Use of Dominant Point detection Feature for Recognition of Online Handwritten Devanagari Script

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Pooja Tewari
801032019

Under the supervision of:
Mr. Karun Verma
Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2012

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, **"Use of Dominant Point Detection Feature in Recognition of Online Handwritten Devanagari Script"**, in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in CSED of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Mr. Karun Verma and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Pooja

(Pooja Tewari)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. Karun Verma)
Assistant Professor
Computer Science and Engineering Department
Thapar University
Patiala

Countersigned by

(Dr. Maninder Singh)
Associate Professor and Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

I am deeply obliged to various personalities for providing the assistance, inspiration, co operation and encouragement throughout the thesis work. The work can't be accomplished without the leadership and guidance of these peoples.

Firstly, I would like to express my honest thanks to my respected supervisor, **Mr. Karun Verma in the Department of Computer Science and Engineering** for sharing his valuable knowledge and providing his precious guidance in carrying out this work. With the encouragement and opportunities presented by him, I have been able to complete this work successfully. The feedback and editorial comments provided by him helped a lot in fruitful completion of this work. He has devoted a lot of time in discovering the terms used in the thesis work.

I would also like to extend my special thanks to **Dr. Maninder Singh Associate Professor Head of CSED** for providing necessary facilities in the department and for being a source of inspiration for me during this work.

I am also grateful to **Dr. S.K Mohapatra, Dean of Academic Affair** for his encouragement and constant support that played a great role in accomplishment of this task.

I am also thankful to all the staff members of the Department for their full support, cooperation and help.

My heartiest thanks are to all who wished me success especially my parents without whom support and constant encouragement, I can't dream of completing my work.

(Pooja Tewari)

Place: TU, Patiala

Date:

ABSTRACT

After the invention of computers, a great amount of work has been done in the field of computer human interface. But the problem of exchanging data between human beings and computing machines is still challenging. An enormous research has been done for efficient character recognition and even now research is going on.

Basically Character recognition techniques associate a symbolic identity with the image of a character. To make the complicated process of online handwritten character recognition easier and more robust, focus should be on salient features of character.

After pre-processing feature extraction is done. Feature extraction is a vital phase of character recognition. Features extracted from character encode the structural characteristics of character shape. One of the methods to get efficient recognition is to extract dominant points of characters.

Dominant points can be used to recognize the character more proficiently. Dominant points are commonly considered as points with local maximum curvature (elevated position). In other words, it can be said that dominant points of a character are those points where the slope value changes noticeably. The dominant points are taken as output from feature extraction phase and input for recognizing the character. The dominant points are extracted for characters and also the distance of these points from centre of character is calculated. On the basis of these features, recognition is accomplished. Recognition is done through SVM.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

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. Character recognition can be categorized into following two parts: -

1. Online Character Recognition

2. Offline Character Recognition

Pen-based interfaces are now increasingly popular and have a vital role in human-computer interaction, which stimulates the greater interests in online handwriting recognition. Provided with both temporal stroke information and spatial information, online handwriting recognition is expected to give better performance than offline handwriting recognition. It is a much easier tool, pen device can take over many functions of the conventional mouse of the computer, but also it is more natural to human writing. Reliably transformation of the handwritten characters into codes understandable by computers is needed for popular usage of pen devices. Due to the mass increase of pen computing applications and various pen input devices, online handwriting recognition is gaining renewed interests.

Basically in offline approach, spatial information is taken into account, while in online handwriting recognition approach temporal trajectories are taken into account.

Instead of several years of research, it is still a research problem to develop an online handwritten Devanagari character recognition system which is robust against the possible global distortions of an input handwriting sample.

The smallest unit of character recognition is stroke. Stroke is basically the movement of pen on the writing pad from the moment the pen is down to the moment the pen is up. Each sentence is made up of large number of words, these words in turn made up of characters and these characters are made up of stroke group. Pixel is the smallest unit of binary image. It is the smallest addressable unit in Display device. Direction of connectivity defines which pixels are connected to each other. The connected pixels in a binary image are called object. Various pen computing devices are available in market for online handwriting recognizing like apple Newton, Go’s tablet computer, palm series, tablet pc.

1.1 Character Recognition

1.1.1 Online character recognition

In the case of online recognition, a time ordered sequence of coordinates, representing the pen movement, is available. This may be produced by any electronic sensing device, such as a mouse, an electronic pen on a tablet. Figure 1.1 shows units of HW.

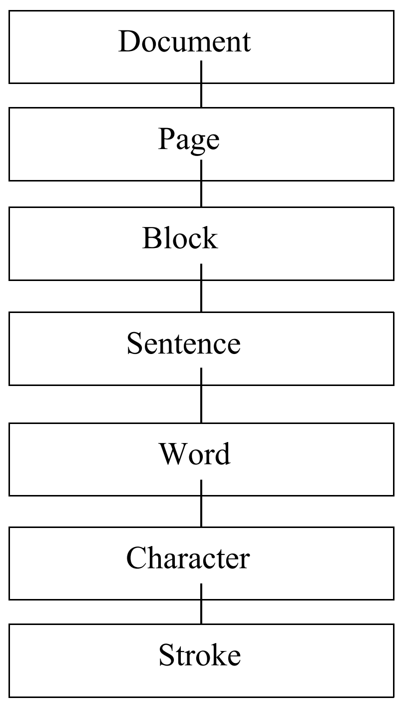


Figure 1.1 Representation of the units of online handwriting

1.1.2 Offline character recognition-In case of offline character recognition only the image of the text is present, which is usually scanned or photographed from paper. Below Figure 1.2 represents types of character recognition.

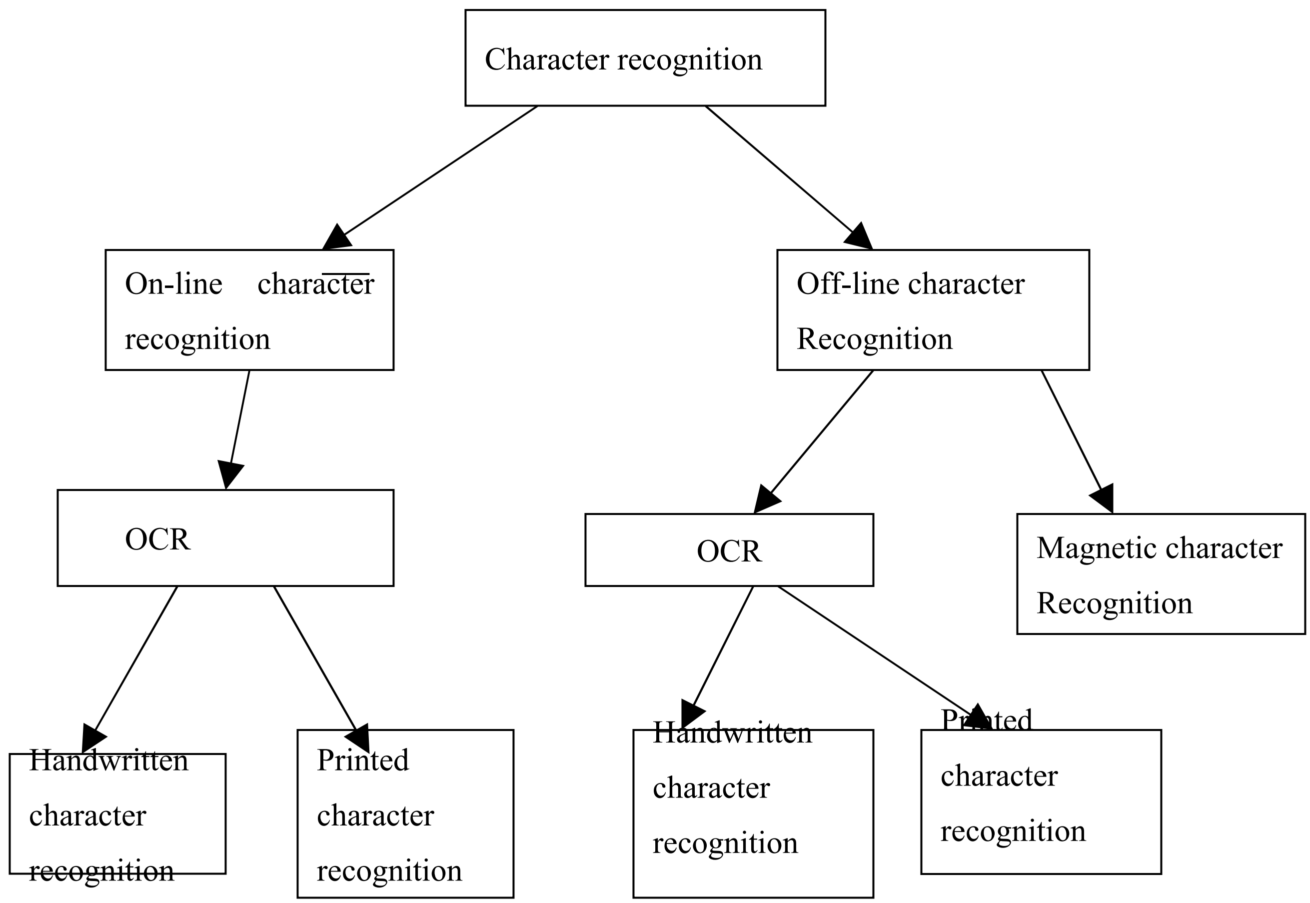

Character recognition
On-line character recognition
Off-line character Recognition
OCR
OCR
Magnetic character Recognition
Handwritten character recognition
Printed character recognition
Handwritten character recognition
Printed character recognition

Figure 1.2 Types of character recognition [11]

Also, it is difficult to recognize scripts such as Chinese and Japanese in computers as these scripts have a large number of alphabets. It is also difficult to input data for computers for scripts like Devanagari and Gurmukhi owing to their complex type nature.

Table 1.1 Comparison between online and offline handwritten characters.

No.	Comparisons	On-line characters	Off-line characters
1.	Availability of strokes	Yes	No
2.	Data Requirement	Samples/second	dots/inch
3.	Way of writing	Digital pen on LCD	Paper document
4.	Recognition Rates	Higher	Lower
5.	Accuracy	Higher	Lower

The two natural ways of communication between computers and users are – (1) through natural handwriting (2) through speech. Speech recognition has limitations in an environment which is noisy and especially where privacy of an individual is required. In present work, focus is on the problem of handwriting recognition only. A variation in handwriting is one major problem and achieving high degree of accuracy is a tough task. These variations are caused by different writing styles. Variation in handwriting occurs since each writer has its own speed of writing, different styles, sizes or positions for characters or text. Variation in handwriting styles also exists even within individual person’s handwriting. This variation may take place due to various reasons, some are as: writing in various situations that may or may not be comfortable to writer; different moods of writer; style of writing same characters with different shapes in different situations or as a part of different words; using different kinds of hardware for writing.

1.2 Handwriting Styles

Handwriting styles can also be constrained or unconstrained. Constrained handwriting, by nature is boxed discrete and spaced discrete. Unconstrained handwriting is cursive or mixed cursive in nature. In boxed discrete handwriting, each character is written inside a special box. Figure 1.3 shows boxed discrete HW.

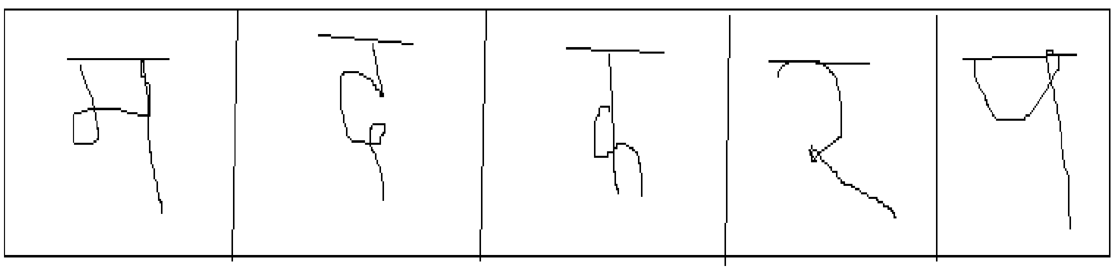


Figure 1.3 Boxed discrete handwriting

Handwriting is called Spaced discrete; when each character is written separately with spaces and no character touches other character is called spaced discrete handwriting. Handwriting is called Run-on discrete if each character is written separately and it also touches other characters, it is referred as run-on discrete handwriting. A Cursive handwriting is when characters in one word are connected and strokes are used more than once in individual character. In aspect of mixed cursive, it is observed that most of the people write in mixed cursive style that include mixture of spaced, run-on discrete and cursive styles handwriting. Spaced discrete, run-on discrete, cursive and mixed cursive handwriting styles are illustrated in Figure 1.4 below

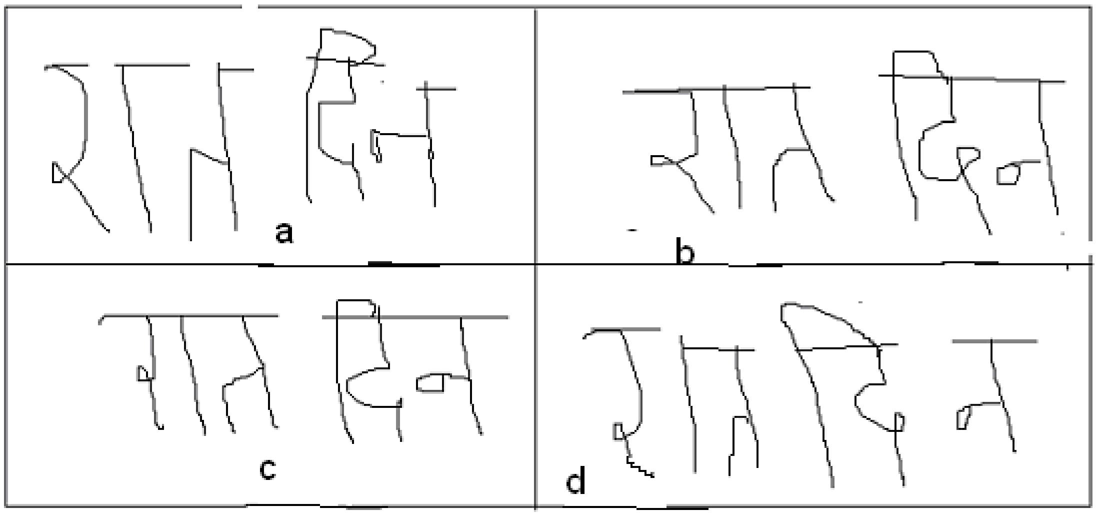


Figure 1.4 various handwriting styles. (a) (b) (c) (d)

(a) Is spaced discrete, (b) is run on discrete, (c) is cursive and (d) is mixed cursive.

It is a difficult task to recognize cursive handwriting due to great amount of variability. The online handwriting recognition has great chances to improve communication between user and computer. Due to different handwriting styles and distortions produced by the digitizing process, even the best handwritten character recognizer is not reliable. The online handwriting recognition technology is used for identification of characters and it is used with devices such as personal digital assistant, cross pad and tablet PCs where a stylus (pen) is used to handwrite on a screen, after which the computer converts the handwritten text into digital text. In order to use these input devices, accuracy achieved by the handwriting recognizer must be sufficiently high so that it is acceptable by the user.

1.3 Commercial Products Based on Pen Computing

Handwriting recognition is widespread as an input method for personal desktop assistants (PDAs). The PDAs like Apple Newton, Go's tablet computer, IBM's ThinkPad tablet computer or the Palm series of PDAs, all of which are used for inputting handwriting, were not quiet successful to popularize handwriting based input over keyboard interfaces. The software providing HCR used in each of the above said devices includes the Newton OS 2.0 for the Apple Newton, Go's Penpoint operating system, IBM's handwriting recognition and IBM's Pen for OS/2.

A modern handwriting recognition system can be seen in Microsoft's version of Windows XP operating system for Tablet PCs. A Tablet PC is a special notebook computer that is outfitted with a digitizer tablet and a stylus, and allows a user to handwrite text on the screen. The operating system recognizes the handwriting and converts it into typewritten text. Windows Mobile OS for PDAs employs a less advanced handwriting recognition system. In recent years, several attempts were made to produce ink pens that include digital elements, such that a person could write on paper, and have the resulting text stored digitally. The success of these products is yet to be determined. Although handwriting recognition is an input form that the public has become accustomed to, it has not achieved widespread use in either desktop computers or laptops. It is still generally accepted that keyboard input is both faster and more reliable. As of now, there have been many PDAs which offer handwriting input, sometimes even accepting natural cursive handwriting, but accuracy is still a problem, and some people still find even a simple on-screen keyboard more helpful and efficient.

Below Figure 1.5 shows some available pen based input devices.

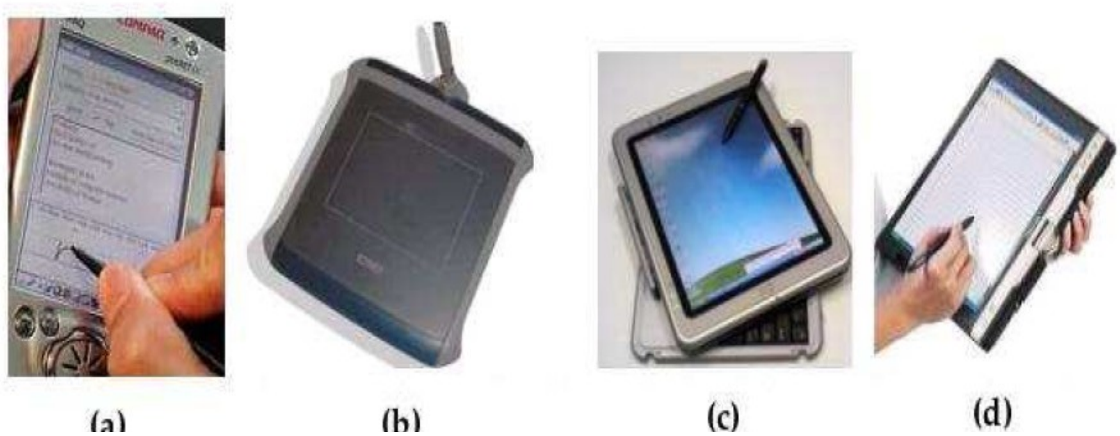


Figure 1.5 pen based input devices.

1.4 Introduction to Devanagari Script

The character sets and the large number of characters differentiate Indian scripts from other writing systems. Indian writing systems basically are derived from the Brahmi script, a phonographic writing system.

1.4.1 Devanagari Writing System

Devanagari is the most widely used Indian writing system and is used for writing Sanskrit, Hindi, Marathi, Sindhi, Bihari, Marwari, Konkani, Bhojpuri, Nepali and Nepal Bhasha. Devanagari script is a logical composition of its constituent symbols in two dimensions. It has eleven vowels and thirty three simple consonants. A horizontal line is drawn on top of all characters which are referred to as the header line or shirorekha. A character is usually written such that it is vertically separate from its neighbors. Devanagari script has many multi-stroke characters. Each character can be read and written in an unambiguous manner. Below Figure 1.6, 1.7 and 1.8 represent different aspect of Hindi.

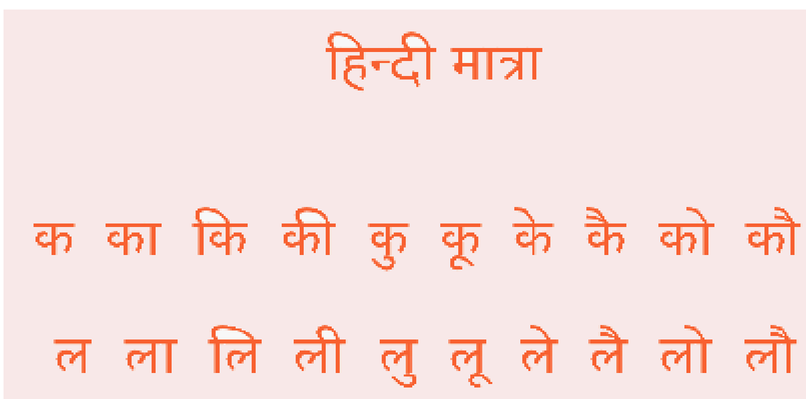
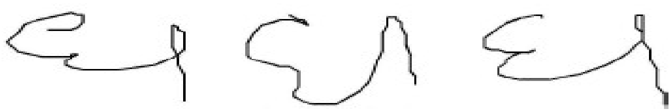


Figure 1.6 Hindi matras [34]

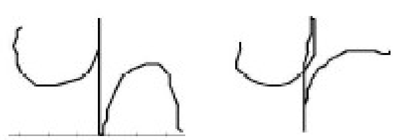
अ	आ	इ	ई	उ	ऊ
ऋ	ए	ऐ	ओ	औ	
V					
क	ख	ग	घ	ङ	
च	छ	ज	झ	ञ	
C					
०	१	२	३	४	
५	६	७	८	९	
Numbers					
का	कि	की	कु	कू	कृ
के	कै	को	कौ		
CV					
गिं	तिः	बॉ	कॉ		

Figure 1.7 Example of different forms of Devanagari characters [8]



(a) dha (b) dya (c) gha

Similar characters that looks alike.



(a) pha (b) half pha

Constant & its modifier look alike.

Figure 1.8 Different representations of character [8]

CHAPTER 2

ONLINE HANDWRITTEN RECOGNITION SYSTEM

2.1 Issues in Online Handwritten Character Recognition for Indian Writing Systems

The On-line handwriting recognition is especially important in the Indian scenario, since it eliminates the need to learn complex key stroke sequences and much faster than other text input mechanisms for Indian languages. Due to the complexity of entering the Indian scripts using a keyboard, handwriting recognition has the potential to simplify and hence revolutionize data entry for Indian languages.

Recognition of Indian script characters poses quite challenges due to style and nature of Indian writing systems. The structure of Indian script characters is much more complex than any other writing systems and it also varies with the context of its co-occurrence with a modifier unit. A character can be written using one or more strokes. Here a stroke is defined as a movement of a writing instrument on a writing surface. Some characters have common constituent stroke units. Hence it involves less number of classes to represent the character set in Indian scripts when using strokes compared to using character units.

2.1.1 Liable to vary

The recognition system should be able to distinguish between structural or shape variations across similar characters and the natural variability that exist when the same character is written by different persons or same character written at different times. The feature extraction and stroke classification address this issue to identify an appropriate representation and methods of classification. There are many examples in Devanagari script to show shape variations [10].Below Figure 2.1 shows shape variations.

Figure 2.1 Shape variations [10]

2.1.2 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 [10].

2.1.3 Presence of vertical appendages of modifiers

Presence of consonant and vowel modifiers affects the horizontal and vertical extent of characters. As a result of this, there is a significant variation in height in Devanagari characters which have vertical appendages of diacritic strokes.

Identification of an appropriate normalization factor for character and stroke sizes is necessary for maintaining uniformity in representation across varying context of occurrence [10].

2.1.4 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 [10].

2.1.5 Variation of the number and writing order of strokes in multi-stroke characters

The variation in the number of strokes is addressed using appropriate rules for character identification. The writing order of strokes is addressed by making the application of rules independent of the order of strokes in the character [10].

2.1.6 Arbitrary pen-lifts in the course of writing a stroke and stray marks

These add to confusability and adversely affect the performance of the recognition system [10].

2.1.7 Character insertions to the left of an already written character

This is an issue in Indian writing systems that are written from left to right. Such an issue never arises in offline character recognition which uses spatial information only [10].

2.2 Steps for Online Handwriting Recognition

The established procedure to recognize online handwritten characters includes following phases or components: data collection, preprocessing, feature extraction or computation of features, segmentation, recognition and post-processing .The output obtained from one phase becomes input for the next phase. Pre-processing involves various steps, normalization, interpolation, smoothing, slant correction and resampling of points. Normalization provides a tremendous reduction in data size. Thinning subpart of normalization extracts the shape information of the characters. Smoothing is done to remove noise. It is basically required to remove jitter in handwriting. Interpolation is used to find the missing points in the image. The slant of handwritten texts varies from user to user. Slant removal methods are used to normalize the all characters to a standard form which are bending in left or right direction. After pre processing, segmentation is done. It is used to divide an image of sequence of characters into sub images of individual symbols. It is a critical phase of the single word recognition process. Segmentation is done in two ways, straight and recognition segmentation. Straight segmentation decomposes an image into sub image, each sub image corresponds to single character. Recognition based segmentation decomposes each character into number of subimages.Third is feature extraction, it is done to extract the features of the character.

Features are basically two types, high level feature and low level features. High level features are long range features. Low level features are local, point oriented features. These are extracted from information of proximal data points around a data point. Features extracted give the structural characteristics of character. These features are given as input to classifier, which finally recognizes the character. Recognition is to identify the sequence of encoded values of character or word.

These phases are illustrated below:

2.2.1 Data collection

Data collection is the first step in online handwriting recognition that collects the sequence of coordinate points of the moving pen. A pen includes two actions, Pen Down and Pen Up. The connected parts of the pen trace between Pen Down and Pen Up is called a stroke. These pen traces are sampled at constant rate, therefore these pen traces are evenly distributed in time. The common names of electronic tablet are personal digital assistant, cross pad (pen tablet) and tablet PC.

2.2.2 Preprocessing

The preprocessing serves many purposes. They are used to remove noisy artifacts from handwriting and also to correct imperfections. These noisy artifacts include irregular size of text, missing points during pen movement collection, jitter present in text, left or right bend in handwriting and uneven distances of points from neighboring position. Basically preprocessing is aimed at enhancement of the classification accuracy of the recognition module. Preprocessing consist of steps as given below-

1. Normalization.

2. Sampling.

3. Resampling.

4. Slant correction.

Size normalization depends on how user moves the pen on writing pad. Centering is required when pen is moved along the border of writing pad. High speed of handwriting may result into missing points which makes identification of character difficult. These missing points can be interpolated using various techniques. Smoothing of input handwriting is required to remove jitter in handwriting. Smoothing usually averages a point with its neighbors. Slant correction and normalizing slant is required to correct the shape of input handwritten character which is bending in left or right directions. Resampling of points refers to the points in the list to be equidistant from neighboring points as far as feasible. It means that new data points are calculated on the basis of the original points of list. These processes are followed in order to make the identification process of characters simpler in nature and also character can be recognized more accurately and with less complexity. Below Figure 2.2 shows preprocessing steps.

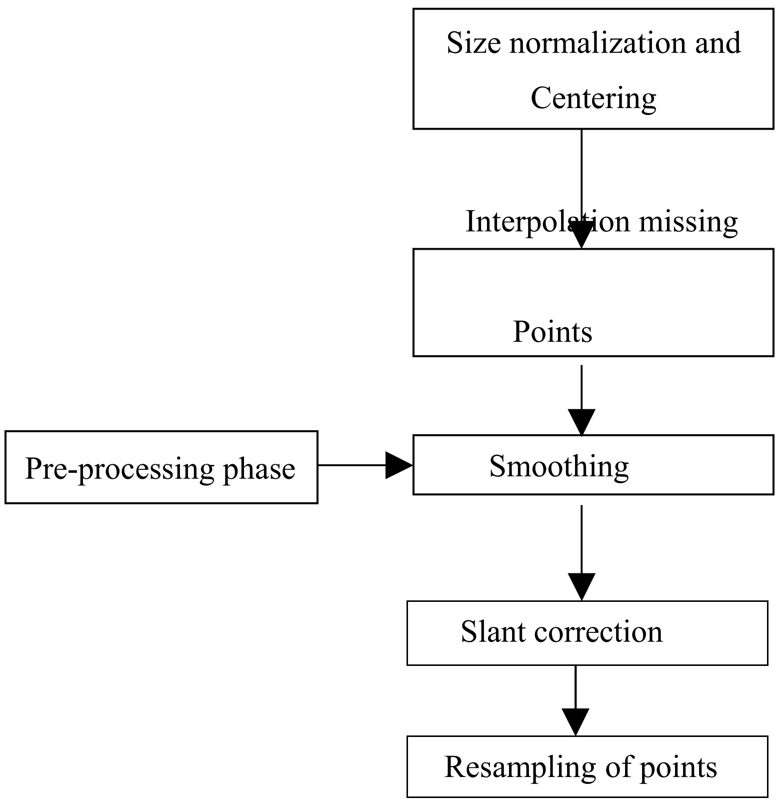


Figure 2.2 Preprocessing steps.

Preprocessing is done –

1. To remove noise.
2. To correct imperfections in handwriting data.

3. To normalize handwritten data.

4. To preclassify delayed strokes.

5. To segment into smaller meaningful units.

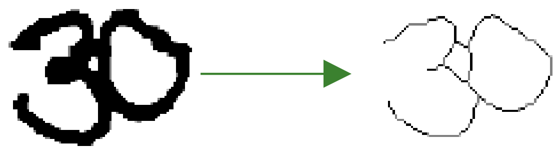


Figure 2.3 Pre-processing.

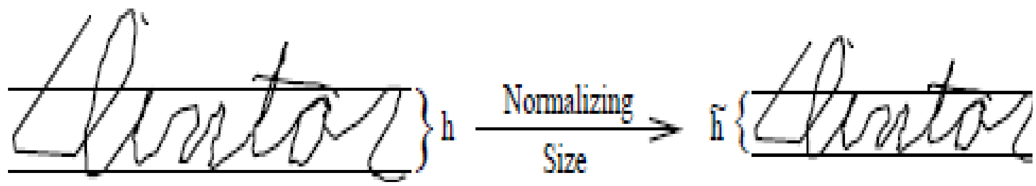


Figure 2.4 Normalizing size [20]

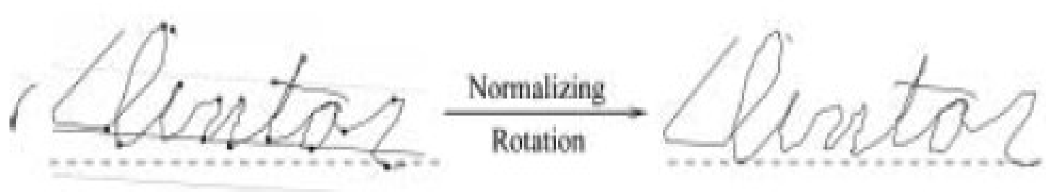


Figure 2.5 Normalizing rotations [20]

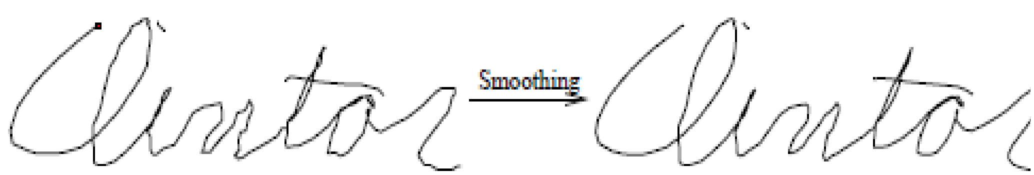


Figure 2.6 Sampling [20]

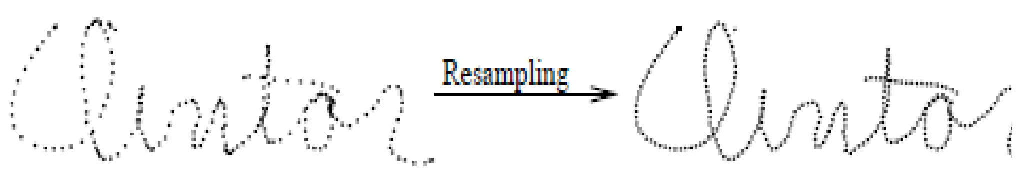


Figure 2.7 Resampling [20]

2.2.3 Segmentation

It is used to divide an image of sequence of characters into sub images of individual symbols. Character segmentation is a main requirement that determines the profits of conventional systems. Different methods exist based on the type of text and strategy being followed. Two of these methods are straight segmentation, recognition-based segmentation. Below Figure 2.8 shows segmentation of character.

Segmentation method has the following properties:

1. Capture perceptually (important groupings or regions) global aspects of the image
2. Segmentation methods should run at speeds similar to edges detection or other visual processing techniques (meaning nearly linear time and with low constant factors).

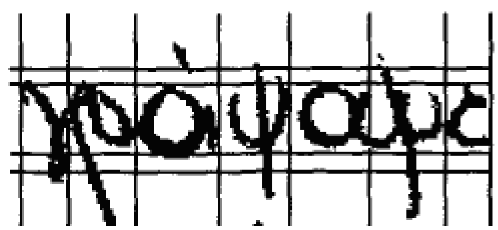


Figure 2.8 Segmentation [35]

2.2.4 Feature extraction

Feature extraction is a very important step in the process of character recognition. The features extracted from the character should encode the local, global and the structural characteristics of the character shape.

Features are classified into two categories-

1. High level features.
2. Low level features

A few neighboring points estimate low-level features where as high-level features are estimated on larger scale. High-level features are those which provide useful information such as loops, crossings, headline, straight line and dots. These high level features are derived on the basis of calculating low-level features such as directions, positions, slope, area and slant *etc* in a stroke.

a. Point oriented features

When the sequence of positional co-ordinates is used for stroke representation; there is no explicit feature extraction step involved. Alternately, a subset of positional coordinates may be selected based on the presence of local structural features in their vicinity. Such points are commonly called as critical points. Local features are computed at multiple points in the image.

b. Global features

Global features combine long range pattern information into a single feature value

c. Comparison of point oriented features and global features

Local features are easier to compute than global features. A comparison is given between Global and local features in the table 2.1 below.

Table 2.1: Comparison between local and global features.

LOCAL FEATURES	GLOBAL FEATURES
Low level, local ,point –oriented features	High level, long range features
Extracted from information of proximal data points around a data point.	Since small distortions can affect global features ,global features are sensitive to natural variations associated with handwriting.
Does not contain information characterizing the stroke as a whole.	

2.2.5 Recognition

Recognition is basically identification of the sequence of encoded values corresponding to the handwritten characters and words in the text. Classification is usually performed at character or word levels. Statistical, syntactical and structural, neural network are the common handwriting recognition methods. In statistical approach, each pattern is represented in terms of features and is viewed as a point in dimensional space. Structural and syntactical methods are related to handwritten patterns where structures and grammar are considered. Neural networks can be viewed as parallel computing systems consisting of a very large number of simple processors with large number of interconnections.

2.2.6 Post Processing

Post-processing refers to the procedure of correcting misclassified results by applying linguistic knowledge. All the possible outcomes of an individual character are studied in terms of graph and the best suitable nature of character is depicted.

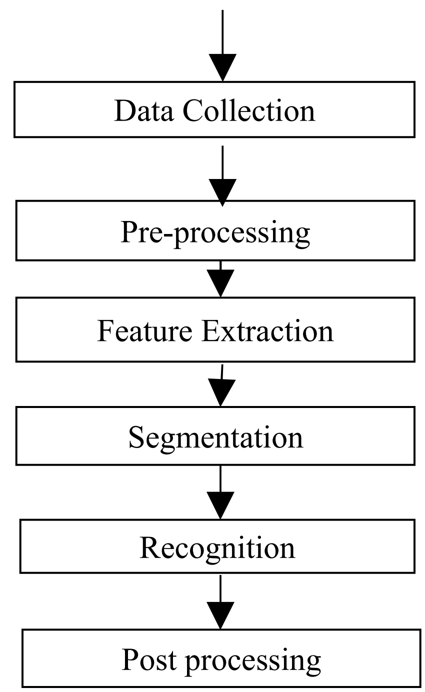


Figure 2.9 Online handwriting recognition phases. [11]

3.1 Preprocessing

Preprocessing includes five common steps, namely, size normalization and centering, interpolating missing points, smoothing, slant correction and resampling of points.

3.1.1 Normalization

Normalization is for reducing variability in character size and pen velocity It can be done by dividing both x and y-dimensions of the stroke by the ‘height’ of the horizontal block. The strokes are then converted onto curve length base and then smoothing is done independently along t-axis using a Gaussian filter [1].

a. Thinning

It is a morphological operation which is used to erase some selected foreground pixels from binary image, like erosion. It can be used for several applications, but is particularly useful for skeletonization. Most commonly it is used to make neat and in order the output of edge detectors by reducing all lines to single pixel thickness. Thinning is normally used on binary images, and produces another binary image as output. Below Figure 3.1 shows thinning.

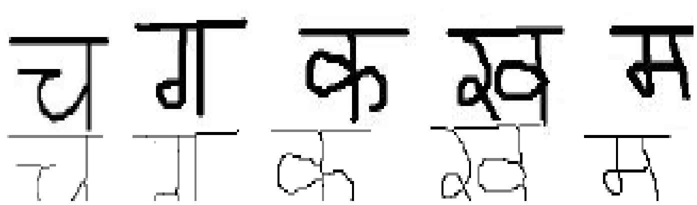


Figure 3.1 Thinning [3]

b. Skeletonization

It is used for reducing foreground regions in a binary image to a skeletal (outline) that largely preserves the extent and connectivity of the original region while destroying most of the original foreground pixels.

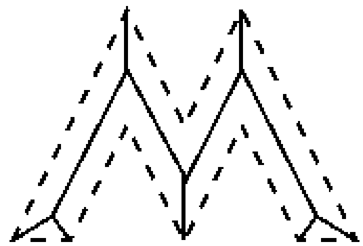


Figure 3.2 Skeletonization [8]

Deepu, Sriganesh and Ramakrishnan. [13] presented a method which uses Principal component analysis. The input from the digitizer is a sequence of points of the form xi, yi. So, pre-processing is required to compensate for the variations in both time and scale. To remove variations in time and scale, pre-processing can be distinguished into two steps, smoothing and normalization.

3.1.2 Smoothing

Smoothing is done to remove noise. Every direction which not makes a block or streams of the same direction value, is considered as a noise, and must be change. It is basically required to remove jitter in handwriting. It can be said it is required to reduce the high frequency noise in the input. Generally, most of the content in the signal is in low frequency. In one scheme, each stroke is smoothed independent of each other using a 5-tap Gaussian low-pass filter. Special care is to be taken to preserve the end points (for elimination of variability, which is occurring due to translation) and to compensate for size differences, rescaling and centrization of character is done [13].

In freeman's direction extraction, smoothing is done by comparing each code with previous and next code. In freeman's direction extraction, an algorithm is used for smoothing the image. The directional features are main consideration in this algo [2].

3.1.3 Resampling

It is performed to obtain a fixed number of points for all characters which are spaced in uniform time (in the input data). For each character, the total length of trajectory is calculated by adding the Euclidean distances between successive points. This trajectory length is divided by the number of intervals required after re-sampling to get the required spacing between successive points in the data which is resampled.The original points are substitute by a new set at this fixed spacing using piece-wise linear interpolation. All training characters which have same number of strokes are taken as a set. The number of points of each stroke is made directly proportional to the average length of strokes obtained from the corresponding set. The result of pre-processing is new set of x and y co-ordinates [13].

3.1.4 De-hooking

It is also part of pre-processing.De-hooking is basically a process to remove the hooking (bent into a curved shape) stroke that may occur during the movement of pen up and down on tablet. Hooking creates problem in the detection of original character. These are mostly found at start and end of stroke.So, it is necessary to remove them. Hooks which are occurring at start and end of stroke are removed through chain code method. Hook is identified on the basis of variation in chain code. If variation in the chain code either at the start or end is less then the standard threshold, then that part is considered as a hook and is excluded by either discarding it or replacing their co-ordinates with the neighboring ones. Hooks not depend on whether writer is experienced or inexperienced. Slant correction is done to normalize the different slants of characters.

3.1.5 Interpolation

It is used to find missing points in characters. Bresenham's line algorithms can be used to estimate the missing intermediate points. Below Figure 3.4 shows interpolation.

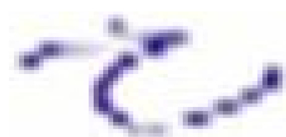


Figure 3.3 Missing points due to speed of writing [14]

Figure 3.4 Interpolation of missing points [14]

The **Bresenham line algorithm** is an algo that determines which points in an n-dimensional raster (graphic image) or bitmap image should be plotted in order to find a close approximation for a straight line between two points. This algo is one of the earliest algo developed in the field of computer graphics. Smoothing is used for data filtering.

It consists of replacing the coordinates of the original point by the use of its neighboring points. In this one, smoothing was applied on the chain code of the stroke as mostly chain code is used to extract features in place of x, y co-ordinates [14].

3.2 Feature Extraction

In the process of handwriting recognition, it is important to identify correct features. Feature extraction techniques have been introduced by various authors.

3.2.1 Structure based feature extraction

One of the feature extraction methods is structure based. One paper presents the structure representation of Tamil language which is also Indian script. Tamil is the most widely spoken language in south India.

K.H, Vidhya, M.kasiranjan, G.Vijay, V.S and Sriganesh. [1, 23, 24] focused on spatial proximity. Closeness of strokes is used to extract features. First pre processing is done. After pre processing, features are extracted on the basis of dot, line terminals (ends of stroke), bumps (tangent exists), cusp. On the basis of these features, stroke is identified. For the identification of unknown stroke, stroke is compared with database of known stroke, this process is called matching. Matching is done on the basis of shape and relative sizes of two stroke pair's. Lastly, character is recognized. In this, first the entire stroke labels corresponding to single block (horizontal) is combined. Then these horizontal blocks are categorized into five categories-

VM=vowel modifier,

V=vowel, X=consonant, Y=consonant vowel (with the help of search table).These categorized blocks are present as input to finite state automata. It identifies character code and signals the character termination. Data of 2000 stroke samples from 15 users was taken, having 96 stroke classes. This achieved 86.1% performance. In future, manual analysis can be automated and stoke labeling can be done by clustering or can say by grouping the strokes. A finite state automaton is used to fully identify the character.

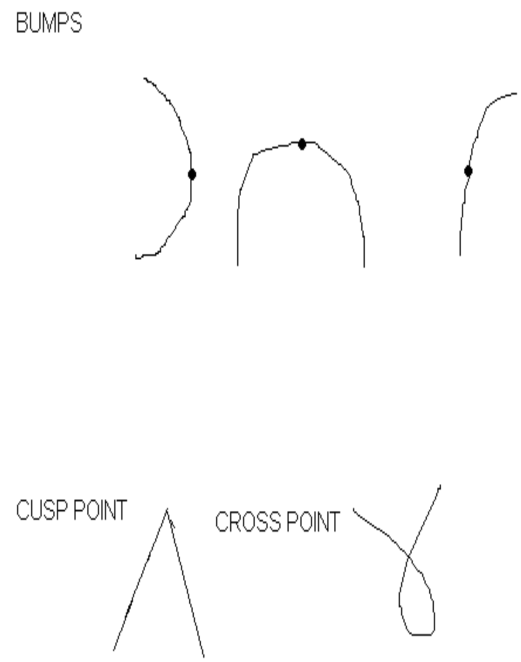


Figure 3.5 Character features [1]

For Devanagari characters four feature extraction techniques can be applied namely, intersection, shadow, chain code histogram and straight line fitting algo. S.Basu, N.Das, R.Sarkar, M.Kundu, M.Nasipuri and D.K Basu [27] proposed shadow features that are computed globally are basically the length of the projection on the sides but on the whole image not on parts of character. Intersection features are the pixel point which has more than two neighboring pixels in 8-connectivity while an open end has exactly one neighbor pixel are computed by decomposing the character image into segments. Also, chain code histogram features (Chain code provides the points in relative position to one another, independent of the coordinate system) and line fitting features are computed by decomposing the character image into different segments feature [3].Below Figures show features extracted.

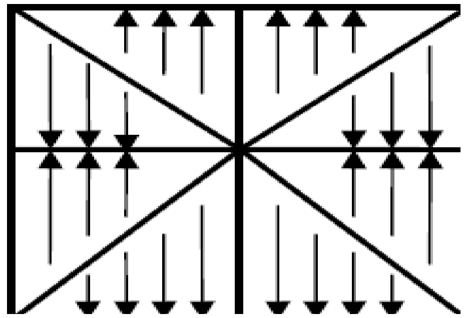


Figure 3.6 Shadow Features [3]

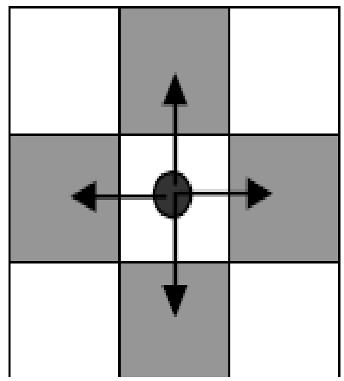


Figure 3.7 Chain coding (4-connectivity) [3]

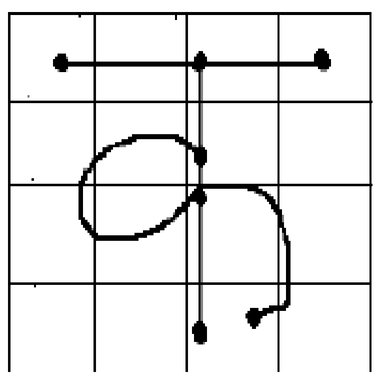


Figure 3.8 Intersection and end points [3].

A method which is independent of size and can extract features from the raw data without resizing can also be used for future extraction**.** The technique contains many characteristics of handwritten characters based on structural, directional and zoning information and merges them to create a single global feature vector [5].Below Figure 3.9 shows proposed technique.

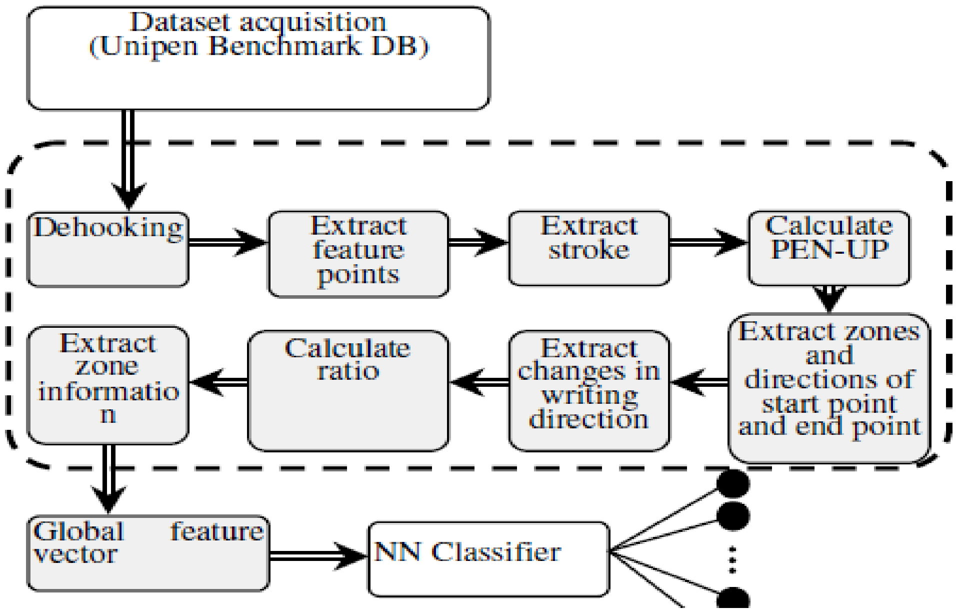


Figure 3.9 Block Diagram of Proposed Technique [5]

A feature based spatial analysis technique can also be used for feature extraction to analyze the structure of mathematical expressions. For mathematical expressions relative positioning and size of one mathematical symbol to another must be taken care for recognition. Also the relationships must be constructed in a manner that shows a proper mathematical expression. So a method be used which can identify the spatial relationship between the symbols. One of the structural techniques used to solve the problem involves the use of bounding box for spatial analysis. But alone it was not so effective.

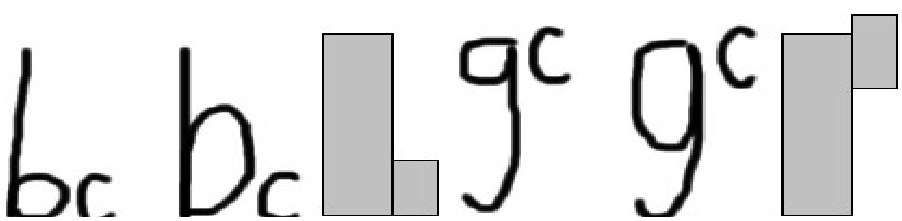


Figure 3.10 Bounding box problems [6]

In the first pair of b and c, relationship is multiplication and in second pair it is subscript relationship. But the bounding boxes are identical for both symbols. This shows that extra care is to taken for implicit relationships.Also, the issue of scalability is not addressed through this method. Many researchers have worked on this problem. Chain and Yeung [6, 17] proposed a method based on definite clause grammar (DCG). It is used to define a set of replacement rules for parsing mathematical expressions. But they did not address the issue of ambiguity resolution, error detection or error correction.

Kosmala and Rigoll. [6, 18] presented a system that is having a benefit of simultaneous segmentation and recognition of symbols based on hidden markov model.

Zanibbi given a tree transformation based method. In this, recursive search identifies linear structure in an expression and produces baseline structure tree. [6, 19]

Feature based fuzzy rule technique is combined with symbol recognition and parsing technique to develop an expression tree for math's expressions. The symbol recognition software used, utilizes fuzzy decision making. Fuzzy logic is used to overcome the imprecision in online handwriting. Information gathered during the symbol recognition phase regarding the features of various symbols is used to determine the limits of fuzzy functions that in turn give relationship confidence. These relationships are investigated on the basis of vertical and horizontal confidence. In some cases relative size is also used to weigh the decision. These relationships are represented by expression tree through semantic analysis .The undesired relationships and regressive entries are detected through expression tree [6].

a. Freeman's chain code

It divides the character into 8 directions values according to point sequence. During preprocessing, these 8 direction values are used to make chain code shorter by removing values which are identical to the preceding ones. It also considers pair of points between pen down and pen up and assigns them directional value. After this, features are extracted by comparing number of codes with threshold. In this first horizontal arcs are find, then vertical arcs, after this line. This priority order increases standard of extraction. Lastly matching is done on the basis of set of models defined for each number. Similarity function is used for matching. The function is used to select model having first primitive from left, similar with shape's first primitive. Data from 10 persons was collected. Here recognition ability is 94.87%.The drawback is that structures of numbers having two types of primitives are difficult to recognize. The advantage is that not any ambiguities among number's. Models and Training is not necessary, so any new model can be easily add any time [2, 7, 25].

b. 12 directional features

One of the techniques for feature extraction is to calculate only twelve directional feature inputs depending upon the gradients. Features extracted from handwritten characters are directions of pixels with respect to their neighboring pixels. After pre processing (noise removing, skeletonization and normalization) gradient of each pixel is computed by using sobel's mask to calculate horizontal and vertical gradient components. These gradient values are mapped to 12 directional values. These 12 directional values of each pixel are given as input to neural network. This increases recognition accuracy as lot of information is available (12 direction feature).It also takes less training time. More advancement can be made by including 12 directional input feature extraction technique, so that special character can also be recognize accurately These inputs are given to a back propagation neural network with one hidden layer and one output layer [7].

c. Star feature method-A star feature method for recognition of online handwritten characters can be applied. The star feature encodes the local, global and structural characteristics of a character. The star feature describes every point of the character, in terms of its relative position with respect to the other points in the character. The experimental results show that the star feature achieves high accuracy [4].

The [4, 26] also taken into account the role of direction pixel and It is also based on Star feature method by encoding of relative position of point with respect to other points of the character. In this calculation in 8 direction is done to find the relative position of given point with respect to other points. First feature vectors fi is calculated for each point in 8 bit binary vector. A point is set if there is a point of intersection of the 8 directional lines with the other parts of the characters. If a point is set, 1 is assigned for that point in 8 bit binary vector, otherwise 0. After this, 8 bit vector for each point is concatenated to form feature vector. The Performance is evaluated on IRONOFF dataset and Tamil dataset .Nearest neighbor classifier is used to calculate accuracy of the recognition. The evaluation on the IRONOFF dataset is done by random division of the data set into three folds and then calculation of the three fold cross validation accuracy. Star feature has 84.9 % accuracy on lower case, 88.6% on upper case characters and 93.5% on numerals and 80 % on Tamil. The benefit is that, it takes less memory to store features as binary feature vector is used. In future the research will be done to extract structural information for complex shapes like loops, cusps. The performance will be evaluated with other classifiers also. Below Figure 3.11 shows directional coding

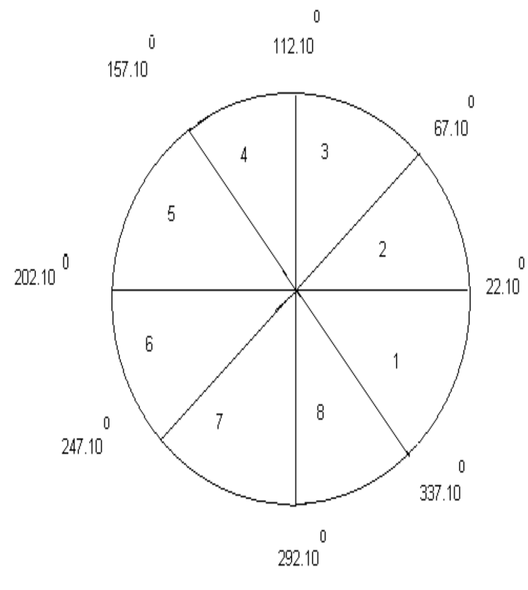


Figure 3.11 Directional coding.

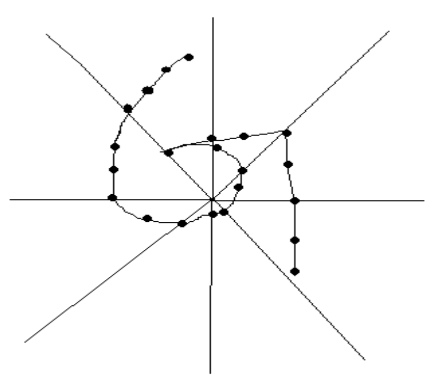


Figure 3.12 Intersection with points.

One paper reveals character based elastic matching technique using local features for recognization .Dynamic Time Wrapping is used with four feature sets:

1. x, y features,

2. Shape Context and Tangent Angle features,

3. Generalized Shape Context feature.

4. x, y normalized first and second derivatives and curvature features.

Preprocessed co-ordinate- The raw coordinates are preprocessed and the preprocessed (x_i) and (y_i) are used as the features.

Shape context-The second is given by S.Belongie, J.Malik and J.Puzicha [28] and it is calculated at every point of stroke. For each point, log of the distance of the point from remaining points (r) and the angle of slope of the line joining the point to every other points (θ) are calculated .Distances (r) are normalized through mean distance to make the features invariant. Finally at each point pi a histogram $h^{\wedge i}$ is built. It is a partial local feature describing the relative position of the whole character with respect to the considering point. Distance $\chi 2$ is used to find the distance between two points using this feature. Below Figures 3.13 and 3.14 shows Shape context features.

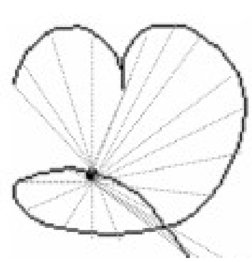


Figure 3.13 Distance from particular point to other points [22]

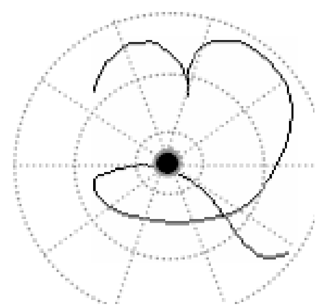


Figure 3.14 SC features at the considered point [22]

TA- At each point p the tangent slope (θ_i) is calculated. The measure of the difference between local tangent angles at two points p_i and p_j is also calculated.

Third is a Global Shape Context feature. It is extended version of above feature. In addition to the Shape context, at each point, the unit length tangent which defines the direction of the edge at that point is calculated.

Normalized Derivative-With respect to x, y,the normalized first derivatives are calculated at each point p_i .

Curvature feature-In case of plane curve, the curvature at each and every point has the magnitude (degree) equal to the reciprocal of the radius of a circle that touches closely the curve at the given point [22].

3.2.2 Dominant point

A dominant point also plays an important and efficient part in extraction of features. Dominant points are commonly considered as points with local maximum curvature or can say elevated position. In the Euclidean plane, curvature can be defined as the rate of change of slope as a function of arc length. It is basically, the amount by which a geometric object deviates from being flat, or straight in the case of a line.

One of the simplest approaches to detect dominal points is by constructing a variable that is ratio of the height to the width of an imagined rectangle, with bottom coinciding with the polygon enclosed by the movement of pen trace. Area of rectangle is equal to the polygonal area.Dominal points are detected when value of given variable exceeds given threshold. Dynamic programming can be used for this purpose but it involves lot of complexity. Also, it cannot respond to pen movement in real-time. Only when all the points representing stroke are captured, this detector performs work. As the sampling intervals are not utilized to conduct computations, respond of the entire system will get delayed. In new method, while pen is moving, dynamically computation of the polygonal area enclosed by the pen movement trace can be done. When a new point comes in, only need is to expand the polygon by adding a new triangle so the cost in computing the new polygonal area is low. As it is known that the high curvature near a turning point will result in a sharp increase of the polygonal area, a variable is constructed which is the ratio of the polygonal area to the square of the corresponding chord, which links the start point and the incoming point. Value of this variable goes on changing according to new point. The computational cost for this search is low and turning points can also be detected in real-time while pen is moving. In simple words variable can be understood as the ratio of the height to the width of an imagined rectangle and whose bottom coincides with the chord of the polygon, and area is equal to the polygonal area. The height of this imagined rectangle reflects the fluctuation amplitude (move up and down) of the pen movement trace. Width represents the scale of the trace.

If the ratio of fluctuation of amplitude versus the scale is small, it means that such a fluctuation does not represent a original turning point. This method can detect dominant points more accurately and hence this is robust process. The computations are less in this method, so computation can be accomplished during the sampling interval. As each new point comes in, a triangle area only need to be calculated, this involves only a few of points. Therefore, the computational costs and also the storage requirements of this method are very low. So the computational cost of the given algorithm is far less than of the Dynamic Processing method [15].Figures 3.17 represent the imagined rectangle and Figure 3.18 represent dominant point detection,

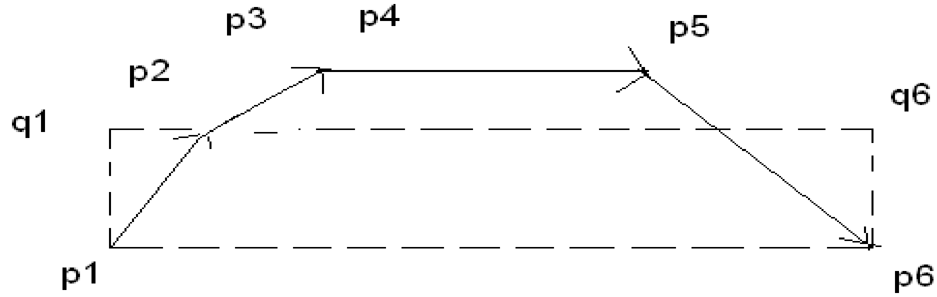


Figure 3.15 a polygon with the imagined rectangle [15]

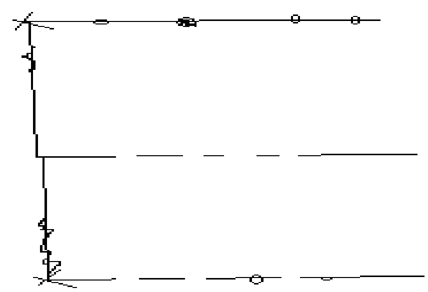


Figure 3.16 Dominant point detection of online script [15]

Method of dominant points can also be applied to alphanumeric characters. To design real time recognizer and reliable recognizer, fast algorithm for feature extraction is require, as feature extraction is very crucial step for fast and efficient recognition. A method for feature extraction must be simple and computationally light. By working on dominant point to extract features, the features are robust with respect to different distortions like size, shape; orientation and noise present in the sample and hence provide high generalization capability. This feature extraction method exploits both sequential and dynamical information for on line recognition. This idea is applicable only for real time system as it uses the time dependent positions of device. Below Figure 3.19 shows feature extraction.

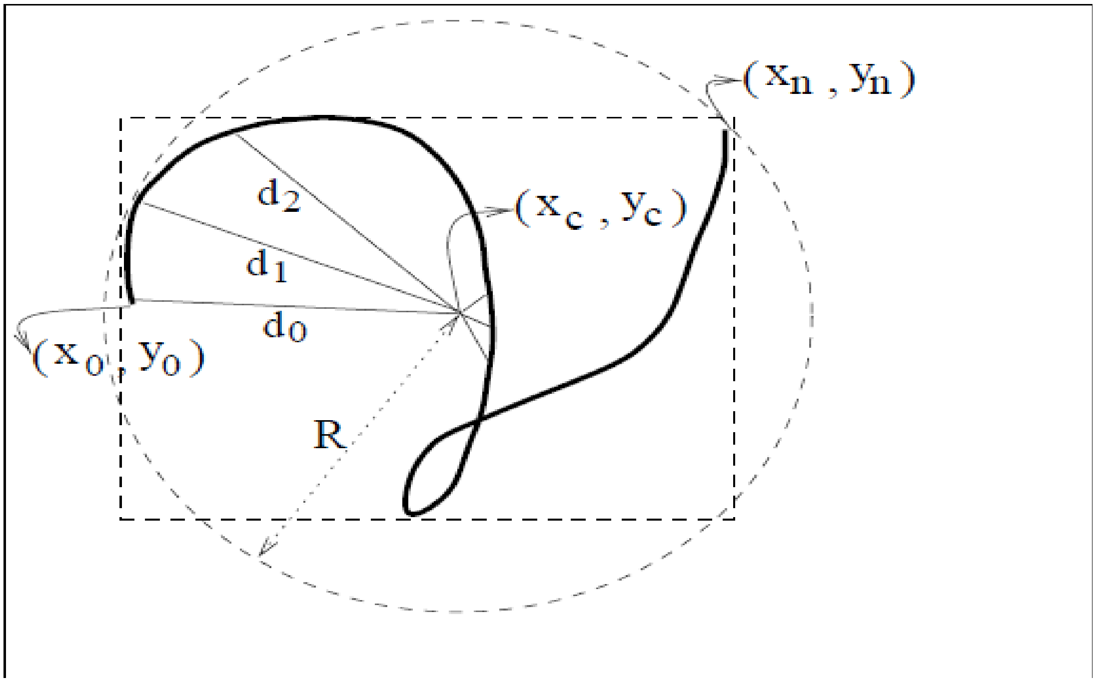


Figure 3.17 Feature extractions for a sample of character of class 2 [16]

To make the features independent of size and orientation, a modified sequence is generated by the use of following algorithm.

(a) A minimum rectangular area that cover *Zth* sample of a particular character class is found from $x_{\min}$, $y_{\min}$, $x_{\max}$ and $y_{\max}$. Here $x_{\min}$ is minimum of all *x* co-ordinates of *Zth* sample. $x_{\max}$ is maximum of all *x* co-ordinates. Similarly $y_{\min}$ is minimum of all *y* co-ordinates of *Zth* sample and $y_{\max}$ is maximum of all *y* co-ordinates of *Zth* sample.

(b) The centre of rectangular area i.e. x_c, y_c and the radius of smallest circle that would enclose the *Zth* sample character is found.

(c) A new sequence is generated from given sequence by the use of two equations from the centre. The point from series of n point is selected through $i = i * n/p$ .The value of i is successively taken as *i= 0, 1, 2, 3….p* and right hand side equation is rounded off to nearest integer value to fetch all p successive selected points. So, we have sequence of distance from centre where each distance is distance between point and centre divided by *R*. Here sequence of distance is normalized with respect to the corresponding radius. By this distance vector becomes invariant of size and orientation.

(d) After calculation of distance vector, sequence of angle is created to capture the angular movement of device and the reference line is taken as line joining starting point and centre. Other angles are found as the line joining the centre and points. Now the sequence of angle and distance vector are used as feature for the classification purpose. Fourier transform of above features are also taken as other two inputs for the next process of recognition [16].

Cho-huak, roland t.chin [20] presented dominal point detection on digital closed curve. The algorithm is compared with various other algorithms. This procedure requires no input parameter and also it remains reliable, when features of different sizes are present on the digital curve. Five dominant point algorithms are discussed here.

Rosenfeld and Johnston algo -This is angle detection procedure by Rosenfeld and Johnston [29].This procedure is comparable to the Rosenfeld thruston edge detection algorithm which finds important enough maxima in average gray-level gradients by the use of degree of smoothing. [20]

Rosenfeld- Weszka- Second dominant point technique is Rosenfeld-Weszka [30] Improved Angle Detection Procedure. After second step of the Rosenfeld-Johnston procedure, the k cosines (k > 1) at each point are smoothed by equations given below. This algorithm is less expensive as it takes extra steps of k-cosine smoothing.

$$\overline{\cos_{ik}} = \frac{2}{k+2}\sum_{j=k/2}^{k}\cos_{ij} \quad \textit{for k=even.}$$

$$= \frac{2}{k+3}\sum_{j=k-1/2}^{k}\cos_{ij} \quad \textit{for k=odd.}$$

Freeman-Davis Corner Finding Algorithm- Third procedure is Freeman-Davis Corner Finding Algorithm. [31] The angular differences between each segment positions are used as a smoothed measure of local curvature along the chain Freeman. First define L_{is} as the straight line segment spanning m chain links and terminating on the node to which link ($\overline{c_i}$) is directed: Lim = $\left\{\overline{c_i}, j = i - s + 1 \ldots i\right\}$. The x and y components of L_{is} are calculated and then angle made by *(Lis)* with the x axis is given. After this incremental curvature is defined as twice the mean over two adjacent angular differences. This is a smoothed measure of curvature. At last, corner is characterized.

Sankar-Sharma Dominant-Point Detection Procedure [32] - In this, the dominant points are computed frequentative as the points of maximum global curvature, which is based on the local curvature of each point with respect to its immediate neighbors. It is observed that each point of a closed curve having two neighbors can be differentiated into three classes based on the local curvature, as shown in Table 2:3.

Table 3.1: Local curvature assignment.

Different Types of curvature	Weight Assigned to each curvature
No curvature	0
Positive curvature	+1
Negative curvature	-1

For all the points having more than two neighbors, let *(m, n)* be the point with k immediate neighbors where k is greater than three. Find all pairs which are possible of 2-neighbor configurations of *(m, n)*. Assign to all possible pair the corresponding local curvature, using Table. From the collection, erase those pairs with zero curvature. Among the all other pairs, if pairs have been assigned positive curvature, then the point *(m, n)* is assigned the weight + 1 and if assigned negative curvature, then point will get weight -1. In some other cases, if some pairs have positive curvature and the rest have negative curvature, then the point *(m, n)* is assigned the weight 0.

Anderson-bezdek vertex detection algorithm [33] - In this basically tangential deflection and curvature of separate curves are found on the basis of geometrical, statistical properties linked with the eigen values and eigen vector structure of the covariance matrices of sample.

Teh-Chin dominant point detection- A new algorithm was constructed called TEH-CHIN DOMINANT POINT DETECTION. Two points must be considered for dominant point detection on digital curves. One is accurate definition of discrete curvature and second is the determination of the region of support for the computation of the curvature. The reliability and exactness of dominant point detection algorithm depends not only on the correct determination of the discrete curvature. It depends also on the accurate determination of the smoothing factor of each point on the basis of local properties of that region. This proposed algorithm requires no smoothing factor. The length of the chord and the perpendicular distance between chord and the point provides a basis for choosing the suitable region of support. The measure of importance of each point is determined by using the neighboring points within the extent. The measure of significance and the region of support of every point is then used to show the way for the selection of points to be removed. The points which remain after the removal process are dominant points.

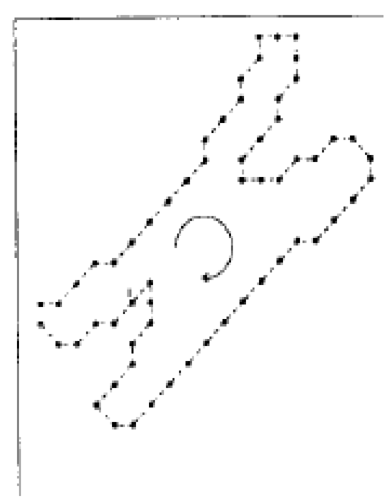


Figure 3.18 a chromosome shaped curve [20]

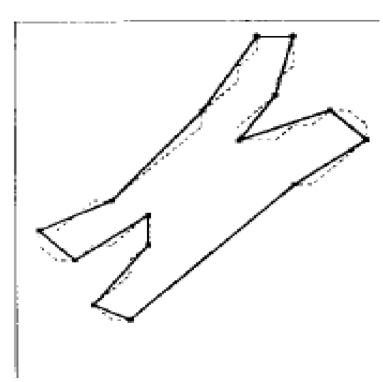


Figure 3.19 Teh-Chin algorithm (k cosine) [20]

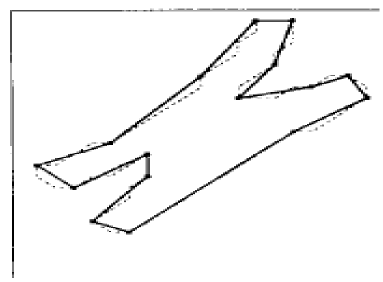


Figure 3.20 Teh-Chin algorithm (k curvature) [20].

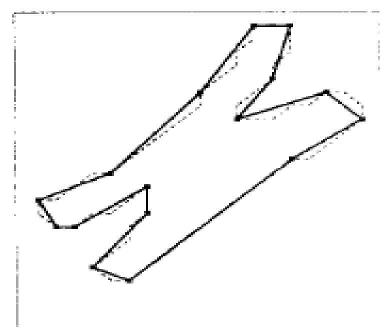


Figure 3.21 Teh-Chin algorithm (1 curvature) [20]

3.3 Recognition

This involves identification of the sequence of encoded values corresponding to the handwritten characters and words in the text. Classification is usually performed at character or word levels.

The SVM intakes a set of input data and forecasts, for each given input, out of two possible classes which forms the input, thus making SVM a non-probabilistic binary linear classifier. An SVM training algorithm makes a model which allocates new examples into one category or in the other category. An SVM model is basically a representable of the examples as points in space, mapped, to make examples of the separate categories divided by a clear gap as wide as possible. Newest examples are then mapped in same space and also predicted to classify in which category they belong based on which side of the gap they go.

A support vector machine formally can be described as hyper plane or a set of hyper planes in an infinite-dimensional space, which can be used for classification, regression etc. A good separation is get by the hyper plane which is at the farthest distance to the nearest training data point of any class, as larger the margin the lower the generalization error. The original problem may be given in a finite dimensional space, it happens may times that the sets to be discriminate are not linearly separable in finite dimensional space. So, it was proposed that the finite-dimensional space be mapped into a much higher-dimensional space, thus making the separation easier in that space. Figure 3.24 shows binary classification.

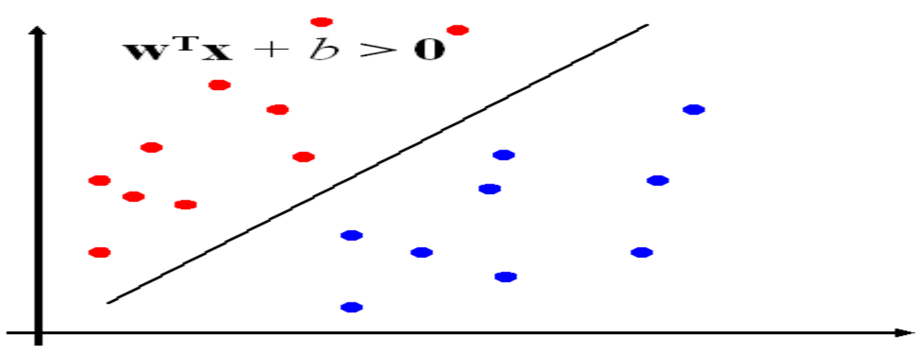


Figure 3.22 Binary classification can be viewed as the task of separating classes in feature space [35]

Basic properties of SVM are:

(a) Flexibility in selecting a similarity function.

(b) Thinly dispersion of solution when interacting with large data sets (only support vectors are used for specifying the separate hyper plane)

(c) Capability to control and properly handle large feature spaces.

(d) Overfittiness can be overcome by soft margin approach.

(e) Feature selection.

SVM has been used in many real-world problems like text (and hypertext) categorization, image classification, bioinformatics (Protein classification, Cancer classification) and hand-written character recognition.

Kernel methods are specific class of algorithms for pattern analysis, whose appropriate known element is the support vector machine (SVM). The general work of pattern analysis is to search and study general types of relations (such as clusters, rankings, principal components, correlations, classifications) in general types of data such as text etc.Four types of kernels are discussed here.

The Linear kernel is the simplest kernel function. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts.

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

The Hyperbolic Tangent Kernel is also known as the Sigmoid Kernel. The Sigmoid Kernel comes from the Neural Networks field, where the bipolar sigmoid function is often used as an activation function for artificial neurons.

The Radial basis function is also known as B-Spline function. The B-Spline kernel is defined on the interval $[-1, 1]$ [35].

A neural network based classifier can be used to recognize the extracted feature (back propagation neural network). It involves a training set of both positive and negative cases. A new sample is classified by calculating the distance to the nearest training case. The sign of that point then determines the classification of the sample. The *k*-NN classifier extends this idea by taking the *k* nearest points and assigning the sign of the majority. [5]

For recognizing single character, the points (score) for a character is calculated by finding through the states of character, an optimal alignment path and then summing the activations along this path. For word recognition, the score of each word is calculated by searching an optimal alignment path by the states of characters composing the word. The final score is developed by summing all the activations along this path. The MS-TDNN is trained with back-propagation in three steps. The first and second step of training operates in an alignment mode which can be called forced, during this MS-TDNN is trained with data whose character boundaries are known for these words. Figure 3.26 below shows training.

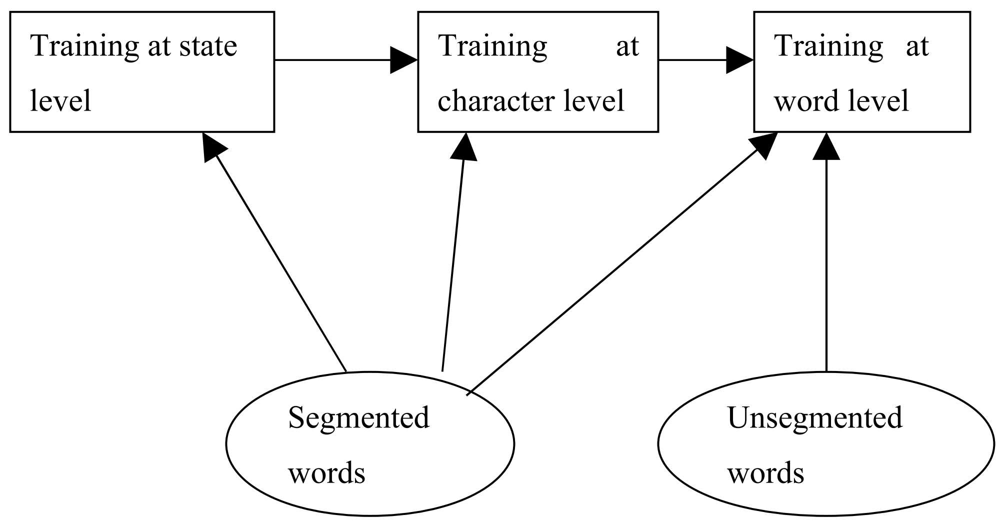


Figure 3.23 Training of a multi state time delay neural network.

Character based elastic matching technique- Character based elastic matching technique can be used for online handwriting recognition. In this, character based elastic matching is applied with local features for the recognization of online handwritten data. Here, Dynamic Time Warping method has been used with different feature sets first is x-y features, second is Shape Context, third feature is Tangent angle, Generalized shape context feature and the last one is the set having x-y, normalized first and second derivatives and also curvature features. Nearest neighborhood classifier is used with dynamic time wrapping distance, as the classifier. Elastic matching algorithm, Dynamic time wrapping is used for matching purpose. It is a method which is used to find optimal alignment in between two time series even if one time series may warp non-linearly due to stretching or shrinking of time series along its time axis. This warping of time series can be used then, to find the similarity between them [22].

In the freeman direction method, firstly smoothing is done called as freeman's direction smoothing. After pre-processing, primitive features are extracted as cusp, loop etc.Feature extraction is followed by matching process. Matching of extracted features is done with the database of features. The proposed Freeman's chain code is used to represent the directional information. For each number, a set of models will be defined. Every model should be distinct from each other, so that there should be no two models having same primitive's type and order. The similarity function is used for matching process. The similarity function selects the models, having their first primitive from the direction left in similar with the shape's first primitive. To find the accurate model from the resulting models, the similarity function does the comparison on the basis of second primitive, and eradicates unmatched models. This whole step is repeated unless more than one model exists, and all the primitives of model are matched with all shape's primitives [2].Figure 3.24 below represent freeman's direction procedure.

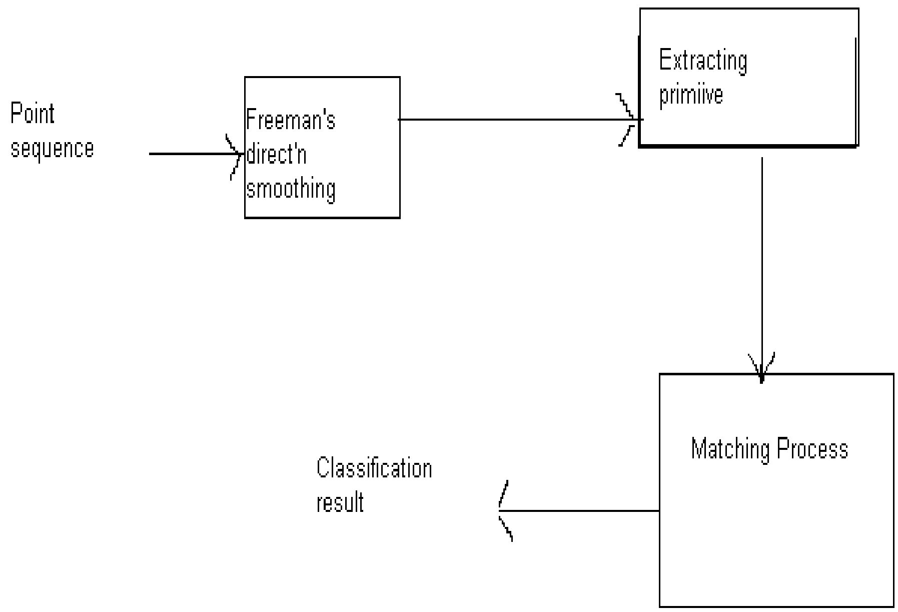


Figure 3.24-Recognition phase in freeman's direction procedure [2]

CHAPTER 4 PROBLEM STATEMENT

4.1 Introduction

On the base of literature review various features are available for recognition of Devanagari characters. An experimental methodology needs to be developed to find the performance of these features in recognition of Devanagari characters. There are in total 48 characters, 14 matras and 10 numerals in Devanagari. All of these characters can be written in various ways using various types of strokes. The focus of this thesis therefore, would be limited to various strokes used in writing matras.

The various techniques have been applied on Devanagari script like:

a) Freeman's chain code method applied on Indian numerals.

b) Multiple feature extraction techniques such as intersection, shadow, chain code histogram and straight line fitting applied on Devanagari characters.

c) Twelve directional features calculation.

d) Structural recognition techniques combined with feature mapping.

But still the recognition of Devanagari characters is complicated process. Various researchers have found dominant point and distance of dominant point from center as good features for recognition of handwriting. The features have not been tested on Devanagari characters. To achieve more accuracy in identification of Devanagari matras, dominant point and distance of dominant point from centre is used in this thesis to recognize character.

Also, Devanagari script is recognized by various methods like neural network method, elastic matching technique but support vector machine is not used till now.

As per the literature survey and gaps listed above, the focus of this thesis would be to

1. To find various strokes for writing matras in Devanagari script.
2. Capturing and identifying dominant point and distance of dominant point from center of the stroke features for each stroke.

3. Training and recognizing the stroke using support vector machine for various kernels (linear, polynomial, radial basis function and sigmoid).

5.1 Data collection

The process has been implemented on the handwritten Devanagari matras.The matras are written by single writer only. The matras are captured with the help of pen device on a writing pad. The sequential points of writing device are noted and used as the data (co-ordinates) for the feature extraction. The points noted are saved on the Microsoft excel sheet. When the writer writes slowly, the points are located densely and when writes fast, the sampled points are densely located. In present work, the writer was asked to write the 100 samples of each matra.There are total 9 matras, so all together 900 samples are created during this work. All these points go under the pre-processing phase as size normalization and centering of strokes are done, after this interpolation is done, then smoothing, slant correction and resampling of strokes is done. As we control the resampled points we get equal points for each matra.This is the first level of experiment. Figure 5.1 shows pre-processing of strokes.

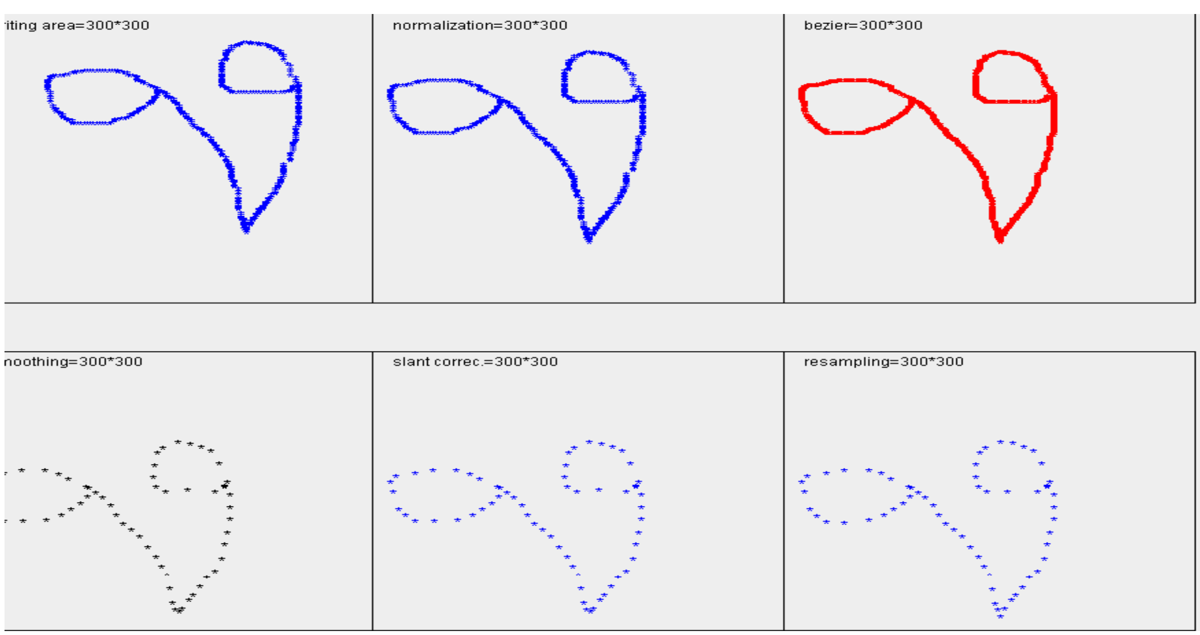


Figure 5.1

File Edit View Insert Format Tools Data Window Help

Arial 10

A1 1084

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1084	117	1086	128	1085	140	1083	146	1078	151	1073	154	1067
2	986	142	987	134	991	128	995	124	999	119	1002	116	1008
3	950	102	956	100	962	100	971	100	980	100	988	100	999
4	923	100	918	104	914	110	910	116	906	122	903	130	900
5	981	100	980	105	979	112	978	117	977	123	977	129	976
6	1063	100	1054	100	1048	101	1040	103	1031	109	1024	116	1015
7	1079	169	1078	176	1072	179	1070	174	1070	166	1070	158	1070
8	931	107	940	107	950	107	959	107	969	108	978	108	987
9	1071	125	1073	130	1077	133	1084	135	1092	133	1094	128	1094
10	1045	110	1040	103	1031	100	1020	100	1015	105	1011	110	1007
11	1094	113	1094	124	1093	135	1091	142	1087	146	1084	150	1079
12	900	114	902	108	908	104	919	102	930	101	938	100	950
13	900	109	909	106	925	104	938	101	957	100	968	100	985
14	934	100	930	105	926	109	923	114	921	120	919	126	918
15	944	100	943	106	942	112	940	118	939	122	937	127	936
16	1064	101	1053	100	1043	100	1033	101	1023	103	1014	108	1003
17	1081	117	1084	124	1086	130	1088	136	1088	144	1088	151	1087
18	900	103	911	103	923	103	935	103	947	103	957	103	968
19	1050	132	1051	139	1057	142	1064	145	1074	147	1085	146	1090
20	1046	107	1039	103	1034	101	1028	100	1023	100	1017	101	1012
21	1080	117	1079	127	1075	132	1071	136	1067	141	1060	144	1053
22	914	127	917	122	925	116	931	113	939	109	946	107	952
23	940	102	948	100	961	100	971	100	982	100	994	100	1005
24	961	100	952	103	945	109	942	116	940	122	939	130	938
25	936	100	936	109	936	116	936	125	936	132	936	140	936
26	1042	107	1036	104	1030	102	1024	101	1016	100	1010	100	1003
27	1087	138	1087	143	1086	148	1085	152	1083	156	1081	159	1076

Figure 5.2 shows data collected on excel sheet.

5.2 Analysis phase

In second level of experiment, we have implemented different partitioning strategies for collected pattern in testing and training pattern. Three strategies have been adopted here. These are (a) 50% data in training, 50% data in testing: (b) 75% data in training, 25% data in testing(c) 90%data in training, 10% data in testing. The data are dominant points of character and the distance of these dominant points from centre. In work, we have used SVM for recognition with various kernels as linear, polynomial, radial basis function and sigmoid. In this we have divided the matras into subparts. The same strokes of two different matras are represented by same numeral like

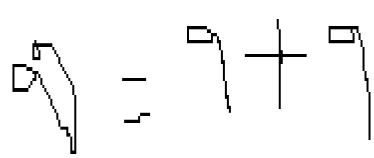
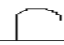
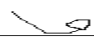
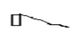
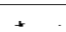


Fig 5.3 joining of two strokes

In this we have represented matras by numbers as given below in table.

Table 5.1

	1
	2
	3
	4
	5
	6
	7
	8
	9

Below Figure 5.4 shows data having dominant points and their distance from centre.

0.491668	1000	292	0.491668	1000	292	0.491668	1000	292	0.491668	1000	292	0.491668	2
0.212819	1048	155	0.212819	1048	155	0.212819	1048	155	0.212819	1048	155	0.212819	3
0.239945	1081	178	0.239945	1081	178	0.239945	1081	178	0.239945	1081	178	0.239945	4
0.418046	1093	273	0.418046	1093	273	0.418046	1093	273	0.418046	1093	273	0.418046	5
0.497035	1094	282	0.497035	1094	282	0.497035	1094	282	0.497035	1094	282	0.497035	6
0.382613	995	174	0.382613	995	174	0.382613	995	174	0.382613	995	174	0.382613	7
0.485305	1095	102	0.485305	1095	102	0.485305	1095	102	0.485305	1095	102	0.485305	8
0.462296	1070	100	0.462296	1070	100	0.462296	1070	100	0.462296	1070	100	0.462296	9
0.5	1095	294	0.5	1095	294	0.5	1095	294	0.5	1095	294	0.5	1
0.426323	1057	290	0.426323	1057	290	0.426323	1057	290	0.426323	1057	290	0.426323	2
0.398651	963	130	0.398651	963	130	0.398651	963	130	0.398651	963	130	0.398651	3
0.473524	1084	215	0.473524	1084	215	0.473524	1084	215	0.473524	1084	215	0.473524	4
0.497029	1093	282	0.497029	1093	282	0.497029	1093	282	0.497029	1093	282	0.497029	5
0.33024	1095	232	0.33024	1095	232	0.33024	1095	232	0.33024	1095	232	0.33024	6
0.199889	968	156	0.199889	968	156	0.199889	968	156	0.199889	968	156	0.199889	7
0.294009	1094	151	0.294009	1094	151	0.294009	1094	151	0.294009	1094	151	0.294009	8
0.438016	1050	156	0.438016	1050	156	0.438016	1050	156	0.438016	1050	156	0.438016	9
0.494872	900	293	0.494872	900	293	0.494872	900	293	0.494872	900	293	0.494872	1
0.5	1095	295	0.5	1095	295	0.5	1095	295	0.5	1095	295	0.5	1
0.498678	1087	294	0.498678	1087	294	0.498678	1087	294	0.498678	1087	294	0.498678	1
0.468901	1093	288	0.468901	1093	288	0.468901	1093	288	0.468901	1093	288	0.468901	1
0.499732	1095	292	0.499732	1095	292	0.499732	1095	292	0.499732	1095	292	0.499732	1
0.441823	1035	289	0.441823	1035	289	0.441823	1035	289	0.441823	1035	289	0.441823	2
0.5	1094	294	0.5	1094	294	0.5	1094	294	0.5	1094	294	0.5	3
0.458804	1089	286	0.458804	1089	286	0.458804	1089	286	0.458804	1089	286	0.458804	1
0.479582	1041	294	0.479582	1041	294	0.479582	1041	294	0.479582	1041	294	0.479582	2
0.273878	987	151	0.273878	987	151	0.273878	987	151	0.273878	987	151	0.273878	3
0.394328	1089	268	0.394328	1089	268	0.394328	1089	268	0.394328	1089	268	0.394328	4
0.104167	1015	212	0.104167	1015	212	0.104167	1015	212	0.104167	1015	212	0.104167	5
0.443017	1051	230	0.443017	1051	230	0.443017	1051	230	0.443017	1051	230	0.443017	6

Figure 5.4 Dominant point, and Distance from Centre features

Support vector per class for the kernels of support vector machine is given below for the 9 ids of matras given in table 5.1.The support vector for ids in 50% ,90% and 75% testing (partitioning) for four kernels of support vector machine are given in below tables.

Table 5.2 Support Vectors per class using various Kernels.

Stroke ID	Kernel	90 % training 10% testing	50% training 50% testing	75% training 25% testing	Stroke ID	Kernel	90 % training 10% testing	50% training 50% testing	75% training 25% testing
1	Linear	57	39	47	6	Linear	67	44	56
	Polynomial	86	51	67		Polynomial	83	51	74
	RBF	86	48	66		RBF	83	51	73
	Sigmoid	94	54	75		Sigmoid	87	53	77
2	Linear	57	41	54	7	Linear	56	30	44
	Polynomial	86	54	74		Polynomial	62	39	51
	RBF	86	53	74		RBF	60	38	49
	Sigmoid	90	55	75		Sigmoid	66	41	53
3	Linear	55	31	49	8	Linear	81	46	58
	Polynomial	73	44	64		Polynomial	93	47	72
	RBF	74	40	63		RBF	92	47	72
	Sigmoid	87	45	71		Sigmoid	83	47	74
4	Linear	68	39	67	9	Linear	74	37	55
	Polynomial	90	48	80		Polynomial	71	39	58
	RBF	90	49	80		RBF	70	40	55
	Sigmoid	91	49	82		Sigmoid	79	44	60
5	Linear	57	31	46					
	Polynomial	77	41	64					
	RBF	74	42	64					
	Sigmoid	81	47	69					

Table 5.3: Recognition accuracy with 900 resampled data

Kernels functions	50% training 50% testing	75% training 25% testing	90% training 10% testing
Linear	84.906%	92.453%	95.339%
Polynomial	71.476%	72.364%	74.362%
RBF	73.918%	77.026%	77.913%
Sigmoid	61.265%	68.923%	67.814%

As above in the table 5.2, the higher no of support vectors means that a slack between two separable classes. The higher no. of support vectors means the slack size is large, and a lower number means a less distance between classes. Linear kernel show less numbers of support vector for each class in comparison to other kernels. Further in table 5.3 the recognition rate of linear kernel is more than the other three kernels, with a recognition rate of 95.339% as the highest with 90-10 dataset.

CHAPTER 6 CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Online handwritten character recognition is difficult problem not only due to variation in human writing but also because of overlapped and joined characters. Characters can be written by using various scripts. To recognize a script involves a various stages. These stages are discussed here. Feature extraction is one of these stages. Feature extraction for Devanagari script is difficult. There are few reasons that creates problem in feature extraction of Devanagari script-

- Some characters are similar in shape.
- Characters can be written at different location on window.
- Large numbers of stroke and character classes are present there.

Techniques are discussed for feature extraction of Devanagari script. Here Dominant point method is discussed. This method has been applied on various foreign languages but not yet applied on Devanagari script. In above work, this method is applied on matras of Devanagari. On providing dominant point feature and distance of dominant point from centre for recognition, recognition accuracy for Devanagari matras shows significant improvement. So, dominant point is feasible method for Devanagari script also. Also, in this support vector machine is used for recognizing with four kernels (linear, polynomial, radial basis function and sigmoid). Linear gives best result out of four kernels applied her. Recognition accuracy of Linear is more than other kernels. Linear gives 95.339% accuracy in 90-10 dataset, 92.453% in 75-25 dataset and 84.906% in 50-50 dataset which is highest of all kernels.

6.2 Future Scope

Dominant point method is efficient but still recognition accuracy can be improved by increasing the size of dataset. Features other than dominant point and Distance of dominant point from center of stroke can be explored. Since, Devanagari script is complex in nature having various numerals, matras and letters, dominant point methods on numerals, characters of Devanagari script can be applied in future.

REFERENCES

[1] Aparna.K.H, Subramanian.Vidhya, Kasirajan.M, Prakash.G.Vijay, Chakravarthy.V.S Madhvanath.Sriganesh, "Online Handwriting Recognition for Tamil", Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition IEEE .Kokobunji,Tokyo,Japan,October 2004, pp 1-6.

[2] Ibrahiem M. M. El Emary, Mustafa M. Hammad, Amma, "On-line Structural Approach for Recognizing Hand-Written Indian Numbers", ICGST-GVIP Journal Volume (5), Issue (7), July 2005, pp 21-26.

[3] Arora.Sandhya,Bhattacharjee.Debotosh,Nasipuri.Mita,Basu.Deepak,Kundu.Mahantapas, "Combining Multiple Feature Extraction Techniques for Handwritten Devnagaricharacters recognition", 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10, pp 1-6.

[4] M.Dinesh, Sridhar.Murali Krishna, "Feature based on Encoding the Relative Position of a Point in the Character for Online Handwritten Character Recognition" ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02 Pages 1014-1017.

[5] Verma.Brijesh,Lu.Jenny,Ghosh.Moumita,Ghosh.Ranadhir,Hampton.Rock,"A Feature Extraction Technique for Online Handwriting Recognition".pp 1-5.

[6] Genoe.Rey, Kechadi.Tahar, "On the Recognition of Online Handwritten Mathematics using Feature based Fuzzy rules and Relationship precedence". In proceeding of: FUZZ-IEEE 2008, IEEE International Conference on Fuzzy Systems, Hong Kong, and China.pp 1641-1646.

[7] Singh. Dayashankar, Singh. Sanjay Kr., Dutta.Dr. (Mrs.)Maitreyee, "Hand Written Character Recognition Using Twelve Directional Feature Input and neural network".2010, International Journal of Computer Applications (0975 – 8887 Volume 1 – No. 3), pp 82-85.

[8] Joshi.Niranjan,G.Sita,Ramakrishnan.A.G,V.Deepu,Madhvanath.Sriganesh,"Machine Recognition of Online Handwritten Devanagari Characters". *Eight International Conference on Document Analysis an Recognition*, 2005. Vol 2, pp 1156-1160.

[9] Long-Long Ma, Cheng-Lin LiuA," New Radical-Based Approach to Online Handwritten Chinese Character Recognition", 2008 IEEE,International conference on pattern recognition,pp 1-4.

[10] Sweatha, thesis submitted on "Online handwritten character recognition of Devanagari and Tamil script using SVM".

[11] Naveen Garg, thesis submitted on "Online Handwritten Recognition of Gurumukhi Characters".

[12] Joshi.Niranjan,G.Sita,Ramakrishnan.A.G,Sriganesh.Madhvanath,"Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition", 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)IEEE.pp 444-449.

[13] V.Deepu, Sriganesh M,Ramakrishnan A. G, "Principal Component Analysis for Online Handwritten Character Recognition".Vol 2,pp 327-330,2004,International conference on pattern recognition.

[14] Razzak.Muhammad Imran, Hussain.Syed Afaq, Sher.Muhammad, Khan.Zeeshan Shafi, "Combining Offline and Online Preprocessing for Online Urdu Character Recognition", International MultiConference of Engineers an Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009.

[15] Yang.Su, Dai. Guozhong, "Detecting Dominant Points on On-line Scripts with a simple Approach", Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02) IEEE."pp 351-356.

[16] Chakraborty.Basabi, Chakraborty.Gautam, "A new feature extraction technique for online recognition of handwritten Alphanumeric characters", Elsevier preprint, 11 march 2002.Vol 148, Issue 1-4, pp 55-70.

[17] Kam.Fai.Chan, Dit.Yan.Yeung, “An Efficient Syntactic Approach to Structural Analysis of On-Line Handwritten Mathematical Expressions”, August 1998.Vol 33, Issue 3, pp 375-384, Elsevier pattern recognition.

[18] A.kosmala, G.Rigoll, “Recognition of Online Handwritten Formulas”, 6th-Int Workshop on Frontiers in Handwriting Recognition (IWFHR), August 12-14, 1998.

[19] Zanibbi.Richard,Blostein.Dorothean &Cordy.James.R, “Recognising Mathematical Expressions using Tree Transformation”, IEEE Transactions on Pattern and Machine Intelligence, Vol. 24, No. 11, Nov 2002,pp 14550-1467.

[20] Chau.Hauk.The, Roland T. Chin, “On the Detection of Dominant Points on Digital Curves”, IEEE Transactions on Pattern Analysis and Machine Intelligence.Vol 11 No. 8, August 1989, pp 859-872.

[21] Jaeger, S., Manke, S., Reichert, J, Waibel A. “Online handwriting recognition: the Npen++ Recognizer”. International Journal of Document Analysis and Recognition, Vol. 3, no. 3, pp. 169-180, 2001.

[22] L.Prasanth,V.Jagadeesh Babu, R. Raghunath Sharma, Dinesh. M, G.V.Prabhakara Rao, “Elastic Matching of Online Handwritten Tamil and Telugu Scripts using Local Features”. International conference on document analysis & recognition, 2007, Vol-2, pp 1028-1032.

[23] Bandyopadhyay.Asok, Chakraborty.Basabi, “Development of Online Handwriting Recognition System: A Case Study with Handwritten Bangla Character”, IEEE, 2009.World congress on nature & biologically inspired computation,pp 514-519.

[24] Connel.S.D, Jain.A.K, “Template-based online character recognition,” Pattern Recognition, vol 34, pp1-14, 2001.

[25] Pavlidis, “Structural Pattern Recognition”, Springer, 1977.

[26] C. Bahlmann. “Directional Features in Online Handwriting Recognition. Pattern Recognition”, 39(1), pp. 115–125, January 2006.

[27] Basu.S, Das.N, Sarkar.R, Kundu.M, Nasipuri.M, Basu.D.K, “Handwritten Bangla alphabet recognition using MLP based classifier”, NCCPB, 2005, pp 285-291.

[28] Belongie.S, Malik.J, and Puzicha.J, “Shape Matching and Object Recognition Using Shape Contexts”, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 4, April 2002, pp. 509-522.

[29] Rosenfeld.A and Johnston.E, “Angle detection on digital curves,”IEEE Trans. Cornput, vol. C-22, pp. 875-878, Sept. 1973.

[30] Rosenfeld.A and Weszka.J.S, “”An improved method of angle detection on digital curves,’’ fEEE Trans. Cornput, vol. C-24, pp. 940-941, Sept. 1975.

[31] Freeman.H and Davis.L.S, “A corner-finding algorithm for chain coded curves,’ IEEE Trans. Cornput, vol. C-26, pp. 297-303, Mar 1997.

[32] Sankar.P.V and Sharma.C.V, “A parallel procedure for the detection of dominant points on a digital curve.” Cornput. Graphics Image Processing, vol. 7, pp. 403-412. 1978.

[33] Anderson.1.M and Bezdek.J.C, “Curvature and tangential deflection of discrete arcs: A theory based on the commutator of scatter matrix pairs and its application to vertex detection in planar shape data,” IEEE Trans. Pattern Anal. Machine Intell, vol. PAMI-6, pp. 27-40, Jan. 1984.

[34] Dongre Vikas J, Mankar Vijay H,”A review on research on Devanagari character recognition”, IJCA, Nov 2012,Vol 12,No. 2,pp 1-8.

[35] Sweatha, thesis submitted on “Online handwritten character recognition of Devanagari and Tamil script using SVM”.

LIST OF PUBLICATIONS

[1] Pooja Tewari, Karun Verma,"Online Handwritten Character Recognition", International journal of computer applications. (Communicated)