RESEARCH ARTICLE OPEN ACCESS

Online Handwritten Sanskrit Character Recognition Using Support Vector Classification

Prof. Sonal P.Patil[1], Ms. Priyanka P. Kulkarni[2]
[1] Assistant Professor, Computer Science, GHRIEM, Jalgaon, Maharashtra, India
[2] Research Scholar, Computer Science & Engg, GHRIEM, Jalgaon, Maharashtra, India

Abstract
Handwritten recognition has been one of the active and challenging research areas in the field of image processing. In this Paper, we are going to analyses feature extraction technique to recognize online handwritten Sanskrit word using preprocessing, segmentation. However, most of the current work in these areas is limited to English and a few oriental languages. The lack of efficient solutions for Indic scripts and languages such as Sanskrit has disadvantaged information extraction from a large body of documents of cultural and historical importance. Here we use Freeman chain code (FCC) as the representation technique of an image character. Chain code gives the boundary of a character image in which the codes represents the direction of where is the location of the next pixel. Randomized algorithm is used to generate the FCC. After that, features vector is built. The criterion of features toinput the classification is the chain code that converted to various features. And segmentation is applied to evaluate the possible segmentation zone. Accordingly, several generations are performed to evaluate the individuals with maximum fitness value. Support vector machine (SVM) is chosen for the classification step.

Key Words: Freeman chain code (FCC), Heuristic method, Support vector machine (SVM),

I. INTRODUCTION

Human can accurately recognize the handwritten characters if they are neat and clean. It is very easy task for human beings. The same can do easily by the kids also. But the same task for machine is very difficult. Various languages use specific script to write. Hindi & Marathi are most commonly used languages by several thousand people [1]. most of the current work in these areas is limited to English and a few oriental languages. The lack of efficient solutions for Indic scripts and languages such as Sanskrit has hampered information extraction from a large body of documents of cultural and historical importance. Sanskrit Character contains complicated curves & various shapes. So recognition of Sanskrit characters is difficult & complicated task[2]. All these considerations make Optical Character recognition (OCR) with Sanskrit script very challenging. The ultimate goal of designing a character recognition system with an accuracy rate of 100 % is quite difficult because handwritten characters are non uniform; they can be written in many different styles. Different writers can be written various sizes of handwritten character. Even there is variation in characters written by the same writer at different time [3] The problem of exchanging data between human beings and computing machines is challenging. Basically character recognition is a process, which associates a symbolic meaning with objects (letters, symbols and numbers) drawn on an image, *i.e.*, character recognition techniques associate a symbolic identity with the image of a character.

1.1 Optical Character Recognition

Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn shown in Figure1.1. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. The more constrained the input is, the better will the performance of the OCR system be. However, when it comes to totally unconstrained handwriting, OCR machines are still a long way from reading as well as humans. However, the computer reads fast and technical deviances are continually bringing the technology closer to its ideal [3].

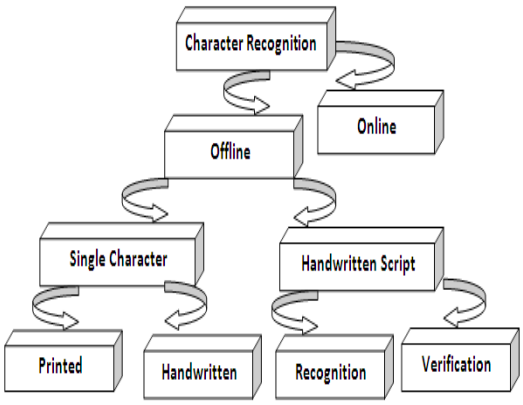

Fig -1: The different areas of character recognition

1.2 classification of character recognition system

1.2.1 Classification According To Data Acquiring Process

A) Online CRS

Online Character recognition system involves electronic digitizer as shown in Figure1.2. A special electronic pen samples the handwriting input and writing is done on electronic surface. Digitizer takes the temporal or dynamic information of writing. This information consist of pen strokes (i.e. the writing from pen down to pen up), the order of pen strokes the direction of writing and the speed of writing within each stroke. The online handwriting signal contains additional information that is not accessible in offline.

Fig -2: Electronic Digitizer

B) Offline CRS

In Offline method a piece of paper is used to write the character and scan directly into the system by a scanner or camera as shown in Figure1.3. In this system, the image of writing is converted into a bit pattern by an optically digitized device such as optical scanner or camera. The bit pattern data is shown by matrix of pixels. The main task of this offline recognition system is to recognize handwritten character on letters or parcels.

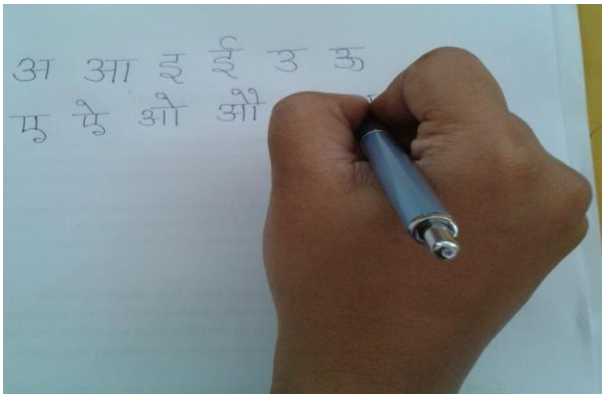

Fig -3: Handwritten Character captured by camera

1.2.2. Classification According To Text Type

A) Printed CRS

Printed text includes all the printed materials such as book, newspaper, magazine and documents. Machine printed text & numbers have uniform nature. For any font printed characters are of same size. The recognition rate is very much dependent on the age of the documents, quality of the paper and ink which may result in significant data acquisition noise.

B) Hand written CRS

There is non-uniformity in handwritten characters. There are different ways for writers to write the characters. & in various sizes. Even there is variation in writing of the same writer at different times. Handwritten character may vary in shape also. So handwritten character recognition is the most difficult part of character recognition [2].

The storage of scanned documents have to be large in size and many processing applications as searching for a content, editing, maintenance are either hard or impossible. Such documents require human beings to process them manually, for example, postman's manual processing for recognition and sorting of postal addresses and zip code. OCR as the short form of Optical character recognition, which translates such scanned images of printed or handwritten documents into machine encoded text. According to this requirements translated machine encoded text can be easily edited, searched and can be processed in many other ways. It also requires small size for storage in comparison to scanned documents. Optical character recognition helps human's simplicity and reduce their jobs of manually handling and processing of documents. Computerized processing to recognize individual character is required to convert scanned document into machine encoded form.

1.3. Character Recognition Architecture

Optical character recognition involves many steps to completely recognize and produce machine

encoded text. These phases are termed as: Pre-processing, Segmentation, Feature extraction, Classification. The architecture of these phases is shown in figure 1.2 and these phases are listed below with brief description [4].

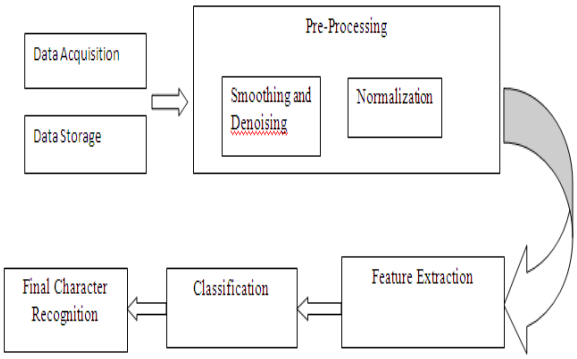

Fig -4: The stages of handwriting recognition system

Pre-processing: The pre-processing phase normally includes many techniques applied for binarization, noise removal, skew detection, slant correction, normalization, contour making and skeletonization like processes to make character image easy to extract relevant features and efficient recognition.

Segmentation: Segmentation phage, which sometimes considered within pre-processing phase itself, involves the splitting the image document into classifiable module object generally isolated characters or modifiers. Generally practiced segmentations are line segmentation, word segmentation, character segmentation and horizontal segmentation to separate upper and lower modifiers particularly in context to most Indian scripts.

Feature Extraction: Feature extraction is used to extract relevant features for recognition of characters based on these features. First features are computed and extracted and then most relevant features are selected to construct feature vector which is used eventually for recognition. The computation of features is based on structural, statistical, directional, moment, transformation like approaches.

Classification: Each pattern having feature vector is classified in predefined classes using classifiers. Classifiers are first trained by a training set of pattern samples to prepare a model which is later used to recognize the test samples. The training data should consist of wide varieties of samples to recognize all possible samples during testing. Some examples of generally practiced classifiers are- Support Vector Machine (SVM), K- Nearest Neighbour (K-NN), Artificial Neural Network (ANN) and Probabilistic Neural Network (PNN).

1.4 Issues of handwritten character recognition

To recognize handwritten documents, either online or offline, the character recognition is much affected by style variations of handwriting by different writers and even different styles of same writer on different times. Distortion and noise incorporated while digitization is also a major issue in character recognition that affects the recognition accuracy negatively[4]. many character recognition issues regarding handwritten character recognition are listed below:

Handwriting Style Variations

Different writers and even same writer have different handwriting styles. Many times a person finds himself/herself unable to recognize his/her own handwriting. Hence, practically it is much difficult to recognize handwriting by machine efficiently. Deformed geometry, slants, skews, overlapping, noise, distortion are inserted by different writers in different ways. Geometric properties like aspect ratio, position and size vary. One can note that some kind of variations also exists in each sample of a character although such samples share high degree of similarities. The shape of a character is also influenced by the word in which it is appearing. Characters can look similar although their number of strokes, and the drawing order and direction of the strokes may vary considerably.

Constrained and Unconstrained Handwriting

The characters to recognize may be constrained or unconstrained. Because unconstrained documents include all possible style variation, so such document are much difficult to recognize. To make this recognition process easier atleast for laboratory work, constrained documents are practiced to recognize.

In constrained document format, the handwritten samples are written in standard format that make the characters easy to recognize. The constrained document has box discrete and space discrete nature of characters or words. In box discrete nature, each character is written in separate standard sized box. In space discrete nature, characters are written have much space from one-another to make the segmentation and recognition easier.

Unconstrained documents consist of touching, overlapping and cursive characters. Cursive writing makes recognition difficult due to stroke variability. Touching characters are difficult to segment the characters from each other and overlapping characters makes this situation worse.

Writer Dependent or Independent Recognition

Writer dependent recognition system is used to recognize the samples of only those writers whose samples are takes to train the recognition system. That is, are specific to a group of writers. In writer dependent system, all possible style variations can be trained to system, hence a higher recognition rate can be obtained.

At the other hand, writer independent system needs the generalization of the recognition system also to recognize the handwritten samples of unknown writers. Hence, it needs to train the system with all possible and commonly used style variations. Hence it need to train the system with large number of samples taken from large number of different types of writers, to make the recognition system generalized. Due to recognition of unknown samples, the recognition rate of writer independent system is comparatively lower. In practice, writer independent systems are in more demand because of generalized application.

Personal and Situational Aspects

Personal factors include writer's writing style which might be affected by handedness- either left handed or right handed and. Many persons are habitual to write random or specific inclined text lines. A good recognition requires neat and clean handwriting and this writing style also depends on profession of writers to some extent. The situational aspects depend on the facts either writer is interested or not to write, how much attention a writer is paying, text is written giving proper time or in hurry, whether there was any interruption while writing, how was the quality of material used for writing etc.

Number of stroke classes

Due to the presence of composite characters in Indian writing systems, a large number of stroke classes are possible. These stroke classes represent consonants, vowels and modifiers or combinations of consonants and vowels. The large number of stroke classes and the shape complexity of various strokes increase the complexity of the recognition system. This is addressed by choosing the efficient recognition algorithm which does not degrade with a large number of classes.

Directionality of writing

There exist big variations in the directionality of writing strokes and stroke segments which could affect the uniformity in stroke representation using certain features. It is necessary to identify writing direction invariant features for representing the stroke .

1.5Application of handwritten character recognition

Handwritten character recognition system is basically used to convert a sample image character to corresponding machine coded character. This basic characteristic can be used to derive many other practical applications. Some of these applications are listed below:

- **Check reading:** The abundant numbers of checks are need to process in banks. Handwritten or printed OCR system can be used for automatically reading the name of recipient, signature verification, filled amount reading and reading all other information.
- **Form processing:** Form processing can be used to process forms and documents of public applications. In such forms handwritten information is written in space provided. This handwritten information can be processed by Handwritten OCR system automatically.
- **Signature verification:** On legal or other documents that include the signatures of authorities and persons, signature can be verified using handwritten character recognition system. The system once can be trained by various samples of signatures of required persons and later on any document their signatures can be verified.

II. LITEATURE SURVEY

2.1Overview

In this chapter literature survey regarding the previous work and related approaches about character recognition is presented. The most advanced and efficient OCR systems are designed for English, Chinese and Japanese like scripts and languages. In context to Indian languages and scripts, recently a significant research work is proposed. Most of Indian work on character recognition is dominated by Devanagari script, which is used in writing of Hindi, Marathi, and Nepali and Sanskrit languages. Devanagari script is dominating because Hindi which is written in Devanagari script is spoken by a mass of Indian population and is national language of India. After Devanagari, second position of recognition research is taken by Bangla script [6]. Recently much research work is also practiced on other Indian scripts like Gurumukhi [7], Gujarati, Tamil, Telugu [8] and many other scripts. Most of the current work in these areas is limited to English and a few oriental languages. The lack of efficient solutions for Indic scripts and languages such as Sanskrit has hampered information extraction from a large body of documents of cultural and historical importance.

In my literature survey I have studied many research approached practiced on many languages and scripts particularly in Indian context. Our

emphasis is to study and analyse the feature extraction approaches, observed or reported results and many other relevant issues considerable to any new research work on OCR. After my detailed literature survey I am able to discover the theme of my proposed work including the techniques I have incorporated in various phases, to implement it and to evaluate my results and conclusions.

2.2 works related to character recognition

The first research work report on handwritten Devanagari characters was published in 1977 [5]. For Indian languages most of research work is performed on firstly on Devanagari script and secondly on Bangla script. U. Pal and B.B. Chaudhury [6] presented a survey on Indian Script Character Recognition. This paper introduces the properties of Indian scripts and work and methodologies approached to different Indian script. They have presented the study of the work for character recognition on many Indian language scripts including Devanagari, Bangla, Tamil, Oriya, Gurumukhi, Gujarati and Kannada.

U. Pal, Wakabayashi and Kimura also presented comparative study of Devanagari handwritten character recognition using different features and classifiers [7]. They used four sets of features based on curvature and gradient information obtained from binary as well as gray scale images and compared results using 12 different classifiers as concluded the best results 94.94% and 95.19% for features extracted from binary and gray image respectively obtained with Mirror Image Learning (MIL) classifier. They also concluded curvature features to use for better results than gradient features for most of classifiers.

A later review of research on Devanagari character recognition is also presented by Vikas Dungre et al. [8]. They have reviewed the techniques available for character recognition. They have introduced image pre-processing techniques for thresholding, skew detection and correction, size normalization and thinning which are used in character recognition. They have also reviewed the feature extraction using Global transformation and series expansion like Fourier transform, Gabor transform, wavelets, moments ; statistical features like zoning, projections, crossings and distances ; and some geometrical and topological features commonly practiced. They also reviewed the classification using template matching, statistical techniques, neural network, SVM and combination of classifiers for better accuracy is practiced for recognition.

Prachi Mukherji and Priti Rege [9] used shape features and fuzzy logic to recognize offline Devanagari character recognition. They segmented the thinned character into strokes using structural features like endpoint, cross-point, junction points, and thinning. They classified the segmented shapes or strokes as left curve, right curve, horizontal stroke, vertical stroke, slanted lines etc. They used tree and fuzzy classifiers and obtained average 86.4% accuracy.

Giorgos Vamvakas et al. [10], [11] have described the statistical and structural features they have used in their approach of Greek handwritten character recognition. The statistical features they have used are zoning, projections and profiling, and crossings and distances. Further through zoning they derived local features like density and direction features. In direction features they used directional histograms of contour and skeleton images. In addition to normal profile features they described in- and out- profiles of contour of images. The structural features they have depicted are end point, crossing point, loop, horizontal and vertical projection histograms, radial histogram, radial out-in and in-out histogram.

Sarbajit Pal et al. [12] have described projection based statistical approach for handwritten character recognition. They proposed four sided projections of characters and projections were smoothed by polygon approximation.

Wang Jin et al. [13] evolved a series of recognition systems by using the virtual reconfigurable architecture-based evolvable hardware. To improve the recognition accuracy of the proposed systems, a statistical pattern recognition-inspired methodology was introduced.

Sandhya Arora et al. [1] used intersection, shadow features, chain code histogram and straight line fitting features and weighted majority voting technique for combining the classification decision obtained from different Multi Layer Perceptron (MLP) based classifier. They obtained 92.80% accuracy results for handwritten Devanagari recognition. They also used chain code histogram and moment based features in [13] to recognize handwritten Devanagari characters. Chain code was generated by detecting the direction of the next in-line pixel in the scaled contour image. Moment features were extracted from scaled and thinned character image.

Fuzzy directional features are used in [14] in which directional features were derived from the angle between two adjacent curvature points. This approach was used to recognize online handwritten Devanagari characters and result was obtained with upto 96.89% accuracy.

2.3 work on Sanskrit character recognition

Namita Dwivedi have described recognition of Sanskrit word using Prewitt's operator for

extracting the features from an image thinning process is applied in pre processing technique Thinning is an important pre-processing step in OCR. The purpose of thinning is to delete redundant information and at the same time retain the characteristic features of the image. Freeman Chain code is one of the representation techniques that is useful for image processing, shape analysis and pattern recognition fields is used with heuristic approach for feature extraction. Genetic algorithm is used for non linear segmentation of multiple characters. This recognition model was built from SVM classifiers for higher level classification accuracy [2]

III. METHODOLOGY

This section describes about the architecture model of HCR in detail.

3.1 Preprocessor

In pre-processing scanned document is converted to binary image and various other techniques to remove noise, to make it ready and appropriate before feature extraction and further computations for recognition are applied. These techniques include segmentation to isolated individual characters, skeletonization, contour making, normalization, filtration etc. Which types of pre-processing techniques will suite; it highly depends on our requirements and also is influenced by mechanism adopted in later steps. Some of the pre-processing techniques are listed and discussed briefly in following sections

A.Gray Scale Image

If the input document is scanned in colored image format, It may be required to first convert it into gray scale, before converting to binary image

B. Binarization

Binarization converts colored (RGB) or gray scale image into binary image. In case of converting colored image it first needs to convert it into gray image. To convert a gray image into binary image we require to specify threshold value for gray level, dividing or mapping range of gray level into two levels either black or white i.e. either 0 or 1 not between it [16].

C. Smoothing and Noise Removal

Smoothing operations are used to blur the image and reduce the noise. Blurring is used in pre-processing steps such as removal of small details from an image. In binary images, smoothing operations are used to reduce the noise or to straighten the edges of the characters, for example, to fill the small gaps or to remove the small bumps in the edges (contours) of the characters [17].

D. Skew Detection and Correction

Deviation of the baseline of the text from horizontal direction is called skew. Document skew often occurs during document scanning or copying. This effect visually appears as a slope of the text lines with respect to the x-axis, and it mainly concerns the orientation of the text lines.

E. Slant Correction

The character inclination that is normally found in cursive writing is called slant. The general purpose of slant correction is to reduce the variation of the script and specifically to improve the quality of the segmentation. To correct the slant presented first we need to estimate the slant angle (θ), then horizontal shear transform is applied to all the pixels of images of the character/digit in order to shift them to the left or to the right (depending on the sign of the θ)[15].

F. Character Normalization

Character normalization is considered to be the most important pre-processing operation for character recognition. Normally, the character image is mapped onto a standard plane (with predefined size) so as to give a representation of fixed dimensionality for classification. The goal for character normalization is to reduce the within-class variation of the shapes of the characters/digits in order to facilitate feature extraction process and also improve their classification accuracy. Basically, there are two different approaches for character normalization: linear methods and nonlinear methods. These methods are described in [15].

G. Thinning (Skeleton)

Thinning is an important pre-processing step in OCR. The purpose of thinning is to delete redundant information and at the same time retain the characteristic features of the image. Thinning is applied to find a skeleton of a character. Skeleton is an output of thinning process. Thinning is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening. It can be used for several applications, but is particularly useful for skeletonization. In this mode it is commonly used to tidy up the output of edge detectors by reducing all lines to single pixel thickness. Thinning is normally only applied to binary images, and produces another binary image as output. A simple example of thinning of a simple binary image is shown in fig 5

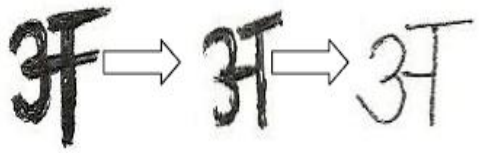

Fig -5: Skeleton produced by thinning process

3.2 Feature Extraction

A. View based features

This method is based on the fact, that for correct character-recognition a human usually needs only partial information about it its shape and contour. This feature extraction method, which works on scaled, thinned diarized image, examines four "views" of each character extracting from them a characteristic vector, which describes the given character as shown in 1.5. The view is a set of points that plot one of four projections of the object (top, bottom, left and right) it consists of pixels belonging to the contour of the character and having extreme values of each block. Thus for 5 × 5 blocks we get 5 × 5 × 8 = 200 features for recognition one of its coordinates. For example, the top view of a letter is a set of points having maximal *y* coordinate for a given *x* coordinate. Next, characteristic points are marked out on the surface of each view to describe the shape of that view [16].

Fig -6: Selecting characteristic points for four views

B. Shadow Features of character

Shadow is basically the length of the projection on the sides as shown in Figure 1.6. For computing shadow features on scaled binary image, the rectangular boundary enclosing the character image is divided into eight octants. For each octant shadows or projections of character segment on three sides of the octant dividing triangles are computed so, a total of 24 shadow features are obtained. Each of these features is divided by the length of the corresponding side of the triangle to get a normalized value [8].

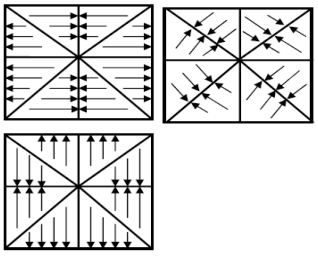

Fig -7: Shadow Features of character

C. Chain Code Histogram of Character Contour

Given a scaled binary image, first find the contour points of the character image. Consider a 3 × 3 window surrounded by the object points of the image. If any of the 4-connected neighbor points is a background point then the object point (P), as shown in Figure 1.6 is considered as contour point.

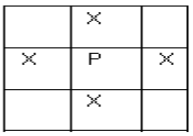

Fig -8: C

The contour follow ontour representation called "chain coding" that is used for contour shown in Figure 1.7. Each pixel of the contour is assigned a different code that indicates the direction of the next pixel that belongs to the contour in some given direction. Chain code provides the points in relative position to one another, independent of the coordinate system. In this methodology of using a chain coding of connecting neighboring contour pixels, the points and the outline coding are captured. Contour following procedure may proceed in clockwise or in counter clockwise direction. Here, we have chosen to proceed in a clockwise direction [8]. The main problem in representing the characters using FCC is the length of the FCC that depends on the starting points. Also, during FCC generation that require traversing each pixel (or node) of the character, it is often to find the problem of finding several branches and revisiting the same nodes. To solve this problem, heuristic is used to generate the FCC correctly to represent the characters [2].

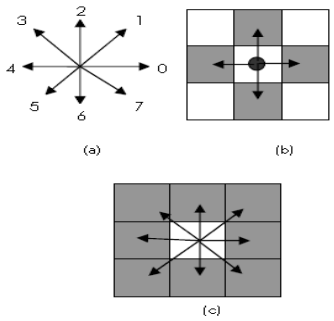

Fig -9: Chain Coding: (a) direction of connectivity, (b) 4-connectivity, (c) 8-connectivity

Heuristic is a method to find a solution that is closed to the best but it does not guarantee that the best will be found. Heuristic methods are proposed which is randomized algorithm. The pseudo code of randomized algorithm is used to generate the FCC of character. The procedure is as following: start from the first node, which is node-method and end-node-method. Node method is to find the first character for every aspects boundary such as left upper, left lower,

right upper and right lower. End node- method is to find the first character based on the end position of a character .In this randomized algorithm, if the number of visited node less than the number of nodes, there would be three kinds of characteristics, which are unvisited, visited and taboo neighbors'. Unvisited neighbors' are nodes that Never went through the route searching. Visited neighbors indicate the nodes that have went through the route searching. Taboo neighbors are used to keep track of the visited search space and revisited node with one step after current node. The criteria of features to input the segmentation stage is the chain code that converted to features are from the calculated values of ratio-upper, ratio-right, ratio-heightweight,ratio-height and number of string character. The ratio-upper is calculated from firstly by cropping the image and then defining the centre of the image character. After that, the total number of upper character is divided with the total number of character. This is done similarly to the ratio-right, ratio-height-weight and ratio-height. The formula of height character as shown in Equation [2]

Height=Height of character/Height of image
Randomized Algorithm (Pseudo code)-
Initialize Data while Termination Not Met do
Select First Node Randomly
{
Node-Method, End –Node-Method
}
while Number of Visited Node<Number of Node do
if there are Unvisited neighbors
Selected one Node Randomly
else if there are visited neighbors
Selected one Node Randomly
else if there are taboo neighbor
Selected one Node Randomly
end if
end while
end while
Display solution

IV. SEGMENTATION

Segmentation partitions the digital image into multiple segments. These segments form the local zones of the image in such a way that is useful to extract features for character recognition. It is a process of assigning a label to each pixel in an image such that pixels with same label share certain visual characteristics. In character recognition generally we need following types of segmentation processes one after another sequentially proceeding to segmentation into smaller objects. These segmentation processes are discussed below. The first three types of segmentation are discussed in [19].

A. Line Segmentation: However straight text lines having enough thoroughly horizontal white space between lines, can be segmented easily, but in practice, particularly in the case of handwritten texts, this technique cannot succeed. In context of India script having header lines a prevalent approach is to detect lines by horizontal projections. Header line is used to have maximum number of pixels while base line is used to have minimum number of pixels in horizontal projections.

B. Word Segmentation: After line segmentation, the text in each line is segmented to detect words. It is called word segmentation. Word segmentation is easier than line segmentation, because there is generally enough space presents between words and each word is bounded by header line having no space within word.

C.Character Segmentation: Now after word segmentation, each word needs to segment into characters. It is called character segmentation. In Indian script having header lines use to remove header line first to have vertical space between characters within the word under consideration.

D.Zone Segmentation: In most Indian scripts like Devnagari, Sanskrit there are modifiers presented either above the header line or below the base line. The basic and conjunct characters are presented in the middle horizontal zone of below the header line and above the base line. Most Indian researchers prefer to segment the text horizontally in three zones namely upper, middle and lower zones. Header line and base line, having maximum and minimum number of pixels in horizontal projections respectively are used to separate these zones.

V. CLASSIFIERS

OCR systems extensively use the methodologies of pattern recognition, which assigns an unknown sample to a predefined class. Various techniques for OCR are investigated by the researchers. The Concept of SVM (Support Vector Machine) was introduced by Vapnik and co-workers. It gains popularity because it offers the attractive features and powerful machinery to tackle the problem of classification i.e., we need to know which belongs to which group and promising empirical performance. The SVM is based on statistical learning theory. SVM's better generalization performance is based on the principle of Structural Risk Minimization (SRM) .The concept of SRM is to maximize the margin of class separation. The SVM was defined for two-class problem and it looked for optimal hyper-plane, which maximized the distance,

the margin, between the nearest examples of both classes, named SVM [2].

At present SVM is popular classification tool used for pattern recognition and other classification purposes. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. The standard SVM classifier takes the set of input data and predicts to classify them in one of the only two distinct classes. SVM classifier is trained by a given set of training data and a model is prepared to classify test data based upon this model. For multiclass classification problem, we decompose multiclass problem into multiple binary class problems, and we design suitable combined multiple binary SVM classifiers [17].

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. The classifier that separates a set of objects into their respective classes with a line. Most classification tasks, however, 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).

SVM Algorithm-

1. Choose a kernel function
2. Choose a value for *C (control over fitting)*
3. Solve the quadratic programming problem (many software packages available)
4. Construct the discriminant function from the support vectors

VI. PERFORMANCE ANALYSIS

Researchers have investigated OCR for a number of Indian scripts: Devnagari, Tamil, Telugu, Bengali, and Kannada, Gurumukhi by using different feature extraction technique and different classifier. Some analyses from different research paper are listed below.

Table -1: Performance analysis

Sr. no	Method Purpo By	Feature Extraction technique	Classifier	Data size	Accuracy Obtained
1	S. Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu, and M. Kundu[1]	1)Shadow feature of Character 2)Chain code histograr of character contour 3) Finding intersection junction in character	MLP	4900	92.80%
2	Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, L. Malik , M. Kundu and D. K. Basu[14]	1)Shadow feature of Character 2)Chain code histograr of character contour 3)View based features 4)Longest run features	1) SVM 2) ANN	1)4900 ISI Kolkata-3430 used for training 2) 2254 Own database-1470 used for training and 784 used for testing	ISI database 1)SVM-94.77% 2)ANN 93.31% Own Database 1)SVM-99.62% 2)ANN 99.51%
3)	Mahesh Jangid[32]	1)Foreground pixel Distribution 2)None density feature 3) background directio distribution feature	SVM	12240 – 314 samples are used	94.89%
4)	Sandhya Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu, M. Undu[18]	1)Shadow feature of Character 2)Chain code histograr of character contour	MLP	1)4900 ISI Kolkata-3430 used for training 2) 2254 Own database-1470 used for training and	90.74%
5)	Ashutosh Aggarwal, Rajneesh Rani, RenuDhir[17]	Gradient Feature Extraction	SVM	7200 – 200 samples of 36 characters. 7200 are used	94 %

VII. CONCLUSIONS

In this paper, we have presented a system for recognizing a handwritten character. This proposes a model for handwritten Sanskrit character recognition. The model starts with pre-processing. The pre-processing stage involves all of the operations to produce a clean character image, so that it is can be used directly and efficiently by the feature extraction stage. Thinning process is used in the pre-processing stage that produced a skeleton of a character. The second step is feature extraction. FCC is generated from the characters that used as the features for segmentation. The main problem in representing the characters using FCC is the length of the FCC that depends on the starting points. Also, during FCC generation that require traversing each pixel (or node) of the character, it is often to find the problem of finding several branches and revisiting the same nodes. To solve this problem, heuristic is used to generate the FCC correctly to represent the characters. The third step is segmentation using the features generated from FCC. Correct segmentation of characters is mandatory for their successful

recognition. This recognition model was built from SVM classifiers for higher level classification accuracy.

REFERENCES

[1] Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu, and M. Kundu, "Combining Multiple Feature Extraction Techniques for Handwritten Devanagari Character Recognition", Proc. IEEE Region 10 Colloquium and Third Intl. Conf. Industrial & Information Systems, Kharagpur (India), pp. 978-1-4244-2806 2008.

[2] Namita Dwivedi, Kamal Srivastava and Neelam Arya," sanskrit word recognition using prewitt's operator and support vector classification", IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013)

[3] Prof. M.S.Kumbhar1, Y.Y.Chandrachud, "Handwritten Marathi Character Recognition Using Neuarl Network", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 9,September 2012.

[4] Shruthi Kubatur, Maher Sid-Ahmed, Majid Ahmadi,'' A Neural Network Approach to Online Devanagari Handwritten Character Recognition"

[5] Jayaraman, A., Chandra Sekhar, C., Chakravarthy, V. S., "Modular approach to recognition of strokes in Telugu script", Proceedings of the Ninth International Conference on Document Analysis and Recognition, September 23-26, pp 501-505, 2007.

[6] U. Pal, B.B. Chaudhury, "Indian Script Character Recognition: A Survey", Pattern Recognition, Elsevier, pp. 1887-1899, 2004.

[7] U. Pal, Wakabayashi, Kimura, "Comparative Study of Devanagari Handwritten Character Recognition using Different Feature and Classifiers", 10th International Conference on Document Analysis and Recognition, pp. 1111-1115, 2009.

[8] Vikas J Dungre et al., "A Review of Research on Devanagari Character Recognition", International Journal of Computer Applications, Volume-12, No.2, pp. 8-15, November 2010.

[9] Prachi Mukherji, Priti Rege, "Shape Feature and Fuzzy Logic Based Offline Devanagari Handwritten Optical Character Recognition", Journal of Pattern Recognition Research 4,pp. 52-68, 2009.

[10] G. Vamvakas, B. Gatos, S. Petridis, N. Stamatopoulos, "Optical Character Recognition for Handwritten Characters".

[11] G. Vamvakas, B. Gatos, S. Petridis, N. Stamatopoulos, "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition", Ninth International Conference on Document Analysis and Recognition(ICDAR), Vol.2, pp. 1073-1077, September 2007.

[12] Sarbajit Pal, Jhimli Mitra, Soumya Ghose, Paromita Banerjee, "A Projection Based Statistical Approach for Handwritten Character Recognition," in Proceedings of International Conference on Computational Intelligence and Multimedia Applications, Vol. 2, pp.404-408, 2007.

[13] Wang Jin, Tang Bin-bin, Piao Chang-hao, Lei Gai-hui, "Statistical method-based evolvable character recognition system",IEEE International Symposium on Industrial Electronics (ISIE), pp. 804-808, July 2009.

[14] Arora, D. Bhattacharjee, M. Nasipuri, M.Kundu, D.K. Basu, "Application of Statistical Features in Handwritten Devanagari Character Recognition", International Journal of Recent Trends in Engineering (IJRTE), Vol. 2, No. 2, pp. 40-42, November 2009.

[15] Mehmet Sezgin and Bulent Sankur, "Survey over image thresholding techniques and quantitative performance evaluation",Journal of Electronic ImagingVol. 13, Issue 1, pp.146–165, January 2004.

[16] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, L. Malik , M. Kundu and D. K. Basu, " Performance Comparison of SVM and ANN for Handwritten Devanagari Character Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010.

[17] Dr.Renu Dhir and Mrs.Rajneesh Rani,"Handwritten Gurumukhi character recognition"

[18] Mahesh Jangid, "Devanagari Isolated Character Recognition by using Statistical features (Foreground Pixels Distribution, Zone Density and Background Directional Distribution feature and SVM Classifier) ",International Journal on Computer Science and Engineering Vol. 3, No. 6, June 2011.