phase of an Handwritten character recognition system. This phase uses the features extracted in the previous stage for deciding that input character belongs to which class. In this work, we have used Support Vector Machine (SVM) classifier for recognition. The SVM is a very useful technique for data classification. [6] SVMs are based on statistical learning theory that uses supervised learning. In supervised learning, a machine is trained instead of programmed to perform a given task on a number of inputs/outputs pairs. The utilization of support vector machine (SVM) [5, 6] classifiers has gained immense popularity in the last years. SVMs have achieved excellent recognition results in various pattern recognition applications.

V. EVALUATION

On the basis of samples of handwritten characters collected from 20 people, the average accuracy of character recognition is 93.06%

Table 3 Recognition Rate

S.NO	CHARACTER	ACCURACY	S.NO	CHARACTER	ACCURACY
1	अ	100%	26	ड	95%
2	आ	100%	27	ढ	90%
3	इ	95%	28	ण	100%
4	ई	90%	29	त	95%
5	उ	100%	30	थ	90%
6	ऊ	100%	31	द	90%
7	ऋ	95%	32	ध	85%
8	ए	80%	33	न	90%
9	ऐ	80%	34	प	95%
10	ओ	95%	35	फ	90%
11	औ	95%	36	ब	95%
12	अं	90%	37	भ	95%
13	अः	100%	38	म	90%
14	क	95%	39	य	90%
15	ख	100%	40	र	90%
16	ग	95%	41	ल	85%
17	घ	80%	42	व	95%
18	ङ	90%	43	श	90%

19	च	95%	44	ष	100%
20	छ	95%	45	स	90%
21	ज	90%	46	ह	95%
22	झ	95%	47	क्ष	95%
23	ञ	95%	48	त्र	95%
24	ट	95%	49	ज्ञ	90%
25	ठ	100%	50	AVERAGE	93.06%

This system is giving 93.06% Accuracy. In Hindi language some characters can be written in different ways, because of this reason system is not giving 100% accuracy.

VI. CONCLUSION AND FUTURE SCOPE

From the results it can be concluded that combination of SVM classifier and diagonal feature extraction approach is best method for the recognition of handwritten characters. The Future work may involve the recognition of words and complete sentences as well as speech can be synthesized for the individual character that is recognized by the system

REFERENCES

[1] Anuj Sharma, R.K. Sharma, Rajesh Kumar, "Online Handwritten Gurmukhi Character Recognition", Ph.D. Thesis, Thapar University, 2009 [Online].

[2] Abdul Rahim Ahmad, Christian Viard-Gaudin, Marzuki Khalid, Emilie Poisson (2004) "Online Handwriting Recognition using Support Vector Machine" TENCON 2004 IEEE Region 10 Conference, Vol. No. 1 , pp 311-314.

[3] http://hindilanguage.info/devanagari

[4] J. Pradeep, E.Srinivasan, S.Himavathi (Oct 2010) "Diagonal Feature Extraction Based Handwritten Character System Using Neural Network." *International Journal of Computer Applications (0975 – 8887)*

[5] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.1, 2.2, 2.2

[6] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000. 1, 2.2, 2.2

AUTHORS PROFILE

Sonika Dogra is currently pursuing Masters of Technology in CSE from Lovely Professional University (LPU), Phagwara, India. She did her Bachelors of Technology (B.Tech) from PTU. Her Research area of interest includes OCR, Pattern Recognition.

Mr. Chandra Prakash is currently working at the rank of Assistant Professor in Computer Science Department of Lovely Professional University (LPU) , Punjab, India. He did his Integrated Masters (BTech and MTech) in Information Technology from Indian Institute of Information Technology and Management Gwalior in 2010. His areas of research are Data Mining, Speech Recognition, Text summarization, Machine learning, Human-Computer Interfaces and Biometrics