HYBRID APPROACH TO CLASSIFY GURMUKHI SCRIPT CHARACTERS

Thesis submitted in partial fulfillment of the requirement for
The award of the degree of
Masters of Science
In
Mathematics and Computing

Submitted by
ANTARPREET KAUR
Roll no. - 30703004

Under
the guidance of
Mr. Rajiv Kumar

JULY 2009

School of Mathematics and Computer Applications
Thapar University
Patiala-147004 (PUNJAB)
INDIA

CERTIFICATE

*I hereby certify that the work which is being presented in the thesis entitled "**Hybrid Approach to classify Gurmukhi Script Characters**" in partial fulfillment of the requirements for the award of degree of Master of Science, School of Mathematics and Computer Applications, Thapar University, Patiala is an authentic record of my own work carried out under the supervision of Mr. Rajiv Kumar.*
The matter presented in this thesis has not been submitted for the award of any other degree of this or any other university.

Antarpreet Kaur

(Signature of Student)

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

Rajiv Kumar

(Mr. Rajiv Kumar)
Supervisor
SMCA, Thapar University
Patiala

Countersigned by:

SSBhatia 15.7.09

Dr. S.S. Bhatia
(Professor & Head)
School of Mathematics & Computer Applications
Thapar University, Patiala.

RK Sharma 22/7

Dr.R.K.Sharma
Dean of Academic Affairs
Thapar University
Patiala.

ACKNOWLEDGEMENT

It gives me immense pleasure to acknowledge my sincere gratitude to my academic supervisor, **Mr. RAJIV KUMAR, Lecturer, School of Mathematics and Computer Applications, Thapar University, Patiala**, for his constant help, encouragement and support throughout the course of this work. His conversations and encouragements will always be remembered. I want to thank him for introducing me to the field of Natural Language Processing and giving me the opportunity to work on this research project. I enjoyed working on this topic.

I would like to extend my special thanks to **Dr. S.S.Bhatia, Prof and Head, School of Mathematics and Computer Applications, Thapar University, Patiala**, providing help and necessary facilities in the department and directly or indirectly encouraging me to work harder during the whole course.

I would also like to give special thanks to my parents and family members who always inspired, encouraged and cheered me throughout my thesis period.

Last but certainly not the least; I would also like to thank my colleagues at Thapar University, whose company and friendship, I dearly cherish. Especially, I would like to thank Ramanjeet kaur and Keshwani Sharma for their support, suggestions and all possible help.

(Antarpreet kaur)

ABSTRACT

Extensive research has been done on optical character recognition in the last few decades. Most of the efforts were made to develop OCR systems for foreign languages which are available in the market. In the context of Indian languages, majority of work has been reported on Hindi and Bangla but a very few reports are available on Gurmukhi script which is used to write Punjabi language, one of the popular languages of northern India. In OCR system, segmentation or more precisely character segmentation is an important preprocessing step for text recognition. The segmentation can be done by various ways. But in this work, we explored one aspect that is not used on Gurmukhi script so far. The first, segmentation stage takes as input an image of a document and separates the different logical parts, like the line of a paragraph, words of a line and characters of a word. The lines and words are segmented according to the horizontal and vertical projection profile respectively. Then, for the segmentation of text into individual characters, water reservoir concept is used. Here, we classify Gurmukhi script characters to subclasses so that segmented part can be tested out whether this represents a correct character or not. For that purpose a hybrid approach is used. This hybrid approach is a combination of water reservoir and feature based approach. The significant point of the scheme is that a character image is tested against only certain subsets of classes at each stage, which increases the computational efficiency.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Transmission and storage of information is done not only through computers but also through paper documents. To integrate these two mediums of information flow, a solution is for computer to "read" paper documents. Machine simulation of human reading is one of the areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier. So, works are still going on this direction.

1.0 NATURAL LANGUAGE PROCESSING

Natural language processing is a field of science and linguistics concerned with the interaction between computers and human languages. Natural language generation systems convert information from computer databases into readable human language. The term "natural" language refers to the languages that people speak, like English and Japanese and Hindi, as opposed to artificial languages like programming languages or logic. "Natural Language processing", programs that deal with natural language in some way or another. The study of human languages developed the concept of communicating with non-human devices.

NLP deals with the Artificial Intelligence under the main discipline of Computer Science. The goal of NLP is to design and build software that will analyze, understand and generate languages that humans use naturally.

There are many applications of Natural Language processing developed over the years. The main are text-based applications, which involves applications such as searching for a certain topic or a keyword in a large document, translating one language to another or summarizing text for different purposes.

1.1 CHARACTER RECOGNITION

Character recognition is the term, which covers all types of machine recognition of characters in various application domains.

The intensive research effort on the field of character recognition was not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents.

A character recognition system can be either "online" or "offline." According to the mode of data acquisition, character recognition methodologies are categorized into two systems as *ONLINE CHARACTER RECOGNITION SYSTEMS* and *OFFLINE CHARACTER RECOGNITION SYSTEMS:*

Online character recognition is the process of recognizing handwriting, recorded with a digitizer, as a time sequence of pen coordinates. It captures the temporal and dynamic information of the pen trajectory. Applications of on-line character recognition systems include small handheld devices, which call for a pen-only computer interfaces and complex multimedia systems, which use multiple input modalities including scanned documents, speech, keyboard and electronic pen. These systems are useful in social environments where speech does not provide enough privacy. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on-line character recognition techniques.

Offline character recognition is the process of converting the image of writing into bit pattern by an optically digitizing device such as optical scanner or camera. The recognition is done on this bit pattern data for machine-printed or handwritten text. Applications of offline recognition are large-scale data processing such as postal address reading; check sorting, office automation for text entry, automatic inspection and identification. Offline character recognition is a very important tool for creation of the electronic libraries. Also, the wide spread use of web necessitates the utilization of offline recognition systems for content based Internet access to paper documents.

According to the text-type, *HANDWRITTEN* and *MACHINE -PRINTED CHARACTER RECOGNITION SYSTEMS* are two main areas of interest in character recognition field:

Machine printed text includes the materials such as books, newspapers, magazines, documents, and various writing units in the video or still image. Machine printed characters are uniform in height, width, and pitch assuming the same font and size are used. These problems for fixed-font, multi-font and omni-font character recognition is relatively well understood and solved with little constraint.

Handwritten text can be further divided into two categories: cursive and hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters are non-uniform and can vary greatly in size and style. Even characters written by the same person can vary considerably. In the location of characters is not predictable, nor the spacing between them. In an unconstrained system, characters may be written anywhere on the page and may be overlapped or disjoint. A typical recognition system will require some sort of constraints, or added information, about the data being processed.

1.2 WHAT IS OCR?

Optical character recognition (OCR) is the process of converting scanned images of machine-printed or handwritten text into a computer processable format. It involves computer software designed to translate images of typewritten text into machine-printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. If you scan a text document, you might want to use optical character recognition (OCR) software to translate image into text that you can edit. When a scanner first creates an image from page, image is stored in computer's memory as a bitmap. A bitmap is a grid of dots; one or more bits represent each dot. The job of OCR software is to translate that array of dots into text that computer can interpret as letters and numbers.

1.3 HISTORICAL BACKGROUND OF OCR

The first conceptual idea of OCR is due to Tauschek in 1929 and handle in 1933. Tauschek obtained a patent on OCR in Germany, followed by handle who obtained a U.S. patent on OCR in U.S.A in 1933.Tauschek was also granted a U.S patent on his method 1935.Machine was mechanical device that used templates. The first commercial system was installed at the Reader's Digest in 1955, which, many years later, was donated by Reader Digest to the Smithsonian, where it was put on display. The United States Postal Service has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rabinow. In 1974, Ray Kurzweil started the company Kurzweil Computer Products, Inc. and led development of the first omni-font optical character recognition – a computer program capable of recognizing text printed in any normal font. He decided that the best application of this technology of this

technology would be to create a reading machine for the blind, which would blind people to understand written text by having a computer read it to them out loud.

1.4 CURRENT STATE OF OCR TECHNOLOGY

Older OCR systems, work by matching the scanned images against stored bitmaps based on specific fonts. The hit or miss results of such pattern recognition systems helped establish OCR's reputation for inaccuracy. Today, OCR software can recognize a wide variety of fonts, but handwriting and script font that mimic handwriting are still problematic. Developers are taking different approaches to improve script and handwriting recognition. Advances are being made to recognize characters based on the context of the word in which the software will use knowledge of the parts of speech and grammar to recognize individual characters. Typical accuracy rate exceed 99%, although certain applications demanding even higher accuracy require human review for errors.

1.5 PHASES OF OCR

Here, in this work, we focus on the methodologies of character recognition system, emphasizing the offline recognition problem. A hierarchical approach for most of the systems would be from pixel to text, as follows:

Pixel ⇨ *Feature* ⇨ *Character* ⇨ *Sub-word* ⇨ *Word* ⇨ *Meaningful text*

The literature review in this field indicates that the above hierarchical tasks are grouped in the stages of the character recognition for image acquisition, pre-processing, segmentation, feature extraction, recognition, post processing.

The process of optical character recognition of any script can be broadly broken down into the following stages:

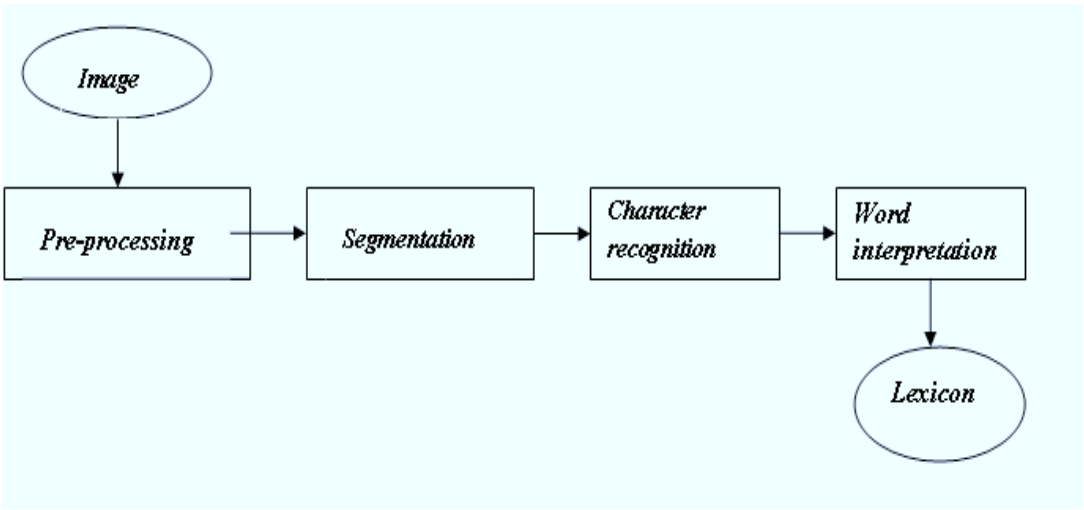


Figure 1.1 Stages of OCR

1. **Image acquisition**: this involves scanning a document and storing it as an image. The resolution (number of dots per inch, dpi) determines the rate of process.
2. **Pre-processing**: process of representing the scanned image for further processing. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the CR systems to operate accurately. It reduces noise and distortion, removes skewness and performs skeltonizing of the image thereby simplifying the processing the rest of the stages.
3. **Segmentation**: the pre-processing stage yields a 'clean' document in the sense that sufficient amount of shape information, high compression and low noise on normalized image is obtained. The next stage is segmenting the document into its sub-components. It separates the different logical parts, like text from graphics, line of a paragraph, and characters of a word. Segmentation is an important stage, because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script. There are mainly two types of segmentation:
    a) ***External segmentation***, which is the isolation of various writing units, such as paragraphs, sentences or words.
    b) ***Internal segmentation***, which is the isolation of letters, specially, in cursively written words.

4. **Feature Extraction**: analyzing its shape and comparing its features against a set of rules stored on OCR engine that distinguishes each character identify a character. The selection of a stable and representative set of features is the main part of pattern recognition system design. It is the most consequential issue in the designing issues involved in building an OCR system.
5. **Classification:** it is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the text segment according to preset rules.
6. **Post-processing:** it is the final stage, improves recognition by refining the decisions taken by the previous stage and recognizes words using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information.

1.6 FACTORS AFFECTING OCR ACCURACY

There are a number of key factors to consider when observing a printed resource and assessing whether it will produce the text resource accuracy desired through OCR technologies. Some of the main ones are listed below:

1.6.1 *Scanning methods possible*

Bit-depth is the number one factor that can improve OCR accuracy once a base level of 300 dpi resolutions is achieved. If the image can be represented as grayscale (8-bit) or better, then this is more likely to improve the OCR accuracy than almost any other scanning mechanism. So if given the choice of increasing resolution or increasing bit depth (i.e. from black and white to grayscale) then go for the grayscale. Remember that all OCR engines will struggle to recognize anything well if the resolution is below 300 dpi and that this is the absolute minimum baseline for scanning. Greyscale is clearly readable thus, the lower the standard of scanned image then the worse the OCR accuracy is likely to be. There are other factors related to the nature of the original that will affect the standard of the scanned image and these should be accounted for as well.

1.6.2 *Nature of original paper*

The original paper on which text appears is critical to the scanned image standard and is often the core reason why grayscale (8-bit) improves the OCR accuracy. If the OCR engine cannot discriminate between the character and the paper background noise then it will be more likely to

misrepresent the character. Grayscale images as opposed to Black &White give the OCR software a better chance of discriminating between text and noise and thus improve the accuracy.

1.6.3 *Nature of printing*

The nature of the printed text in the original may make a significant difference to OCR accuracy. Obviously if the text is printed poorly or if it was typed and characters are broken, faded or have indistinct edges then this will affect the ability of an OCR engine to recognize patterns and differentiate between similar shaped characters. So the clarity of the printing is a factor to consider. Some fonts may also have improved print clarity over others and also by the use of larger point sizes. Character sizes of below 6 points in the original will limit the accuracy likely to be achieved, although increasing resolution in the scanned image to 600 dpi and using grayscale may improve matters.

1.6.4 *Formatting complexities*

Variations in font size and type face may result in misunderstanding the characters. Broken character and touching character stemming from excess ink or paper degradations may not be recognized.

These are reasons why OCR vendors never claim their software to be 100% accurate.

1.7 APPLICATIONS OF OCR

A common goal in the field of artificial intelligence is to replicate the function of human beings and automate tasks that normally require manual labor or at least supervision. Optical character recognition (OCR) is a typical example of such a problem. OCR has been the subject of a large body of research because there are numerous commercial applications for this technology. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Some of the most significant applications include

- Speeding up the data entry: For many documents, input tasks, OCR is the most cost effective and speedy method available.
- To reduce data entry errors.
- To reduce the storage space required by paper documents. Each year, the technologies free areas of storage space given over to fill cabinets and boxes full of paper documents.

Some practical application potentials of OCR system are:

- Reading aid for the blind

- Automatic text entry into the computer for desktop publication
- Library cataloging
- Ledgering
- Automatic reading for sorting of postal mail
- Bank cheques and other documents
- Document data compression: from document image to ASCII format
- Language processing

1.8 GURMUKHI SCRIPT

Gurmukhi script is word primarily for the Punjabi language, which is the world's 14th most widely spoken language. Gurmukhi script is a logical composition of its constituent symbols in two dimensions. It is an alphabetic script. Gurmukhi has 12 vowels and 41 simple consonants. Besides the consonants and the vowels, other constituent symbols in Gurmukhi are a set of vowels modifiers called matra placed to the left, right, above or at the bottom of a character or conjunct), pure consonants forms corresponding to some consonant (also called half letters) which when combined with other consonants yield conjuncts.

Vowels and Vowel diacritics (Laga Matra)

ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
a	ā	i	ī	u	ū	e	ai	o	au
[ə]	[ɑ]	[ɪ]	[i]	[ʊ]	[u]	[e]	[æ]	[o]	[ɔ]
ਕ	ਕਾ	ਕਿ	ਕੀ	ਕੁ	ਕੂ	ਕੇ	ਕੈ	ਕੋ	ਕੌ
	ਕੰਨਾ	ਸਿਹਾਰੀ	ਬਿਹਾਰੀ	ਔਂਕੜ	ਦੁਲੈਂਕੜ	ਲਾਂਵਾਂ	ਦੁਲਾਂਵਾਂ	ਹੋੜਾ	ਕਨੌੜਾ
	kannā	sihārī	bihārī	auṅkaṛ	dulaiṅkaṛ	lānvāṁ	dulānvāṁ	hōṛā	kanauṛā
ka	kā	ki	kī	ku	kū	ke	kai	ko	kau

Other symbols

◌ੱ	ਅਧਕ (adhak) - doubles the consonant before which it appears	ਹੁੱਟੀ	huṭṭī [hʊʈʈi] - tired
◌ਂ	ਬਿੰਦੀ (bindī) - indicates nasalization. Used with all vowels except a, i and u	ਸ਼ਾਂਤ	šāṃt [ʃãt] - peaceful
◌ਃ	ਵਿਸਰਗ (visarg) - used very occassionally to represent an abbreviation or to add a voiceless 'h' after a vowel.	ਕਃ	kaḥ
◌ੰ	ਟਿੱਪੀ (ṭippī) - indicates nasalization. Used with a, i and u, and also with ū when in final position	ਤੰਦ	tamd [tə̃d] - strand
◌੍	ਹਲਨਤ (halant) - silences the inherent vowel. Sometimes used in Sanskritised text and dictionaries.	ਕ੍	k
ੴ	ek oṅkar - often used in Sikh literature. It literally means 'one God'.		

Consonants (Vianjans)

ੳ	ਊੜਾ (ūṛā) u, ū, o	ਅ	ਐੜਾ (aiṛā) a, ā, ai, au	ੲ	ਈੜੀ (īṛī) i, ī, e	ਸ	ਸੱਸਾ (sas'sā) sa [sə]	ਹ	ਹਾਹਾ (hāhā) ha [hə]
ਕ	ਕੱਕਾ (kakkā) ka [kə]	ਖ	ਖੱਖਾ (khakhkhā) kha [kʰə]	ਗ	ਗੱਗਾ (gaggā) ga [gə]	ਘ	ਘੱਗਾ (ghaggā) gha [gə]	ਙ	ਙੰਙਾ (ṅaṅṅā) ṅa [ŋə]
ਚ	ਚੱਚਾ (caccā) ca [ʧə]	ਛ	ਛੱਛਾ (chachchā) cha [ʧʰə]	ਜ	ਜੱਜਾ (jajjā) ja [ʤə]	ਝ	ਝੱਜਾ (jhajjā) jha [ʤə]	ਞ	ਞੰਞਾ (ñaññā) ña [ɲə]
ਟ	ਟੈਂਕਾ (ṭaiṅkā) ṭa [ʈə]	ਠ	ਠੱਠਾ (ṭhaṭhṭhā) ṭha [ʈʰə]	ਡ	ਡੱਡਾ (ḍaḍḍā) ḍa [ɖə]	ਢ	ਢੱਡਾ (ḍhaḍḍā) ḍha [ɖə]	ਣ	ਣਾਣਾ (ṇāṇā) ṇa [ɳə]
ਤ	ਤੱਤਾ (tattā) ta [tə]	ਥ	ਥੱਥਾ (thaththā) tha [tʰə]	ਦ	ਦੱਦਾ (daddā) da [də]	ਧ	ਧੱਦਾ (dhaddā) dha [də]	ਨ	ਨੱਨਾ (nannā) na [nə]
ਪ	ਪੱਪਾ (pappā) pa [pə]	ਫ	ਫੱਫਾ (phaphphā) pha [pʰə]	ਬ	ਬੱਬਾ (babbā) ba [bə]	ਭ	ਭੱਬਾ (bhabbā) bha [bə]	ਮ	ਮੱਮਾ (mam'mā) ma [mə]
ਯ	ਯੱਯਾ (yayyā) ya [jə]	ਰ	ਰਾਰਾ (rārā) ra [rə]	ਲ	ਲੱਲਾ (lallā) la [lə]	ਵ	ਵੱਵਾ (vavvā) va [wə]	ੜ	ੜਾੜਾ (ṛāṛā) ṛa [ɽə]
ਸ਼	ਸ਼ੱਸ਼ਾ (śaśśā) śa [ʃə]	ਖ਼	ਖ਼ੱਖ਼ਾ (ḵẖaḵẖḵẖā) ḵẖa [χə]	ਗ਼	ਗ਼ੱਗ਼ਾ (ġaġġā) ġa [ɣə]				
ਜ਼	ਜ਼ੱਜ਼ਾ (zazzā) za [zə]	ਫ਼	ਫ਼ੱਫ਼ਾ (faffā) fa [fə]	ਲ਼	ਲ਼ੱਲ਼ਾ (laḻḻā) ḻa [lə]				

The writing style is from left to right and the concept of upper/lower case (as in English) is absent. Most of the characters have a horizontal line at the upper part. Mostly this line, called headline connects the character of words. A word in Gurmukhi script can be partitioned into two horizontal zones. The upper zone denotes the region above the headline, while lower zone reprocess units. The area below the headlines, the major part of the character, is located in center zone.

a) Upper zone from line number 1 to 2 b) Middle Zone from line number3 to 4
c) Lower zone from line number 4 to 5

Figure 1.2: Three zones in Gurmukhi script

Sample text in Punjabi (Gurmukhi alphabet)

ਸਾਰੇ ਇਨਸਾਨ ਆਜ਼ਾਦ ਅਤੇ ਹੱਕ ਤੇ ਇੱਜ਼ਤ ਦੇ ਲਿਹਾਜ਼ ਨਾਲ
ਬਰਾਬਰ ਪੈਦਾ ਹੁੰਦੇ ਹਨ| ਉਹ ਅਕਲ, ਸਮਝ ਤੇ ਚੰਗੇ ਮੰਦੇ ਦੀ ਪਛਾਣ
ਅਤੇ ਅਹਿਸਾਸ ਰੱਖਦੇ ਹਨ, ਇਸ ਲਈ ਉਹਨਾਂ ਨੂੰ ਇੱਕ ਦੂਜੇ ਨਾਲ
ਭਾਈਚਾਰੇ ਵਾਲਾ ਸਲੂਕ ਕਰਨਾ ਚਾਹੀਦਾ ਹੈ|

1.9 GURMUKHI SCRIPT FROM OCR VIEW POINT

OCR for Indian languages in general is more difficult than for European languages because of the large number of vowels, consonants and conjuncts (combination of vowels and consonants) segmentation has to deal with the positioning of the conjucts and positioning of the conjucts half syllables. These factors coupled with the inflectional and agglutinative nature of Indian languages make the OCR task quite challenging, language models and computational linguistics as pertains to Indian languages is an area of recent research. Extensive research has been done on optical character recognition in the last few decades. Most of the efforts were made to develop OCR systems for foreign languages are available in the market.

In the context of Indian languages, majority of work has been reported on Hindi and Bangla but a very few reports are available on Gurmukhi script which is used to write Punjabi language, one of the popular languages of northern India. Very little work is seen in the literature on recognition of Indian languages in general & Gurmukhi in particular. Some of the papers dealing with machine

recognition of Indian language scripts have been presented in [18] and [20].To the best of our knowledge ,there has been no reported published research on OCR of Gurmukhi script though some research papers on pre-processing and classification techniques for OCR of Gurmukhi script are reported [25-26] and the first paper on Gurmukhi script recognition is given by G.S.Lehal & Chandan Singh[3].Research on different stages of OCR of Gurmukhi script is being carried out by the students at Punjabi university, Patiala.

Preliminary work was done by Sanjeev Kumar [25] and Khushwant kaur [26], developing a feature based Gurmukhi recognition script system. In this, count & location of local features such as endpoints, T-points, cross-points and loops were used to identify isolated Gurmukhi characters.

Work has also been reported on the recognition of degraded documents by R.K.Sharma, G.S.Lehal & M.K.Jindal [2]. In this, different kinds of degradation available in printed Gurmukhi script have been identified. After identifying the different kinds of degradation, problems associated with each kind of degradation have been discussed; some possible solutions have also been discussed.

1.10 ERRORS THAT CAN OCCUR IN OCR SYSTEMS

Some of the errors which may reduce the performance of OCR systems are:

A) Confusion because of shape similarity. This is caused by mapping. For example `d` onto `F`, and `B` as `f`. this is a relatively minor transgression because human readers distinguish such pairs by context.

B) Confusion caused by printing. Ink spreads might cause the closing of gaps, faint reproduction might cause gaps. It is tempting to evaluate OCR systems only on high quality input but that is utterly unrealistic.

C) Confusion caused by digitization. various types of digitization distortion are:
   - the point spread function of the scanner
   - non-uniform illumination
   - multiplicative noise (non- uniform paper reflection)
   - additive noise due to electronics

   Due to such errors, the performance of an OCR system can be degraded because the shape of the characters has been severely distorted by digitization. Since the averaging of the

input by the point spread function of the scanner is the major source of trouble. There has been a trend towards higher scanner resolution.

D) Confusion because of error in future selection. Even if a character is printed and scanned perfectly, it may be misread if the features are not computed correctly.

E) Confusion because of classifier design. This will happen if the classifier has not been trained in all possible forms of characters.

1.11 ASSUMPTIONS

We have considered following assumptions while developing algorithms for this thesis work:

1. We have assumed that the text is clean printed non-italic and non-decorative regular font is used.
2. Text has been differentiated from other non text area like graphics and images.
3. Noise cleaning, orientation and skew detection and corrections have already been done.
4. Documents should contain only Gurmukhi characters.

1.12 LITERATURE REVIEW

The need of some form of automated or semi-automated OCR has been recognized for decades. Today, there are numerous algorithms that perform this task, each with its own strengths and weaknesses. In this survey, a few papers are presented which are related to the present work.

R.G.Casey and Eric Lecolinet [1] aims at providing an appreciation for the range of character segmentation techniques that have been developed. Here, segmentation is listed under four headings. Classical approach consists of methods that partition the input image into sub-images, which are then classified. The second class of methods segments the image either explicitly by classification of pre-specified windows, or implicitly by the classification of subsets of spatial features collected from the image as a whole; the third strategy is a hybrid of first two, employing dissection together with recombination rules but using classification to select from the range of admissible segmentation possibilities offered by these sub-images. Finally, holistic approach, that avoids segmentation by recognizing entire character strings as units.

G.S.Lehal and Chandan Singh [2] describes a feature extraction and hybrid classification scheme for machine recognition of Gurmukhi characters, using binary decision tree and nearest neighbor. Classification process completes in three stages, where in the first stage, the characters are grouped into sets depending on their zonal positions. In the second stage, the characters in middle zone set are further distributed into smaller sub-sets by a binary decision tree using a set of robust and font independent features. In the third stage, the nearest neighbor classifier is used using the special features distinguishing the characters. The significant point of this scheme is that a character image is tested against only certain subsets of classes at each stage, which enhances the computational efficiency.

G.S. Lehal and Chandan Singh [3] present a system for recognition of machine printed Gurmukhi script. Character recognition in Gurmukhi script faces major problems mainly related to the unique characteristics of the script like connectivity of characters on the headline, a larger number of similar characters and two or more characters in a word having intersecting minimum bounding rectangles. A set of very simple and easy to compute features is used and a hybrid classification scheme consisting of binary decision tree and nearest neighbors is employed.

Neena Madan, Sunil Madan and Hardeep Singh [4] suggested a new approach to segment machine printed Gurmukhi text. To resolve the issues of touching characters, a two pass mechanism is used. In pass one; it approximates the segmentation point, while in pass-two the cutting point is optimized. This approach has been very successful in segmenting a pair as well as triplets of touching characters. This approach can be easily extended to the other Indian languages scripts such as Devnagri and Bangla, which have horizontal lines at the top called headlines.

M.K.Jindal, R.K.Sharma and G.S.Lehal [5] identified different kinds of degradation available in Gurmukhi script. After identifying the different kinds of degradation, that is, touching characters, broken characters, heavy printed characters, faxed documents and typewritten documents and problems associated with each kind of degradation have been discussed and some possible solutions have also been discussed.

U.P.Pal and Ariban Sarkar [6] deals with an Optical Character Recognition system for printed Urdu. Here, the document image is captured using a flatbed scanner and passed through skew correction, line segmentation and character segmentation modules. These modules are developed by combining conventional and newly proposed techniques. Next, individual characters are recognized using a combination of topological, contour and water reservoir concept based features. The feature detection methods are simple and robust. This approach achieves 97.8% character level accuracy on average.

N.Tripathy and U.Pal [7] proposed a scheme based on the water reservoir concept, for the segmentation of unconstrained Oriya handwritten text into individual characters. At first, the text image is segmented into lines, and then lines are segmented into individual words, and words are segmented into individual characters. For line segmentation, the document is divided into vertical stripes .Analyzing the heights of the water reservoir obtained from different components of the document; the width of a stripe is calculated. Stripe wise horizontal histograms are then computed and the relationship of the peak valley points of the histograms is used for line segment. Based on vertical projection profile and structural features of Oriya characters, text lines are segmented into words. For character segmentation, at first, isolated and connected characters in a word are detected. Using structural, topological and water reservoirs concept based features, touching characters of the word are then segmented.

U.Pal, A.Belaid and C.Choisy [8] deals with a new scheme for automatic segmentation of unconstrained handwritten connected numerals. This approach is mainly based on water reservoir. A reservoir is a metaphor to illustrate where the region numerals touch. Reservoir is obtained by considering accumulation of water poured from the top or from the bottom of the numerals. At first, considering reservoir location, and size, touching positions are decided. Next, analyzing the reservoir boundary, touching position and topological features of the touching pattern, the best cutting point is determined. Finally, combined with morphological structural features the cutting path for segmentation is generated.

U.Pal and Sagarika Datta [9] proposed a robust scheme to segment unconstrained handwritten Bangla texts into lines, words and characters. For line segmentation, at first, the text is divided

into vertical stripes. Stripe width of the document is computed by statistical analysis of the text height in the document. Next, the horizontal histogram of these stripes and the relationship of the minimal values of the histograms are used to segment text lines. Based on the vertical projection profile lines are segmented into words. For segmentation of characters, water reservoir principle is used. At first, isolated and touching characters in a word are identified. Next touching characters of the word are segmented based on the reservoir base area points and structural feature of the component.

U.Pal and N.Tripathy [10] proposed a scheme towards the recognition of Indian stylistic documents. Here, at first, using water reservoir concept based features the characters are segmented from the stylistic documents without any skew correction. Next, individual characters are recognized. For recognition, contour distances of the outer contour points of the characters of the characters are calculated from the centroid. These contour distances are then arranged in a particular order to get size and rotation invariant feature. Finally, computing statistical feature on these arranged contour distances the input character is recognized.

Fatos T.Yarman and Nafiz Arica [11] serves as a guide and update for the readers working in the character recognition area. First, an overview of the character recognition systems and their evolution over time is presented. Then, the available CR techniques with their superiorities and weaknesses are reviewed. Finally, the current status of CR is discussed and directions for future research are suggested. Special attention is given to the offline handwriting recognition, since this area requires more research to reach the ultimate goal of machine simulation of human reading.

R.K.Sharma and Dr.Amardeep Singh [12] proposed and implemented some algorithms to segment handwritten Gurmukhi Text, which have shown encouraging results. Algorithms have been proposed to segment the touching characters

B .Anuradha Srinivas, Arun Agarwal, and C. Raghavendra Rao[13] gives an overview research in optical character recognition system for Indian Language scripts. The aim of this paper to provide a starting point for the researchers entering in to this field.

Trier, O., Jain, A. and Taxt, T.[14] present an overview of feature extraction methods for offline recognition of segmented (isolated) characters. Different feature extraction methods are designed for different representation of the characters.

CHAPTER 2
SEGMENTATION

Pre-processing stage yields a 'clean' document in the sense that sufficient amount of shape information, high compression and low noise on normalized image is obtained. Pre-processing includes noise removal; skew detection and binarization .The next stage is to segment the document into its sub-components.

2.0 SEGMENTATION

Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.

Text segmentation is a process in which the text image is segregated into units of patterns that seem to form characters. All recognition algorithms depend on the segmentation algorithm to break up the image into individual characters. It is the process of dividing written text into meaningful units, such as words, sentences, or topics. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing.

Segmentation process involves three steps namely line segmentation, word segmentation and character segmentation.

Line segmentation is the process in which from the image, we extract only lines or differentiate the lines. Horizontal projection of a document image is most commonly employed to extract the lines from the document. If the lines are well separated, and are not tilted, the horizontal projection will have separated peaks and valleys, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between lines.

Word segmentation is the process in which from the line segmentation we extract only words. As we know that there is a distance between one word to another word this concept is used for word segmentation. Word segmentation is the problem of dividing a string of written language into its component words. Word splitting is the process of parsing concatenated text (i.e. text that contains no spaces or other word separators) to infer where word breaks exist. A vertical projection profile gives the column sums. One can separate lines by looking for minima in horizontal projection profile of the page and then separate words by looking at minima in vertical

projection profile of a single line. Valleys in the vertical projection of a line image can be used in the extraction of words in a line, as well as extracting individual characters from the word.

Character segmentation is the process in which from the word segmentation we extract only characters. Character segmentation is a crucial step of OCR systems as it extracts meaningful regions for analysis. This step attempts to decompose the image into classifiable units called character. A poor segmentation process produces misrecognition or rejection segmentation process carried after out only the pre-processing of the image.

According to Casey and Lecolinet,

"Character segmentation is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols".

Character segmentation is a necessary preprocessing step for character recognition in almost all OCR systems. Character segmentation has been a well investigated field over the last decade and its main aim was to provide individual character to optical character recognition (OCR) algorithms.

Automatic segmentation is the problem in natural language processing of implementing a computer process to segment text.

It is important to note that this step (segmentation) only produces a sequence of fragments, while the segmentation of characters is confirmed at the classification stage.

The digitized image is processed to lines and words using appropriate horizontal and vertical projection profiles. The projection profile gives valleys of zero height for these OFF pixels between the text lines.

The two operators of projection profiles are:

- Horizontal projection: For a binary image of size H* W where H is the height of the image and W is the width of image, the horizontal projection is defined as HP (j), j=1, 2 …H.

  This operation counts the total number of black pixels in each horizontal row.

- Vertical Projection: For a binary image of size H*W where H is the height of the image and W is the width of the image, the vertical projection has been defined as VP (k), k=1, 2, ...W.

  This operation counts the total number of black pixels in each vertical column.

How to segment lines from a document image?

Horizontal projection of a document image is most commonly employed to extract the lines from the document. If the lines are well separated, and are not tilted, the horizontal projection will have separated peaks and valleys, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between lines.

How to segment words from lines?

Here, the vertical projection profile gives the column sums. Lines can be separated by looking for minima in horizontal projection profile of the page and then separates words by looking at minima in vertical projection profile of a single line .Valleys in the vertical projection of a line image can be used in the extraction of words in a line, as well as extracting individual characters from the word.

How to segment characters from a word?

The words are then segmented into smaller parts and given to the classifier to recognize the characters. For English and some other foreign languages the normal projection profile method is enough to segment the machine printed characters.

But in the script with which we are dealing, that is, Gurmukhi script, we have to use some different approach, so as to check whether the segmented area gives a character or not. If segmented area is inaccurate, then the segmentation is done again.

2.1 SEGMENTATION IN GURMUKHI SCRIPT

The text is segmented first into lines, each line into words and finally each word is segmented into its constituent characters

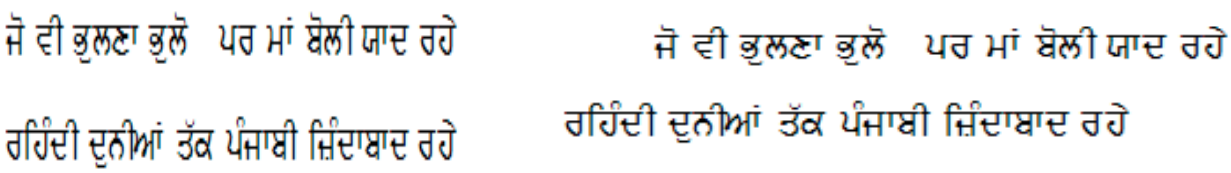


Figure 2.1. (a) Line segmentation

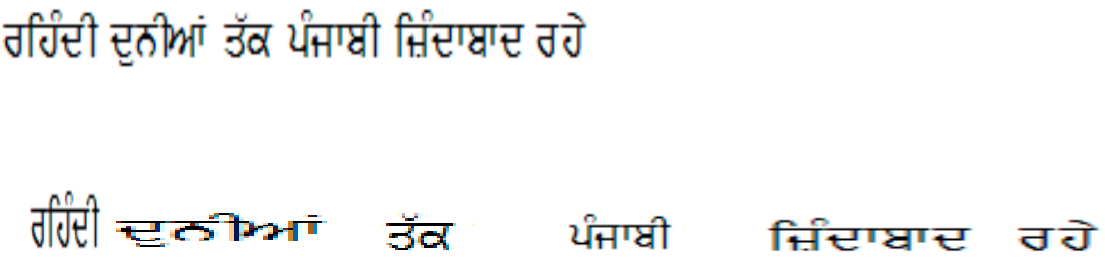


Figure 2.1. (b) Word segmentation

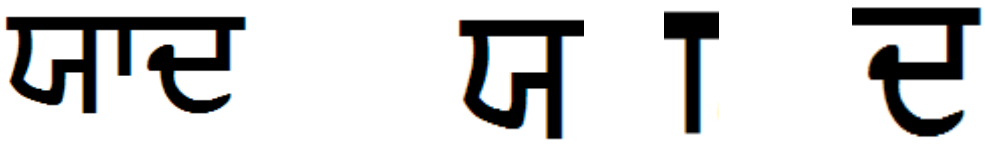


Figure 2.1 (c) Character segmentation

2.2 SEGMENTATION STRATEGIES

The segmentation stage takes in a page image and separates the different logical parts, like lines of a paragraph, words of a line, and characters of a word. Different methods used can be classified based on the type of text and strategy being followed like ***straight segmentation method, recognition based segmentation*** and ***cut classification method***

1. Straight segmentation method: each word is segmented into several characters and the character recognition techniques are applied to each segment. It is simple but it depends on the accuracy of detection of segmentation points. It is known as classical or dissection approach.
2. Recognition based segmentation method: a number of segmentation points are found in the touched characters. The candidates are confirmed by using recognition results. This method is fully dependent on the performance of the recognizer.
3. Cut classification method: It is based on a classifier deciding whether it represents a cut hypothesis or not, for each column of the character image. It is known as holistic approach

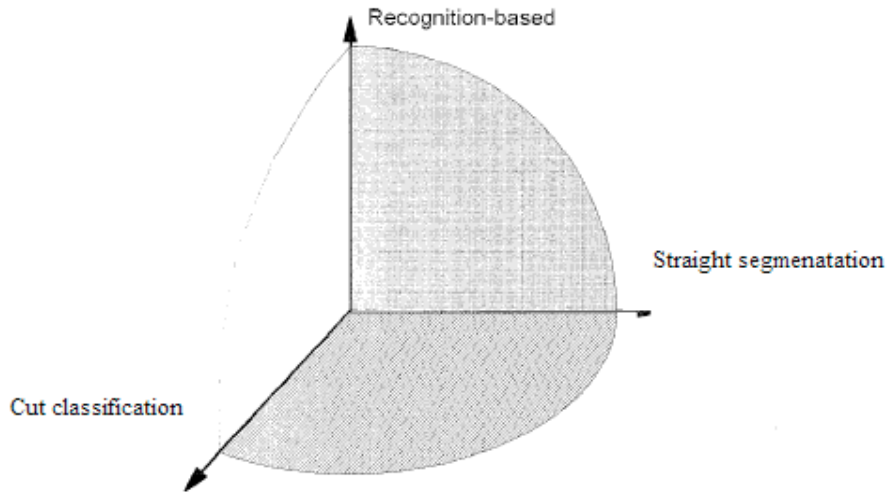


Figure 2.2 The view of segmentation strategies in 3 D space

2.2.1. Straight segmentation method for segmentation:

In this technique, we usually decompose the image into sequence of sub images using general feature like approximate character size, pitch and white space.

Dissection is an intelligent process that analyses the image; however classification into symbol is not involved at this point.

- Using white space and pitch: In machine printed, vertical white space often serves to separate successive characters but its not there with handwritten document, but we can provide boxes for handwritten in which to print individual symbols. Especially, where OCR is concerned, we provide additional space between characters because it is easy to segment characters separated by white space. Pitch or number of characters per unit of horizontal distance, provides the basis for estimating segmentation points. In many machine printed applications involving limited font set, each character occupies a block of fixed width so applying this rule permits correct segmentation in case where several character along the line are merged or broken.
- Projection analysis: The vertical projection of printed line consists of simple running count of black pixels in each column. It can serve for detection of white space between successive letters. Moreover, it can indicate locations of vertical strokes in machine prints or region of multiple lines in handprint. Thus, the vertical histogram is the basis of segmentation process, but it fails, to make clear distinction between merged characters .We can use second difference of projection for such separation. Sometimes, image is transformed by the AND of adjacent column with the difference function applied to transformed image, so the separation of such character can be possible to higher extent.
- Connected component processing: The methods described above are generally not used for hand printed characters or proportional fonts. Thus, the pitch based methods cannot be used for such documents because there may be different number of character. Similarly, pitch based methods are not used of slanted characters or when inter character connection and inter connection stroke have similar thickness. So, technique for segmentation of hand printed text calls for a two –dimensional analysis even for Non-touching characters. A general approach used is determining black regions or blobs then these blobs are broken into characters. The approaches used are
    - i) Bounded box analysis
    - ii) Splitting of connected components
- Dissection with contextual processing (graphemes): The input images can be divided into sub images that are not necessarily forming characters. The preliminary shapes called Graphemes or

pseudo-characters are intended to fall into readily identifiable classes. A contextual mapping function from grapheme to symbols can then complete the recognition process. In doing so, the mapping function may combine or split grapheme classes, that is, implement a many to one or one to many mapping. This amounts to re-segmentation or as over-segmentation when the intent to leave no composite characters.

2.2.2. Recognition based segmentation:

This is another approach for segmentation, which usually employs another method that is; the image is segmented into sub images without considering their contents. There are two steps for recognition based segmentation:-

i) windowing step or generation of segmentation hypothesis

ii) verification step or choice of best hypothesis

A new technique called shortest path segmentation having features of dynamic programming and neural net recognition was proposed, this technique selects the optimal consistent combination of cuts from a pre-defined set of windows.

2.2.3. Holistic Approach:

It recognizes the entire word as a unit. So there is no requirement for the character segmentation. The main flaw is that, it requires a pre-defined lexicon because this is not dealing with characters but with words. For this approach, training session is a must to modify the lexicon. So, this kind of methods is usually suitable for applications having static lexicon where there is a limited domain of the OCR that is cheque recognition, passport reader, paper checking etc. There are two steps for this approach:

I. first is, feature extraction

II. second is, comparison of the word with the words in the lexicon

In actual practice, combination of the above approaches is used to segment the space. One of the recognition based approach is water reservoir approach. Before moving ahead, it is better to discuss this approach first. But firstly, we discuss the recognition based segmentation in detail followed by the water reservoir concept.

2.3 RECOGNITION BASED SEGMENTATION

This is the segmentation in which the system searches the image for components that match classes in its alphabet. In this, no feature based dissection algorithm is employed. Rather, the image is divided systematically into many overlapping pieces without regard to content. These are classified as part of an attempt to find a coherent segmentation /recognition result. Systems using such a principle perform "recognition based segmentation.

Methods considered here also segment words into individual units (letters). Letter segmentation is a by-product of letter recognition, which may itself be driven by contextual analysis. The main interest of this category of methods is that they bypass the segmentation problem. No complex "dissection" algorithm has to be built and recognition errors are basically due to failures in classification.

The basic principle is to use a mobile window of variable width to provide sequences of tentative segmentations which are confirmed (or not) by character recognition. Multiple sequences are obtained from the input image by varying the window placement and size. Each sequence is assessed as a whole based on recognition results. The windowing process can operate directly on the image pixels, or it can be applied in the form of weightings or groupings of positional feature measurements made on the images.

Recognition based segmentation consists of the following two steps:

1) Generate of segmentation hypothesis (windowing step).
2) Choice of the best hypothesis (verification step).

There were many algorithms given for this segmentation method. One algorithm was reported, that is, recursive algorithm for machine printed characters. this algorithm also based on prototype matching, systematically tests all combinations of admissible separation boundaries until it either exhausts the set of cut points, or else finds an acceptable is one in which every segmented pattern matches a library prototype within a pre-specified distance tolerance.

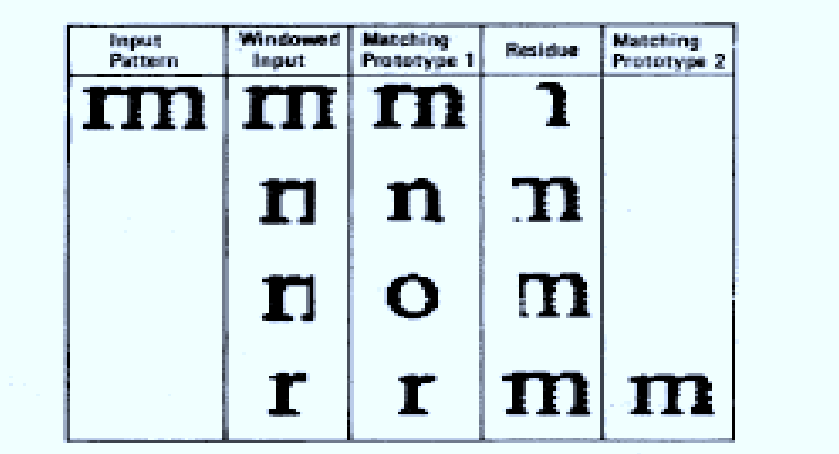


Figure 2.3 Recursive segmentation.

The example shows the results of applying windows of decreasing width to the left side of an input image. When the sub image in the window is recognized (in this case by matching a prototype character stored in the system's memory), then the procedure is recursively applied to the residue image. Recognition (and segmentation) is accomplished if a complete series of matching windows is found. In the top three rows, no match is obtained for the residue image, but successful segmentation is finally obtained as shown at the bottom.

Here, various features and their positions of occurrence are recorded for an image. Each feature contributes an amount of evidence for the existence of one or more characters at the position of occurrence. The positions are quantized into bins such that the evidence for each character indicated in a bin can be summed to give a score for classification. These scores are subjected to contextual processing using a pre-defined lexicon in order to recognize words. The method is being applied to text printed in a known proportional font.

2.4 WATER RESERVOIR CONCEPT

One of the recognition based technique for character segmentation is the water reservoir technique. The water reservoir principle is as follows. If water is poured from a side of a component, the cavity regions of the component where water will be stored are considered as reservoirs of the component.

The following are some terms related to this concept

1. ***Top Reservoir***: The reservoirs obtained when water is poured from top of the component. A bottom reservoir is visualized as a top reservoir when water is poured from top after rotating the component by 180 º.
2. ***Bottom Reservoir:*** The reservoirs obtained when water is poured from bottom of the component. A top reservoir is visualized as a bottom reservoir when water is poured from bottom after rotating the component by 180 º.
3. ***Left Reservoir***: If water is poured from left side of a component, the cavity regions of the component where water will be stored are considered as left reservoirs. A left reservoir of a component can also be visualized as a top reservoir when water is poured from top after rotating the component by 90º clockwise.
4. ***Right Reservoir:*** If water is poured from right side of a component, the cavity regions of the component where water will be stored are considered as right reservoirs. A right reservoir of a component can also be visualized as a top reservoir when water is poured from top after rotating the component by 90º anti-clockwise.

   Here top, left, right, bottom reservoirs of some Gurmukhi characters are shown respectively.

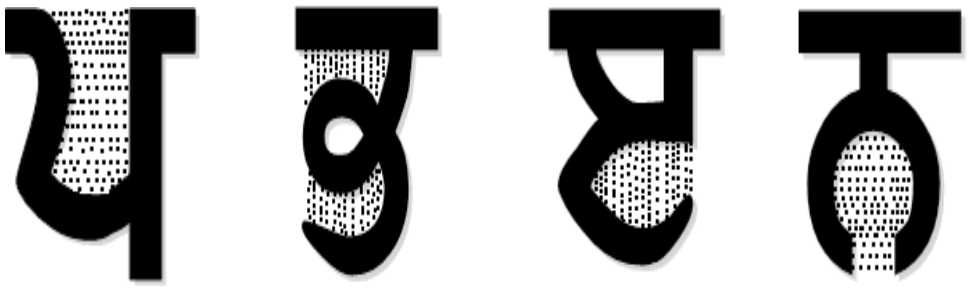


Figure 2.4: The Top, Left, Right, & Bottom reservoirs

5. ***Water flow level***: The level from which water overflows from a reservoir is called as water flow level of the reservoir.
6. ***Reservoir base-line***: A line passing through the deepest point of a reservoir and parallel to water flow level of the reservoir is called as reservoir base-line.
7. ***Height of the reservoir***: The height of the reservoir is the depth of the reservoir.

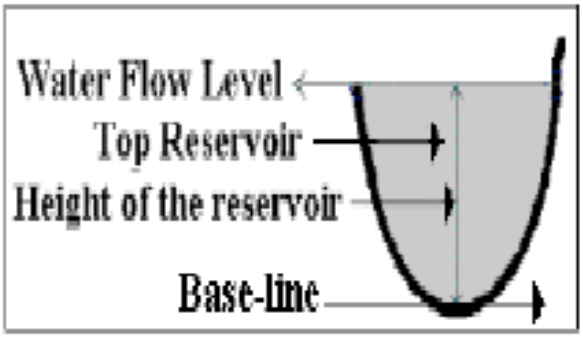


Figure 2.5: Water reservoir and its features

We first segment a character, then classify it according to water reservoir concept and according to the rule of recognition based segmentation. In the next chapter, we would discuss the feature extraction and classification according to which these segmented areas are being classified.

CHAPTER 3

FEATURE EXTRACTION AND CLASSIFICATION

3.0 INTRODUCTION

For the character recognition, we consider some principal features. These features are chosen with the following considerations:

(a) Robustness, accuracy and simplicity of detection

(b) Speed of computation

(c) Independence of size and fonts

(d) Tree classifier design need

After considering these features, initial classification of characters is done. Here, in this work we first extract some basic features of Gurmukhi script on the basis of which further we classify them using binary tree classifier.

3.1 FEATURE EXTRACTION

After an image has been segmented into regions, it is ready to enter the next level that is, the feature extraction stage. Feature extraction is defined as problem of "extracting from the raw data the information, which is most relevant for classification purposes, in the sense of minimizing the within class pattern variability while enhancing the between class pattern variability". Feature extraction can be considered as finding a set of parameters (features) that define the shape of the underlying character a precisely and uniquely as possible. The features have to be selected in such a way that they help in discriminating between characters.

It should be clear that different feature extraction methods fulfill this requirement to a varying degree, depending on the specific recognition problem and available data. A feature extraction method that proves to be successful in one application domain may turn out not to be useful in another domain.

Feature Extraction is an important step in achieving good performance in OCR systems. The choice of feature extraction methods limits or dictates the nature and output the pre-processing step. In practice, the requirement of a good feature extraction method makes selection of the best method for a given application a challenging task.

After feature extraction, classification process is discussed next, which is done on the basis of feature extraction.

3.2 CLASSIFICATION

It is important to note that segmentation only produces a sequence of fragments, while the segmentation of characters is confirmed at the classification stage. The classification stage in an OCR process assigns labels to character images based on the features extracted and the relationships among the features. In simple terms, it is this part of the OCR which finally recognizes individual characters and outputs them in machine editable form.

The classification process is carried out at the final stage to recognize the character. It assigns an input character to one of many pre-specified classes, which are based on the extracted features and their analysis. Classification stage uses the feature extracted in the feature extraction stage to identify the text segment according to preset rules. A tree classifier determines the classification of a point in feature space by starting from the root node, the classifier tests a particular feature or a set of features associated with that node and decides whether to branch to the next left node or the next right node. The process is continued until the classifier traces a path to a particular terminal node and returns the classification associated with that terminal node.

3.3 BINARY TREE CLASSIFIER

The design of a tree classifier has three components:

(1) a tree skeleton or hierarchical ordering of the class labels,
(2) choice of features at each non-terminal node,
(3) The decision rule at each non-terminal node.

The developed tree is a binary tree where the number of descendants from a non-terminal node is two. Binary decision tree classifier is an established technique in the repertoire of pattern recognition researchers. While traversing the tree, only one feature is tested at each non-terminal node. The selection of the feature at a particular non-terminal node is done by considering its impact on the length of the tree. If the set of patterns at a non-terminal node can be sub-divided into two subgroups by examining a feature so that the number of elements of one group is roughly equal to that of the other group, the resulting binary tree will have zones.

The binary tree classifier has the advantage of speed, since the maximum number of comparisons required for the classification of a character is equal to the height of the classifier tree, which is

not more than 10 in most of the cases. Thus at most only 10 comparisons are needed for the classification of a character using a binary classifier tree.

Disadvantage of binary classifier trees

- they are sensitive to noise and fonts
- If once a wrong decision is made at one of the nodes, there is no coming back and a wrong path will be followed from that point onwards and ultimately an incorrect decision will be made about the classification of that character.

3.4 PROPOSED SCHEME

As explained in the previous chapter, there are three fundamental strategies of segmentation, which are 'classical approach', 'recognition based' and 'holistic method'. But here, we will go further with the water reservoir technique, which is one of the recognition based method, for segmenting the characters.

The classification stage uses the features extracted in the feature extraction stage to identify the text segment according to the preset rules. After classifying characters according to reservoir concept, we further classify them on the basis of three other features, that is, number of components, according to the sidebar and according to the loop.

Then, we designed a strictly binary decision tree in which the leaf nodes correspond to the classification of the character in one of its subclasses. A tree classifier determines the classification of a point in feature space, by successively narrowing the region in which it is expected to lie. Starting from the root node, the classifier tests a particular feature or a set of features associated with that node and decides whether to branch to the next left node or the next right node. The process is continued until the classifier traces a path to a particular terminal node and returns the classification associated with that terminal node. Only one feature is tested at each non-terminal node for traversing the tree. The decision rules are binary that is the presence/absence of the feature. Here are the proposed methods, at first we classify characters according to the four reservoirs one by one and the second method describes their further classification on the basis of three features of Gurmukhi script.

In the following methods, we are using some abbreviations to make it less complex. These are left curvature –LC, right curvature-RC, top curvature- TC, bottom curvature-BC, number of components-NC

3.5 METHOD BASED ON THE WATER RESERVOIR CONCEPT

As already discussed there are 41 (35 + 6) characters in Gurmukhi script. This set is shown below and let us call it as Gurmukhi character set (say) C_0 .

Get a character say c from C_0. Then check whether c has left curvature (LC), and then on the obtained two paths which lead to two sub classes. Now on these subclasses we apply check for the presence of RC. Further, two alternate paths are available which are followed by the decision rule to check whether the TC exists in the character or not. After these combined checks this is the turn to check the presence of bottom curvature on each available subclass. Finally, after the combination of all these four checks, we got seven subclasses at the end along with two individual characters.

And applying the above discussed method, manually, we get the following seven subclasses

Sr. No.	Class	Class member
1	SC1	l L n
2	SC2	k | t T F x d P v
3	SC3	G m
4	SC4	K p ^
5	SC5	g j X S Z z
6	SC6	h c J f q Q D b B r
7	SC7	a e C \

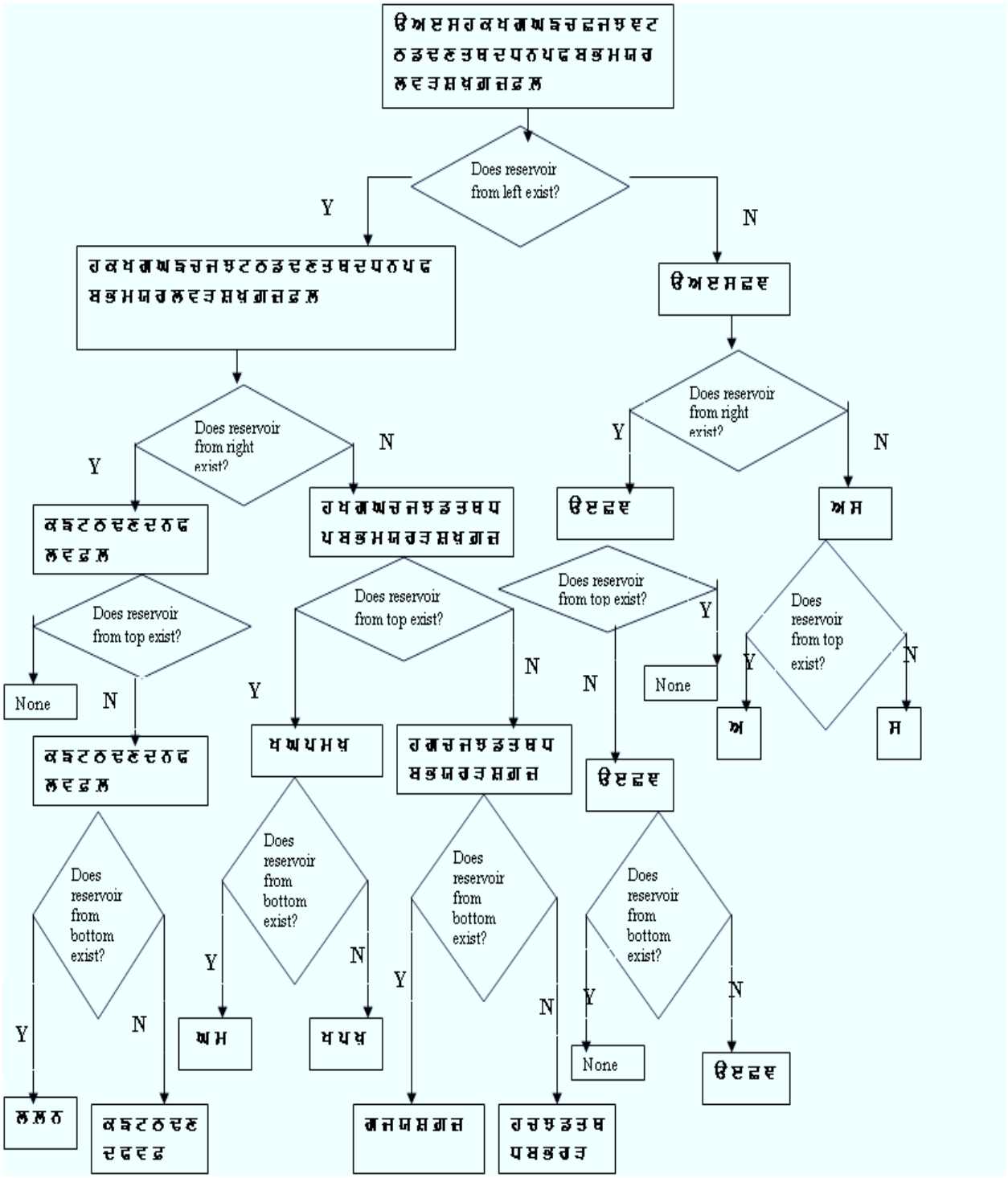
Table 3.1: Classes obtained after Water Reservoir method

Sr. No.	Individual class	Individual Character
1	IC1	**A**
2	IC2	**S**

Table 3.2: Individual Characters obtained after Water Reservoir method

SCn stands for subclass n. ICn stands for individual characters.

The above mentioned procedure is diagrammatically shown in flowchart 3.1. This representation is shown by using the binary tree classifier.



Flowchart 3.1: Water reservoir based classification of Gurmukhi Script characters

After a careful analysis of these subclasses (SC1 – SC7), we found that some of classes consists of more than three or more characters. It was felt that, there should be some mechanism by which one can further classify these sub classes to have more precise results. So, we decided to apply feature based approach to classify these classes. The features of Gurmukhi script which are used to classify SC1 – SC7 are as follows:

- Sidebar: the presence or absence of a sidebar is a robust feature for classifying the characters. For example m, X, r have a sidebar while k, | and C do not have it.
- Loops: the presence of a loop in the character is another important classification feature. We consider a loop only if headline is not part of that loop. For example, X does not have a loop since headline is involved while r has a loop.
- More than one component: if the characters have a dot at their bottom, they are considered having more than one component. For example L has more than one component but l has a single component.

3.6 METHOD FOR FEATURE BASED CLASSIFICATION

Get any character c from SC1- SC7, and then apply a check to classify it according to the number of components it have. Further, we apply check for the sidebar on the obtained subclasses. From here we got two paths, one having the characters which have the sidebar in it, others having the characters which do not have the sidebar. On these subclasses, a rule is applied to test them for the loop. After classifying according to these three features, we get 11 subclasses and 7 individual characters.

Here, after applying this, we get the following subclasses:

Sr. No.	Class	Class member
1	S1	k | T F P
2	S2	t x d v
3	S3	G m
4	S4	g j X
5	S5	S Z z
6	S6	c Q b r
7	S7	h D
8	S8	f B
9	S9	J q V
10	S10	a C
11	S11	e \

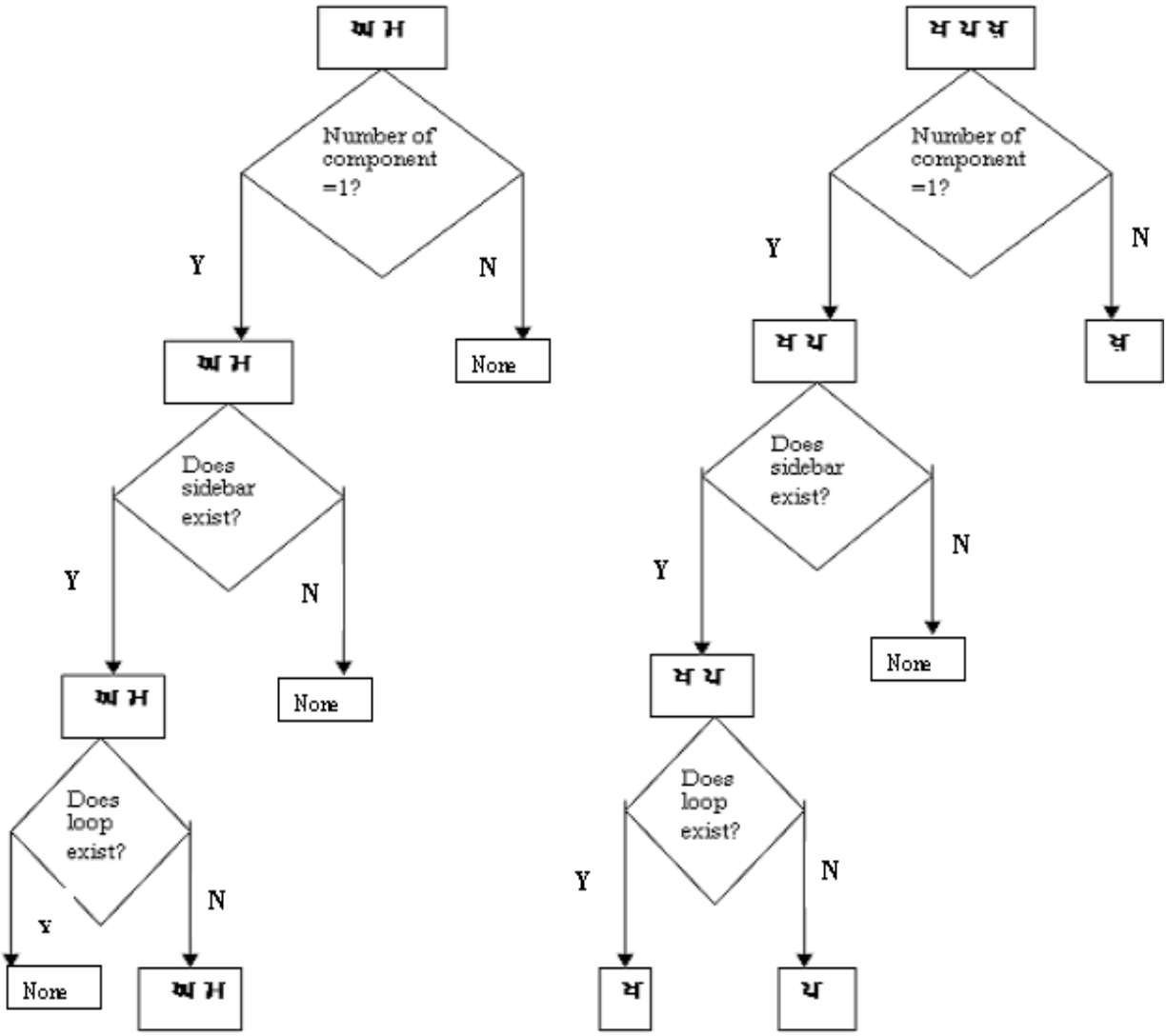
Table 3.3: Classes obtained after Feature Extraction method

Sr. No.	Individual class	Individual Character
1	IC3	L
2	IC4	n
3	IC5	l
4	IC6	K
5	IC7	p
6	IC8	^
7	IC9	&

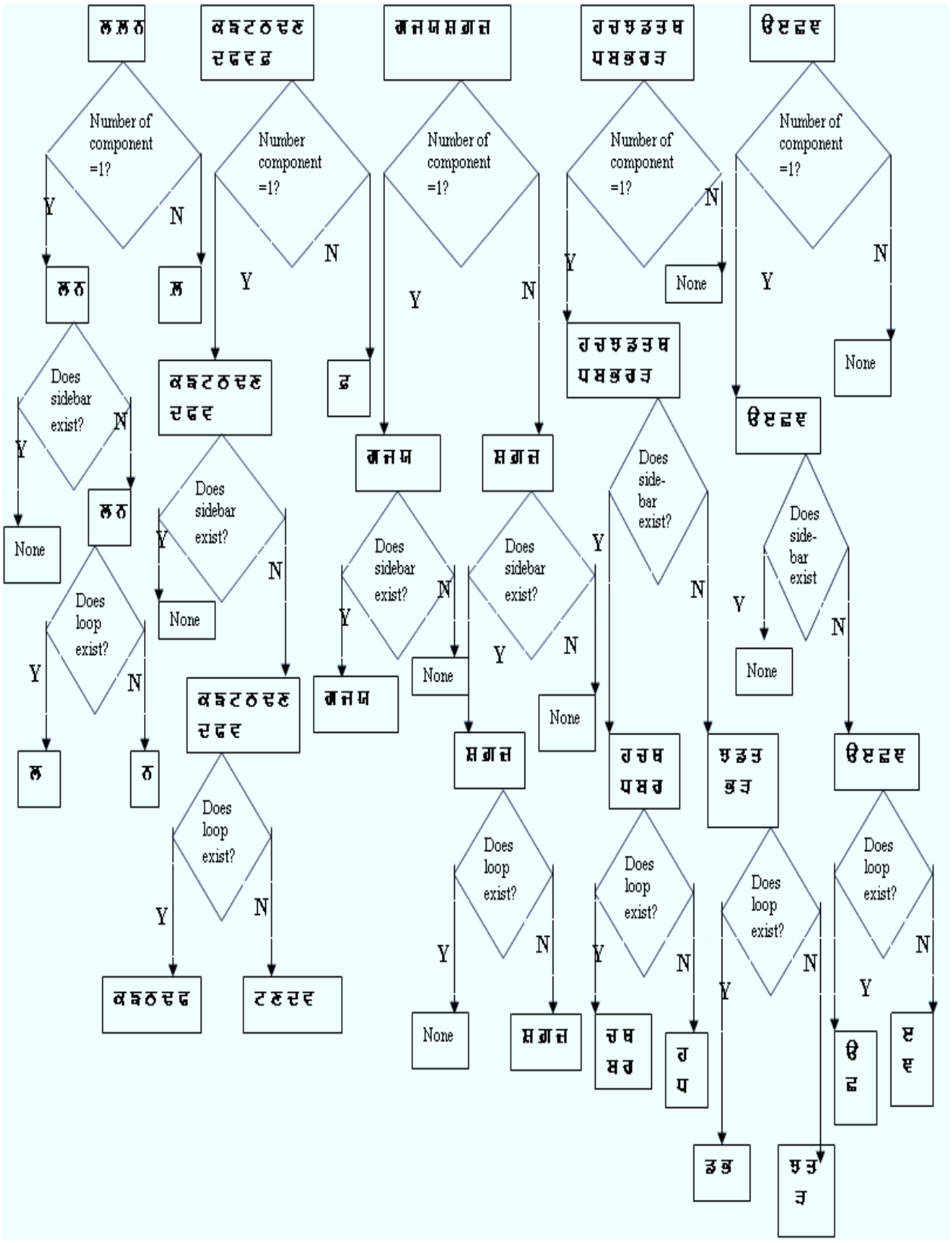
Table 3.4: Individual Characters obtained after Feature Extraction method

So, from Table 3.2 and 3.4 it is clear that overall 9 specific characters are found by these two methods.

The above mentioned procedure is diagrammatically shown in flowchart 3.2. This representation is shown by using the binary tree classifier.



Flowchart 3.2 (a): Feature based classification of Gurmukhi Script characters



Flowchart 3.2(b): Feature based classification of Gurmukhi Script characters.

These flowcharts give the diagrammatic representation of the algorithms given above. With the help of these flowcharts; we can classify Gurmukhi characters into some sets of classes based on their features. Some characters can individually be separated. At last, we get 11 subclasses and 9 individual characters which make our recognition process much easier.

CHAPTER 4
CONCLUSION AND FUTURE SCOPE

CONCLUSION AND FUTURE SCOPE

Optical character recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format. The practical importance of OCR applications, as well as the interesting nature of OCR problem, has led to great research interest and measurable advances in the field. But very limited research is reported on OCR of the scripts of Indian languages. There are a few research papers on Gurmukhi script reported till now.

Since Gurmukhi script is used primarily for the Punjabi language, which is the world's 14^{th} most widely spoken language. But in this field of character recognition, Gurmukhi script faces many problems related to unique characteristics of the script like connectivity of characters on the headline, a large number of similar characters in a word having intersecting minimum bounding rectangles.

Here, in this work, a set of very simple and easy to compute features is used and a hybrid classification scheme is employed based on water reservoirs and feature based approach. By classifying according to water reservoir and feature based approach, we are able to obtain characters in different classes, having similar characteristics.

There are 41 characters (35 consonants and 6 two component characters) in Gurmukhi script. Without any particular approach; it becomes very cumbersome job to recognize each character. But with the application of proposed hybrid approach, the efforts done for character recognition reduce tremendously. Firstly, we classify the segmented part as per the left reservoir, and then checks for right, top and bottom reservoirs are applied respectively using the first algorithm3.4 and then it is shown in flowchart 3.1. Since there were many similarities found in the obtained subclasses yet so we further apply Algorithm 3.5 to classify them as per the Gurmukhi script features. The diagrammatic representation of algorithm 3.5 is shown in flowchart 3.2.

As per algorithm analysis and design theory [20], the overall height of the binary tree gives the time taken to find a particular item. Here in present case, the height of the tree for water reservoir is 4 shown in flowchart 3.1 and for feature base approach it is 3, as shown in flowchart 3.2. So, finally we can say that overall height is 7, which means we have to apply overall 7 checks, to

find out a particular class for a segmented area. Therefore, with this hybrid approach, at last we are left with only 11 subclasses and 9 individual characters. In this way, recognition of Gurmukhi characters may become an easy task. One can easily establish a character by finding out to which class it belongs. Finally, it will also increase the computational efficiency of the overall OCR system.

This work reported on Gurmukhi language script may be extended in several directions. Here, we discussed only recognition of the 41 characters in Gurmukhi script. But this work can be extended to matras and other symbols used in Gurmukhi script also. Here, the font used is Gurbanilipi (true type), but since many fonts are available for Gurmukhi script, so it can be extended in the field of other fonts also.

BIBLIOGRAPHY

1) R.G.Casey and E.Lecolinet, "A Survey of Methods and Strategies in character segmentation", *IEEE Trans on PAMI, Vol 18, pp 690-706, 1996.*

2) G.S.Lehal and C.Singh, "Feature Extraction and classification for OCR of Gurmukhi script", *Vivek, Vol 12, No.2, pp2-12(1999).*

3) G S Lehal and Chandan Singh, "A Gurmukhi Script Recognition System", *15th International Conference on Pattern Recognition (ICPR'00) - Volume 2 p. 2557.*

4) N.Madan , Sunil Madan and H.Singh, " Hybrid Approach to character segmentation of Gurmukhi Script characters", *Proceedings of the 32nd applied imagery pattern recognition workshop,2003.*

5) M. K. Jindal, G.S. Lehal and R.K. Sharma, "A Study of Touching Characters in degraded Gurmukhi Script", *International Conference on Pattern Recognition and Computer Vision, PRCV 2005, pp. ?, 25-27 February 2005, Istanbul, Turkey*

6) U.Pal and Anirban Sarkar (2003), "Recognition of Printed Urdu Script", *IEEE, Proceeding 7th ICDAR.*

7) N.Tripathy and U.Pal,"Handwriting Segmentation of Unconstrained Oriya text",*Proceedings of the 9th workshop on Frontiers in Handwriting Recognition.IEEE, 2004*

8) U.Pal , A.Belaid and C. Choisy, " Touching numeral segmentation using water reservoir concept", *Pattern Recognition Letters, Vol.24,pp.261-272,2003*

9) U. Pal and Sagarika Datta. "Segmentation of Bangla Unconstrained Handwritten Text".*Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), 2003.*

10) U.Pal and N.tripathy, "Recognition of Indian Multi-oriented and Curved Text", *IEEE Trans, proceedings, ICDAR*

11) N.Arica and F.T. Yarman-Vural,"An Overview of Character Recognition Focussed on Offline Handwriting" *IEEE Transaction on system ,Man , and Cybernetics-Part C ;Applications and Reviews 3(2),216-233* .

12) Rajiv K. Sharma & A Singh , "Segmentation of Handwritten Text in Gurmukhi Script", *International Journal of Computer Science and Security, volume (2) issue (3)*

13) B.A. Srinivas, A.Agarwal and C.R. Rao (2008) "An Overview of OCR Research in Indian Scripts ".*International Journal of Computer Sciences and Engineering Systems, Vol.2, No.2.*

14) O.Trier, A.Jain and T.Taxt (1996) "Feature Extraction Methods for Character Recognition-A survey", *Pattern Recognition, Vol.29, No.4, pp.641-662.*

15) R.Teke, S.Malhotra"Recognition of Handwritten Devanagri Numerals".*International Journal of Computer processing of oriented languages*

16) D.Das and R.Yasmin (2006) "Segmentation and Recognition of Unconstrained Bangla Handwritten Numerals", *Asian Journal of Information Technology 5(2), pp.155-159.*

17) Veena Bansal and R.M.K. Sinha. "Segmentation of touching and Fused Devanagari characters ", *Pattern recognition, vol. 35: 875-893, 2002.*

18) U. Pal and B. B. Chaudhuri, "Indian script character recognition: A Survey", *Pattern Recognition*, *vol. 37, 2004, pp. 1887-1899*.

19) Rajean Plamondon, Sargur N. Srihari. "On – Line and Off – Line Handwritting Recognition: A Comprehensive Survey", *IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol 22(1). Janurary, 2000*.

20) U. Pal, S. Sinha and B. B. Chaudhuri. "Multi-Script Line identification from Indian Documents", *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR) 2003.*

21) Prashant Beri, " Noise Removal and segmentation of handwritten in Gurmukhi script using complete fill method", *an M.tech thesis submitted to Department of Computer Science and Engg, Punjabi University,Patiala,2005*

22) Sunil Madan, " Skew correction and character segmentation for printed text in Gurmukhi Script", *An M.Tech thesis submitted to Department of computer science and Engg, Punjabi University ,Patiala,1997*

23) B.Verma "A Contour Code Feature based Segmentation for Handwriting Recognition". *Proc. 7th ICDAR., 2003*

24) N.Tripathy and U.Pal,"Handwriting Segmentation of Unconstrained Oriya text",*Proceedings of the 9th workshop on Frontiers in Handwriting Recognition.IEEE, 2004*

25) S.kumar, "A technique for recognition of printed text in Gurmukhi script", *M.Tech thesis, Punjabi University, (1997).*

26) K.Kaur, "An Approach towards the recognition of printed Gurmukhi Script", *M.Tech thesis, Punjabi University, (1999).*