ANALYSIS AND RETRIEVAL OF SCANNED DOCUMENTS

USING WORD SPOTTING TECHNIQUES

BY

MUHAMMAD RASHID HUSSAIN

2012-NUST-DirPHD-CSE-18

THE DISSERTATION

SUBMITTED TO

NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SOFTWARE ENGINEERING

SUPERVISOR

DR ASIF MASOOD

DEPARTMENT OF COMPUTER SOFTWARE ENGINEERING

MILITARY COLLEGE OF SIGNALS
NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY PAKISTAN

2017

Acknowledgements

Allah be Praised for His Blessings and Guidance. This thesis is the result of four years of dedicated work which won't have been possible without the support of many.

First of all, I would like to extend a word of thanks to my very knowledgeable and outstanding professional supervisor, Dr. Asif Masood, who is ranked amongst the best in the domain. I was always led by his skillful guidance and helpful suggestions, which made this all possible. Patience, dedication, and constant encouragement of my supervisor remained an asset for me which I will cherish throughout my life.

Secondly I would like to thank my GEC Members; Dr. Fahim Arif, Dr Imran Ahmed Siddiqi and Dr Khurram Khurshid. I am lucky that I had such distinguished professionals and extraordinary human beings who guided me all the way along and without their appreciation, guidance and support; this won't have been possible. They were always there; whenever I needed them. I have no words to thank them for their contribution in making this possible.

Thirdly I would like to thank the Commandant, faculty at Military College of Signals (MCS) for providing extremely conducive environment of learning and research. My prayers and best wishes for complete team of MCS.

Lastly, certainly not the least; Thanks to my family, parents, brothers and sister; for their prayers.

Dedicated to my mother

whose unwavering faith in me has been the driving force in my life, and her unconditional love and prayers remain my unremitting source of strength

Publications

1. **Muhammad Rashid Hussain**, Ahsen Raza, Imran Siddiqi, Khurram Khurshid, and Chowki Djeddi, "A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation," *Eurasip Journal of Image and Video Processing*, 2015. ***(Impact Factor 1.06).***

2. **Muhammad Rashid Hussain**, Asif Masood, Haris Ahmad Khan, Imran Siddiqi, Khurram Khurshid 'Language Independent Keyword Based Information Retrieval System of Handwritten Documents using SVM Classifier and Converting Words into Shapes' Journal of Engineering and Applied Sciences, 2016. ***Category X Journal.***

3. **Muhammad Rashid Hussain**, Haris Ahmad Khan, Imran Siddiqi, Khurram Khurshid, Asif Masood 'Keyword based Information Retrieval System for Urdu Document Images' 11th

IEEE International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2015, ***Thailand***.

4. **Muhammad Rashid Hussain**, Asif Masood 'Word Segmentation of Handwritten English Text for Improvement of Word Spotting Results' International Conference on Image Processing, Production and Computer Science, ICIPCS 2016, ***UK.***

Contents

List of Figures

List of Tables

List of Abbreviations

SVM	Support Vector Machine
IT	Information Technology
SOP	Standard Operating Procedure
RLSA	Run Length Smoothing Algorithm
OCR	Optical Character Recognition
DTW	Dynamic Time Warping
PCA	Principal Component Analysis
ROC	Region of convergence
CNN	Convolutional Neural Networks
BCC	Basic Connected Component
HOG	Histogram of oriented Gradient

ZOI	Zones of Interest
CCs	Connected Components

Abstract

Writing is a codified system of standard symbols: the repetition of agreed-upon simple shapes to represent ideas. Language using symbols is assumed to be universal which is easier to interpret and efficient to use. Handwriting has remained one of the most frequently occurring patterns that we come across in everyday life. Handwriting offers a number of interesting pattern classification problems including handwriting recognition, writer identification, signature verification, writer demographics classification and script recognition etc. There is a dire need to address these problems and all out efforts be made to devise a script independent framework that can be applied globally to maximize the advantages of wealth of knowledge contained in the form of handwritten scripts. Lot of research in this area is ongoing. The work presented here is a document indexing and retrieval system using word spotting as the matching technique. Word spotting presents an attractive alternative to the traditional Optical Character Recognition (OCR) systems where instead of converting the image into text, retrieval is based on matching the images of words using pattern classification techniques. Proposed system relies on extracting words from images of handwritten documents and converting each word into a shape represented by its contour. Conversion of words into shapes is an innovation proposed in our framework that will set new avenues of research; as this work has not been experimented before in the history of word spotting. A set of multiple features is then extracted from each shaped word and instances of the same word are grouped into clusters. These clusters are used to train a multi-class Support Vector Machine (SVM) which learns different word classes. The documents to be indexed are segmented into words and the closest cluster for each word is determined using the SVM. An index file is maintained for each word cluster which keeps information on the documents containing the respective word along-with the word locations within each document. A query word presented to the system is matched with the clusters in the database and the documents containing occurrences of the query word are presented to the user. The system evaluated on the handwritten images of IAM database reported promising precision and recall rates. Enhancement of feature vector space by introducing new set of features is also a major contribution. Study has also been carried out to analyze the contribution and significance of different features employed in our study. Use of most relevant feature vector through employment of Principal Component Analysis (PCA) has also been applied to condense the dimensionality. The proposed framework has also been successfully tested in extremely challenging / cursive Urdu language scripts. Promising results in both English and Urdu scripts amply proves script independence that can be applied globally.

Chapter -1
Introduction

Evolution of mankind and development of diverse cultures is a history spanned over millions of years. Putting ideas in writing and saving it for future even in today's modern era is the best form of expression. Recognizable systems of writing developed in 3 major cultures within 1200 years of each other. Sumerian cuneiform developed around 3000 BC, Egyptian hieroglyphs around 2800 BC, and the precursor to Kanji Chinese around 1800 BC. The development of writing allowed cultures to record events, history, laws, theories in math, science, medicine; create literature and more. Simple pictographs were used to represent people, places and things. As the needs for communication expanded, different pictographs were combined to represent ideas, and required knowledge to interpret the new symbols. These became ideographs: abstract symbols that evolved beyond the original drawings. With the tremendous advancement in the field of

Information Technology, human cultures and ways of living are being transformed in parallel. Various techniques are being used / researched to make life relatively easier. One such difficulty arises once humans write certain things using their own handwriting / language and want to take maximum advantage of their effort. There are roughly 6,500 spoken languages in the world today. However, about 2,000 of those languages have fewer than 1,000 speakers. Lots of information wealth exists in handwritten documents which have been preserved using scanning thus ultimately converting them to images. Various techniques exist in making such images searchable. One such technique is using Optical Character Recognition (OCR) which has gained popularity in recent past. OCRs are being developed for different languages but the applicability is mostly limited to English language with some accuracy. Research is ongoing in this field. So a need arises to develop a framework which should be script independent and can be used for various languages around the

globe. This work presents a document indexing and retrieval system using word spotting as the matching technique. In word spotting, retrieval is based on document images by employing pattern classification techniques. Segmentation is undertaken wherein each word is extracted after removing errors / mistakes and the extracted word is innovatively converted into shape represented by its contour. Multiple features are then extracted from each word and instances of the same word are grouped into clusters. Clusters are used to train a multi-class Support Vector Machine (SVM) which learns different word classes. Indexing the document takes place wherein words are segmented and closest matched cluster is found for each word using the already trained SVM. Index file is maintained for each word cluster which keeps information on the documents containing the respective word along-with the word locations within each document. A query word presented to the system is matched with the clusters in the database and the documents containing occurrences of the query word are presented to the user.

Various dynasties ruled the Indian sub-continent and left behind enormous and rich cultural heritage that also included intellectually enriched research in the shape of various documents scripted in Urdu. In order to provide efficient access to this knowledge, analysis through digitizing the existing work is the need of hour. In addition to digitization, efficient search mechanisms are also needed to be implemented to provide users a rapid access to the queried information. In most cases, the digitized documents are complemented by manually assigned tags which not only are a time consuming task but also provides a very limited search facility. Automating the transcription of these documents using Optical Character Recognition (OCR) systems is also challenging due to the very complex cursive nature of Urdu text. Keyword spotting based information retrieval system for document images has been introduced to overcome these limitations. The proposed technique relies on two major modules, document indexing and retrieval. Images of documents are segmented into partial words (ligatures) and identical partial words (PWs) are grouped into

clusters. Our proposed technique introduces the concept of considering each (partial) word as a unique shape and a set of shape descriptors is extracted to characterize the PWs. The clusters of PWs are used to index a given set of documents. During retrieval, the query word presented to the system is matched with the clusters in the database and all documents containing instances of the query word are retrieved and presented to the user.

Proposed framework has been evaluated on IAM and Urdu Database and realized promising precision and recall rates. Our proposed framework ensures script independence and can be applied globally as each script can be easily converted into shape represented by its contour and matching / retrieval can be done with profound efficiency.

1.1 **Motivation**

Handwritten and cursive fonts like Urdu pose a number of challenges which reduces their usability and ultimately results in their limited applicability. In this information era, such challenges needs to be minimized and hidden wealth of information in the shape of scanned / handwritten images be made available to the researchers around the globe. Our proposed framework is efficient, script independent information retrieval system that can be used with any type of script. Introduction of innovations like converting words into shapes for superior matching and using newly added features difference in area (delta feature) and shape feature explained in subsequent chapters have made the system more robust and reliable.

1.2 **Research Challenges**

Working with handwritten documents and cursive fonts is an extremely challenging area. Some of the research challenges are:-

a. Inter and intra class variability in writer's handwritten texts.
b. Cursive nature, difficult word formation approach, appearance of characters within words, overlapping of partial words, dots and diacritic marks, segmentation and varying styles poses enormous challenges.
c. Indexing and search is either not present or is not reliable.
d. Content is not structured and often lacked the desired quality.
e. One book / publication may be divided into a sizable number of digital media storages, thus increase in storage area.
f. Divided contents on different media making it difficult to cross reference.

1.3 **Online Access to Digital Documents**

Presently, in order to aide users across the globe, various organizations / companies provide access to information through digitizing books mostly through scanning / meta-tagging and searching through content available inside the image is non-existent. One of the biggest projects is being undertaken by renowned search engine company Google and is named as Google Book Search (Figure 1). Initially scanning is undertaken and then search option / facility is provided to subscribers / users. OCR Technology is used which converts images into text.

Complete process is stored in database.

Figure 1- Screen shot of the Google book search service where a book is opened and its text can be searched using the text search option

Additional resources that are currently providing online sevices for document search and retrieve options include ‘Universal Digital Library 8’ operated by Carnegie Mellon University USA and Europeana 9 sponsored financially by European Union. Keeping in view its enrichd wealth of knowledge; many Libraries around the globe are trying to make these resources accessible online for better utilization. Few examples are Gallica (bibliothèquenationale de France) , BVH11 (Bibliothèques Virtuelles Humanistes) of the Centre d'Études Supérieures de la Renaissance Tours and Medic@ of the bibliothèque interuniversitaire de medicine Paris [BIUM].

1.4 **Context of the Work being Presented**

Scanned documents hold immense experiences and knowledge and inspire educated / noneducated people. Searching through the use of query word, thus getting relevant answers, saves time and effort of scanning a complete book / document. Proposed framework offers a detailed analysis of numerous retrieval methods for scanned scripts / images that permits queries in the

form of a word images. Based on the matching of features after training by classifier, closest matched cluster and its instances will be retrieved and presented to the user.

1.5 **Problem Statement**

Motivation of this work stems from the work already done by Khurshid et al. [1] in the same field. However, existing work suffers from following limitations:-

a. Existing work is script dependent and primarily focused on French language. Script independence has not been achieved.

b. In the existing work carried out, query words are matched with each word in each document in the database. Concept of clustering is not present.

c. Feature set employed is not robust and needs enhancements.

d. Work done was not tested on handwritten documents and mostly printed documents were used.

1.6 **Contribution Objectives / Research Goals**

Effort has been made to contribute in all phases of research. Lots of research work has already been undertaken in binarization so our endeavor was to use the already binarized database and focus on challenges faced by handwritten / scanned images. However, we did carry out binarization on printed documents and tested prevalent techniques to accrue their efficiency. Development of script independent word retrieval framework was the major focus. Our Contribution will be focused in following major areas which required improvement in the existing work:-

a. Detailed analysis of existing handwritten databases. Lots of databases exist and it becomes extremely difficult for a new researcher to decide which database is relevant to his research

area. To this end successful accomplishment of comprehensive survey of handwritten document benchmarks: structure, usage and evaluation have been undertaken. Publication in Eurasip Journal of Image and Video Processing, 2015 has been undertaken. This journal is ISI Indexed and current impact factor is 1.06.

b. Development and implementing script independent framework for document retrieval system. Framework includes, segmentation, indexing, enhancement of feature vector space, matching and retrieval. Journal publication has been successfully completed in Journal of Engineering and Applied Sciences, 2016.

c. Segmentation plays a very important role in word spotting. Removal of errors present in printed documents particularly in handwritten images is a major challenge. Our proposed framework addresses all major issues like, varying style of writers, slant lines, punctuations, overlapped characters etc. Paper on the subject has been published in International Conference on Image Processing, Production and Computer Science ICIPCS 2016.

d. Urdu language is extremely cursive and poses enormous challenges in word spotting. Formation of clusters, indexing, matching and retrieval stages were successfully implemented using script independent framework. Publication has been done in 11th IEEE International Conference on Signal-Image Technology & Internet-Based Systems SITIS 2015.

e. Our proposed approach will use the concept of clustering and classification for efficient searching mechanism. Instead of matching each word with all words of the document, matching it with contents of training clusters has been undertaken. Complete scanned document is segmented and uploaded in the database and then Indexing is done by realization of each word location. Improvement has been made in this particular approach.

Cluster contains all possible combinations of a word / character, so for each word there is a separate cluster. Query word is matched with the cluster only; and once the matching part is done, than all the words contained in the document will be highlighted through reference maintained in the database.

f. New Features have been added to make feature vector more robust. Converting words into shapes, calculating difference in area before and after the application of structuring element etc. have made enhanced our research.

1.7 **Organization of Thesis**

The thesis is organized in a sequential way starting with chapter 2, wherein, the explanation of state of the art approaches in document retrieval, methods classification is deliberated in detail to give a complete idea of research work prevalent in the field. In chapter 3, discussion on pre-processing (binarization and segmentation) has been undertaken, wherein, problems faced in segmenting handwritten and Urdu fonts alongwith our proposed resolution techniques will be presented . Chapter 4 will focus on feature extraction and Clustering. Enhanced Features set is also part of discussion. Chapter 5 deals with matching and retrieval process. For matching two words, multi-class Support Vector Machine Classifier has been used.

Chapter 6 describes experimental results obtained from our proposed framework. Results of Binarization, Segmentation, and Retrieval using various criteria's have been discussed. Principle component analysis has also been undertaken. Chapter 7 includes conclusion and future work. It is worth mentioning here that work done on handwritten documents in English language and corresponding application in Urdu language have been described in parallel to aide the user in getting a feel of proposed framework alongwith its intricacies and applicability.

Chapter -2
Literature Review

Automation of handwritten textual documents is an active research area [2, 3] wherein searching for a particular word in complete document might be necessitated[4]. Application of a particular search algorithm in an enterprise wide environment does not hold good for handwritten documents due to varying styles of writing, overwriting, joint words etc. Although latest technologies like OCRs have yielded good results in printed documents, yet its performance in handwritten documents remains a challenge. Due to non-availability of OCR for most of languages, its applicability is currently limited in scale. Word Spotting technique has been introduced to overcome the limitations of OCR technology and is subdivided into two categories; holistic and analytical recognition techniques. Holistic word recognition techniques employ words as a whole in which a local feature sequence (pixel length of each word) is extracted and then matched with remaining words of document[5, 6]. In Analytical (character based) Techniques a word is sub-divided into finer grains called characters which operate independently or in groups[7-10]. The basic technique is to extract features from handwritten documents which should be robust to various writing styles and then apply supervised classification by employing classifiers like Neural Networks [11-13] and Support Vector machines (SVM)[14, 15].

State-of-the-art work in this field includes systems that are based on Keywords; in which learning phase is used with promising results [16, 17] but they suffer from major limitation of requiring enormous dataset. An efficient retrieval mechanism was proposed for unconstrained handwritten document retrieval[18], Comprehensive empirical evaluation of handwritten text line detection methods was undertaken by evaluating four different handwritten text line detection algorithms,

on four different databases and using three different metrics[19]. Identification of script of handwritten text by employing Directional Discreet Cosine Transform was proposed in[20].

Lots of work has been done and plenty is in progress in the domain of information retrieval in printed documents. Recent research area shows applicability of Convolutional Neural Networks (CNN) architecture for word spotting. By using the recently proposed Pyramidal Histogram of Characters (PHOC) as labels, this CNN is able to achieve state-of-the-art performance in Query-by-Example as well as Query-by-String scenarios on different datasets [21]. Word images and text strings are embedded in a common vectorial subspace, thus resulting in faster computation especially for comparison purpose [22]. Handwritten text-to-speech system has been proposed using query-by-speech keyword spotting system [23]. Query-by-string methods has been proposed for full segmentation-free decoding framework that does not require any presegmentation on word or line level [24]. Issues and problems related to printed documents which are discussed in detail in[25-27]. These include the physical issues such as quality of the documents, the marks and strains of liquids, inks and dust, etc. and the semantic issues such as foreground entity labeling. Perspective falsification alongwith shadow is generally observed once huge volumes of historical books are scanned resulting in deformation of characters / words [27]. Analogous type of spurious components are also detected in old documents [28]. Such problems yield opportunity for researchers to make efforts for better performance in this active research area.

Diverse methods have been suggested to make document images more interesting / productive and work encompasses various stages like segmentation (including binarization, text/graphic segmentation, text words extraction, etc.) as well as for information retrieval. In current chapter, brief discussion on work done for document image segmentation and information spotting has

been undertaken. Section 2.1 analyzes the different categories of document segmentation methods and section 2.2 focuses on the different information retrieval methods.

The above mentioned issues are addressed with various techniques. Manually annotating the pages / documents and creating an index is one choice. Transcription is another technique in which full text search is made possible through text search engines commercially available. Transcription suffers from the limitation of degradation in performance with the increase in the database. Optical Character Recognition (OCR) systems have been devised to cater for reading and digitizing the printed / scanned documents. Commercially available OCR software's works well with good quality documents. Ancient / historical documents suffer from limitations like torn pages, faded ink, poor quality pages etc, thus making these OCRs virtually impracticable for usage.

Document images have many issues [25-27] including low quality, marks and strains of liquid, foreground entity labeling, perspective distortion, warping of words in shadow[27]. Old manuscript have numerous spurious components[28].

One of the techniques is page segmentation method and it's sub-divided in three main categories: bottom-up, top-down and hybrid[29-33]. In bottom up approaches, moving from smaller details towards bigger pictures is done. Pixels are recursively used to connect the components and move towards building the structures. Bottom-up approaches use techniques like neural networks [27] connected component analysis[34] active contours[35], region-growing methods include runlength smoothing algorithm[36]. These methods normally suffer from accumulation of errors impediment. In Top-down approach, moving towards finer grains from bigger picture is done.

The most well-known methods are form definition languages, histogram analysis[37], rule based systems[38], or space transforms (Fourier and Hough transform)[39], projection methods [26] and variations like rectangulation, white streams[40]. Top-down methods are faster in execution, but prior knowledge of document class and type is required for efficiency. In hybrid technique, mix

and match of bottom-up and top-down approach is used. Hybrid techniques include Gabor filter method[41]. Wavelet and fractal analysis, auto correlation function[33], texture-based methods[31, 42]. Hybrid methods work well for graphic / text segmentation but lack the desired degree of accuracy once it comes to finer level of segmentation as prevalent in historical books. Due to the limitation of OCR with respect to handwritten scripts, word spotting technique has been employed on handwritten manuscripts[43]. Word spotting can be divided into sub categories based on a number of distinctions. First on the list is segment-based or segment free methods [43-48]. Rothfeder et al. [49]divided all word spotting in two main categories – image and feature based techniques. Image based techniques use template matching using correlation by working directly on pixels[50], on the other hand feature-based matching methods computes features before matching.

Word spotting is subdivided into two categories; holistic and analytical recognition techniques. Holistic word recognition techniques [43, 44, 51] are segmentation free technique capable of handling noise effectively, operates on a word as a whole; keeping in view the fact that humans recognize the words easily [52] whereas analytical techniques are segment based and operates on characters extracted from words.

2.1 **Research Based on Holistic (word based) Techniques**

Li *et al.* [53]proposed a retrieval algorithm based on holistic, in which a local feature sequence (pixel length of each word) is extracted and then matched with remaining words of document. Andreev and Kirov [54] used a customized Hausdorff distance for matching two words in image space. Harris Corner detector is used for finding out the points of interest for each word image and then matching is done in[49]. Marinai *et al.* [55]used self-organizing maps based on clustering and principle component analysis. Run-Length Smoothing Algorithm, RLSA has been

used for words extraction based on their aspect ratio and clustering is performed on each subset of partition. The work has been extended by Marinai et al. [56]through building a framework for document retrieval in digital libraries. Zagoris *et al.* [57]used a combination of Fourier transform and scalar characteristics for matching based on Euclidian distance. Similar research was undertaken by [45] for Greek ancient documents Cross correlation matching has been used for holistic technique by[58]. Adamek *et al.* [44]proposed a closed contour matching technique using DTW. Shunyi Yao et all [59] proposed histogram of Oriented Gradient (HoG) based two-directional Dynamic Time Warping (DTW) matching method for handwritten word spotting. Word shape coding technique was suggested by Bai *et al.[60]*, Bertolami *et al.* [61]and Adamek *et al.[44]*. Kluzner *et al.* [62]proposed an adaptive OCR system based on clustering by using FineReader engine performing simultaneous recognition for a cluster.

2.2 **Research Based on Analytical (character based) Techniques**

In this technique, a word is segmented into finer grains called characters which operate independently or in groups[8, 10, 63]. Choice of a particular segmentation point is dependent on accuracy by which a character is recognized. Unlike holistic technique where word itself becomes a segment point, in analytical approach multiple methodologies are being proposed to meet this challenge. One simple way is to break a word into smaller units and then these subunits are recognized[64]. Analytical methods are more flexible being letter oriented [48] as compared to holistic due to size and nature of lexicon. Vamvakas *et al.* [8]generated ASCII files based on training set. Image is first segmented into words, and through bottom up approach, characters are extracted using connected component analysis and skeleton features. Each character image is, first, normalized to fit in a pre-defined window size and then is represented by a fixed length feature vector based on the character's zone and area properties.

Moghaddam and Cheriet [9] presented a connected component based method for word spotting on cursive Arabic scripts by generating a basic connected component (BCC) library which is clustered based on SOM. New connected components are matched with existing library. DTW along with histogram matching is used. Terasawa *et al.*[10] presented Histogram of oriented Gradient (HOG) method where HOG Feature is calculated through rectangular sliding window in the direction of textual information. Histogram is calculated with orientation evenly spaced. It was found that that most effective number of bins is obtained to be either 12 or 16. Authors presented two alternate approaches for textual indexation of old documents i.e. Word Spotting and computer assisted transcription. Computer assisted transcription is based on character pattern redundancy in document images. Characters are first segmented and compared to one another to create a pattern dictionary. All characters within a class have a same label which removes the requirement to recognize all the characters of a class. Operations are performed on the word image to get the guides and the enlarged bounding. Manual transcription of 50% of the pattern dictionary, the authors achieved a correct transcription of 80% of an entire book (200 pages and 2000 characters per page) in 3 hours.

For word spotting, the authors experimented a number of differential features and designated the gradient angle based on best P-R curve. Applicable Zones of Interest (ZOI) were selected to escalate efficiency / accuracy. Enlarged bounding boxes and guides were obtained through the application of morphological operations.

Template is divided into pieces using the ZOIs and the distance and orientation between the centers of these zones are kept. The first ZOI is compared chronologically with each ZOI of the test image while the later ZOIs of the template are matched through lesser displacement possibility. The authors report results of two experiments. The first one is carried out on 185 images of two column pages with 20 lines per column where the search of the word ‘fyon’ found 28 good hits out of 28

and the time of execution was 260 minutes. On the same document images, the authors then searched the word ‘egypt’ and found 15 occurrences out of 15 with the first bad hit at rank 13 and the last good hit at rank 68 and this search took 400 minutes. Overall, results achieved by this approach are satisfactory but the execution time is too long which makes this approach less lucrative. Having presented an overview of some of the recent and notable studies in the area, comparative analysis of these methods is undertaken in the following section.

2.3 **Comparison of the Retrieval Methods – Discussion**

There have been no standard criteria for the evaluation of different information retrieval / word spotting algorithms. Terasawa *et al.* [10]proposed an automatic evaluation framework for word spotting algorithms which provides certain guidelines and protocols for obtaining a uniform standard in the evaluation process. If followed globally, it can prove to be useful in future for a better and fair comparison of different methods. Traditionally, retrieval techniques have generally been evaluated by running a set of queries and analyzing the list of retrieved words. Two most common measures for judging the quality of the retrieved-words list are recall and precision. But as all these methods have been evaluated in different conditions on different proprietary databases sets with different number of test and query images, comparing the listed precision and recall rates portrays no true picture of the performance and efficiency of a particular method with respect to others, thus rendering a quantitative analysis irrelevant. A way though could be to program all these algorithms and then run the tests for the different methods in exact same computational conditions on one common data set. But due to many real life constraints, this is not feasible for us. So we will limit ourselves to a brief qualitative analysis of the different methods that has already been discussed. Two main categories of the document image retrieval methods which are holistic or segmentation-free techniques and analytical or segmentation-based techniques are most

noteworthy. Both the techniques have their pros and cons. Holistic analysis methods compare a sequence of observations derived from a word image with similar features for the words in the database. There are many factors that make this approach very attractive and natural. These factors mainly concern the poor quality and printing of historical documents that result in high level of noise, irregularity in printing and different font variations in ancient printed texts. All these factors can complicate the character segmentation process which itself is not an easy task. By using a holistic approach, all the character segmentation issues can be avoided and matching with whole words directly with good recognition rates using the different methods discussed earlier in the chapter. On the other hand, analytical methods look for the best match between consecutive sequences of primitive segments or characters of a word. Though segmentation of characters is a very difficult problem, especially for ancient documents[66], the analytical approaches usually tend to give better recognition results as compared to their holistic counterparts. It is due to their ability to focus on the local intrinsic characteristics of words which give more in depth details of a word, thus differentiating between two different words becomes easier. Another important point is the low level matching, which takes more intrinsic details into the matching process, making the systems robust. Thus similar feature set and matching distance could give better results when these features are defined for characters as compared to same features defined for words as evaluated in[67].

Another major advantage of all segmentation-based methods is their flexibility with respect to the size and nature of the lexicon, which is a result of these methods being letter-oriented. From the above discussion, it can be concluded that although a lot of work has been undertaken in this active research area, yet a lot of scope exists for researchers. Handwritten scripts alongwith cursive writing like Urdu font is yet another domain which is best dealt with word spotting . Script

independence is the need of hour and our proposed framework amply contributes in all phases of word spotting.

Chapter -[1]
Overview of Proposed Framework

3.1 Introduction

Careful study of literature reveals certain important conclusions which form the basis of this research work. One of the major conclusions is strong dependence on script for undertaking any meaningful work particularly in the field of hand writing. This infers that there is a need to develop a script independent system using robust features that ensures language independence. Keeping this important conclusion / limitations in the existing research; an endeavor has been made to develop a script independent framework with capability and capacity to contribute in all phases of framework.

---

[1] .2 Overview of Proposed Framework

The complete system architecture is designed in two phases. In first phase, reference database is acquired through segmentation of training images and clusters are formed in a semiautomated manner. Features are extracted from clusters and are used to train the SVM Classifier to create the reference dataset as illustrated in Figure 2. In the second stage indexing and retrieval of words is performed where documents that are to be indexed are pre-processed to extract words from them and later extract features from each word. Every word being segmented-out is classified on the basis of reference database obtained from stage-1, and is assigned to a unique cluster. Each cluster contains an Index file which is updated every time when indexing of a new document is performed. This index file acts as reference to specific word in the document which is vital for word detection in large documents database. In the

retrieval system, features of query image are extracted and classified using SVM which allocated the query word to a specific cluster. Information contained in the indexed file at that cluster is used to retrieve all documents from database which contain the query word as illustrated in Figure 3.

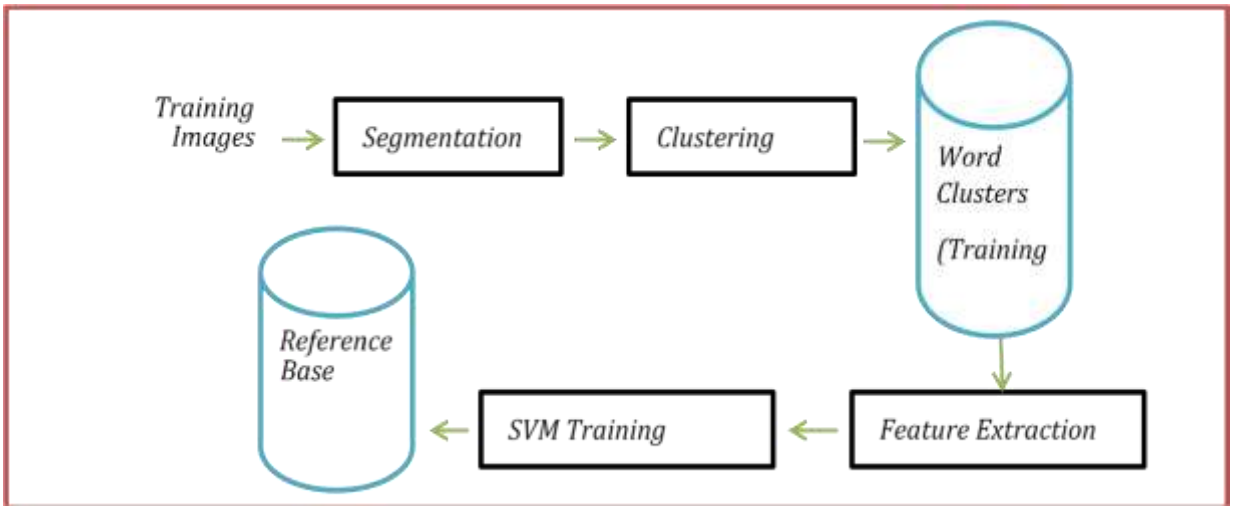


Figure 2- Overview of Proposed Methodology (Obtaining Reference Base)

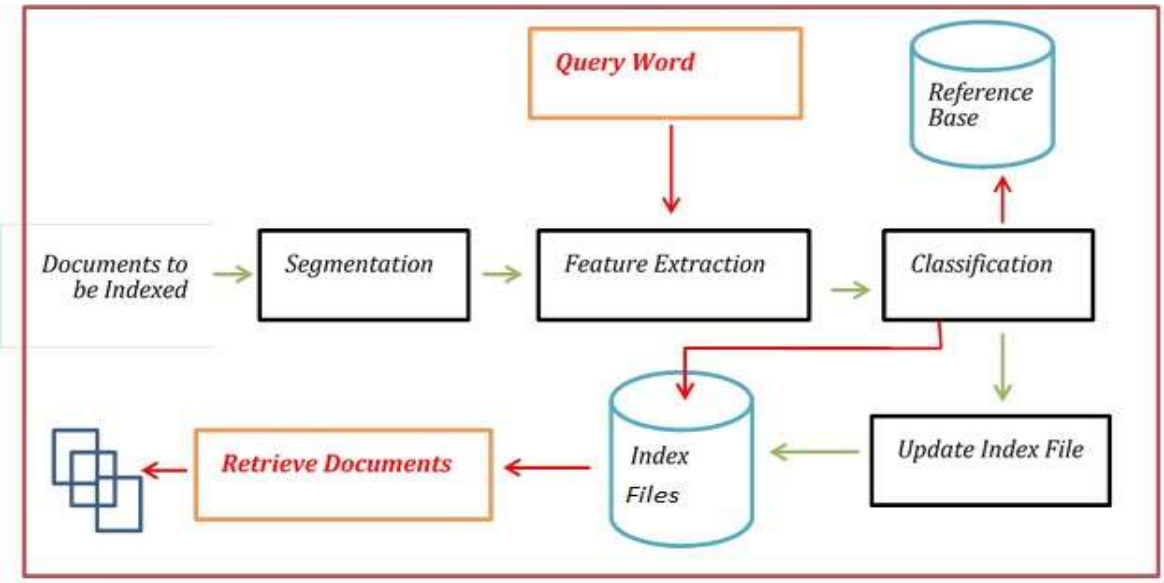


Figure 3 – Overview of Proposed Methodology (Indexing and Retrieving Images)

3.3 **Description of Each Phase of Framework**

3.3.1 **Segmentation**

Segmentation is a critical phase of framework; wherein, words are extracted from scanned document / image after removal of errors present in that image. Proper segmentation is necessary as any error introduced in this stage will have pronounced results in later stages.

Detailed discussion on the segmentation is covered in Chapter 4.

3.3.2 **Clustering**

Clustering refers to placement of each extracted word in a unique cluster containing instances of similar characteristic words. Words that have common features will be placed in a single cluster. Correct assignment is essential to reduce errors in later stages particularly in retrieval phase. Each unique cluster has its own index file that contains references to the location of extracted words in the documents. Detailed discussion on clustering is covered in Chapter 5.

3.3.3 **Feature Extraction**

Features are the identification marks of any word. Each extracted word / query word has its unique feature set which is obtained through employment of various innovative techniques both global and local. In our proposed framework; existing features as well as newly added features have been used to their best utility. Robust feature set is essential for indexing and matching. Detailed discussion on clustering is covered in Chapter 5.

3.3.4 **Classification**

Classifiers are used to train the dataset and ultimately retrieve the query word basing on parameters. Numbers of varying classifiers are prevalent in today's research world. There is no single classifier that suits all; and the choice depends on implementation framework. In our

proposed methodology; SVM Classifier has been used due to its strong capability to handle multi class variability. In our implementation we have each word / cluster as a unique class which requires separability for identification / matching. SVM has also probability estimation function which is extremely helpful in making a decision in identification / discarding any query word. Discussion on use of SVM is covered in Chapter 5.

3.3.5 **Indexing**

After SVM is trained on training data set; documents that are to be indexed passes through segmentation, feature extraction phase and are assigned to appropriate cluster. Index file present in each cluster is updated with each incoming word. Discussion in this regard is covered in Chapter 6.

3.3.6 **Matching and Retrieval**

Once the documents are indexed and index file has been updated; any query word can be presented to the system. In order to facilitate the user, a GUI has been designed. Query is given as an input image. Features are extracted and trained SVM looks up in the reference database that was obtained as a result of training. Nearest match and cluster is indicated that contains the instances of closely matched image. If the probability estimation function is within the desired threshold, framework reads the index file present in that cluster and retrieves the documents that contain the query word. If threshold is not met; the query word is discarded and 'No Match' message is displayed. Detailed discussion alongwith Results are covered in Chapter 6 & 7.

3.4 **DataSet Used in the Proposed Framework**

In this research work, first challenge was to carry out a detailed analysis of existing handwritten databases. To this end successful accomplishment of comprehensive survey of handwritten document benchmarks: structure, usage and evaluation has been undertaken [68] as shown in Table 1 & 2. Using the above knowledge IAM handwriting database [69] has been used.

Total of approx. 1000 clusters dataset have been used either for training the classifier or retrieval of relevant results. Sufficient number of dataset was used to train the classifier both for English and Urdu languages and then based on training, a different set of clusters were used for querying purpose. For Urdu language use of 100 handwritten documents and also downloaded Urdu text from websites has been done. The idea was to use max cursive words for accurate results. Sample Images are attached as Annexure-A.

Database	Year	Language	Content	Mode	Writers	Statistics
PE92	1993	Korean	Isolated characters	Offline	500+	235,000 characters
CEDAR	1994	English	Words, Characters, Digits	Offline		10,570 Words, 27,835 characters, 21,179 digits
NIST	1995	English	Isolated Digits	Offline	3000	810,000 digits
JPCD	1997	Japanese	Characters	Online	80	1,227 character categories
MNIST	1998	English	Isolated Digits	Offline		70,000 digits
Al-Isra	1999	Arabic	Sentences	Offline	500	500 sentences,37,000 words,10,000 digits
IRONOFF	1999	French	Words, Characters, Digits	Online		50.000 words, 32000 characters
Firemaker	2000	English	Paragraphs	Offline	250	4 samples/writer
GRUHD	2001	Greek	Text, symbols	Offline	1000	1,760 forms, 667,583 symbols, 102,692 words, 123256 digits
IAM	2002	English	Sentences	Offline	657	1,539 forms, 5,685 sentences, 115,320 words
IFN/ENIT	2002	Arabic	Words	Offline	411	2,265 forms,26,449 city names
Checks DB	2003	Arabic	Check amounts	Offline		7,000 cheques, 29,498 subwords, 15,000 digits
AHDB	2004	Arabic	Sentences, check amounts	Offline	100	105 forms
ARABASE	2005	Arabic	Sentences, words, letters	On/Off	400	400 forms
TAM-On DB	2005	English	Sentences	Online	221	1,700 forms; 86,272 words
Numerals DB	2005	Bangla Devanagari	Digits	Offline		45,948 numerals
IAUT/PHCN	2008	Farsi	Isolated Words	Offline	380	1,140 forms; 34,200 words

RIMES	2008	French	Sentences	Offline	1,300	12,723 Pages

Database	Year	Language	Content	Mode	Writers	Statistics
IFN Farsi	2008	Farsi	Words	Offline	600	7,271 words; 23,545 sub words
CENPARMI-A	2008	Arabic	Words, characters, digits	Offline	328	13.439 digits; 21,426 characters; 11,375 words
LMCA	2008	Arabic	Words, characters, digits	Online	55	30.000 digits; 100,000 characters; 500 words
CENPARMI-U	2009	Urdu	Words, Characters, Digits	Offline		18,000 words
FHT	2009	Farsi	Sentences	Offline	250	1,000 forms; 106,600 words; 8,050 sentences
HCL2000	2009	Chinese	Characters	Offline	1,000	3,755 Characters
CENPARMI-F	2009	Farsi	Words, letters, digits	Offline	400	432,357 images
IAM-OnDO	2010	English	Text, Drawings, Tables etc.	Online	200	1,000 documents
RODRIGO	2010	Spanish		Offline		1,853 pages
ADAB	2011	Arabic	Words	Online	170	20,000+ words
CASIA	2011	Chinese	Text, Characters	On/Off	1.020	3.5M isolated characters; 1.35M characters in text
SCUT-COUCH	2011	Chinese	Characters	Online	190	3.6M characters
Indonesian TDB	2011	Indonesian	Sentences	Offline	200	200 forms
AMHCD	2011	Amazigh	Characters	Offline	60	25,740 characters
MRG-OHTC	2011	Tibetan	Characters	Online	130	910 character classes
KHTD	2011	Kannada	Sentences	Offline	51	4,000 lines; 26,000 words

KHATT	2012	Arabic	Sentences	Offline	1,000	1,000 forms
Devanagari DB	2012	Devanagari	Digits, Characters	Offline	750	5,137 isolated numerals
Database	**Year**	**Language**	**Content**	**Mode**	**Writers**	**Statistics**
UHSD	2012	Urdu	Sentences	Offline	200	400 forms
QUWI	2013	Arabic English	Sentences	Offline	1,017	4,068 forms
HaFT	2013	Farsi	Sentences	Offline	600	1,800 images
CVL	2013	English German	Sentences	Offline	311	2,163 forms
Tamil DB	2013	Tamil	Words	Offline	500	265,00 city names
AHTID-MW	2015	Arabic	Text lines	Offline	53	3,710 lines

Table 1 - An overview of handwritten databases

Task	**Databases**
Offline handwriting recognition	1AM, lAM-HistDB, RIMES, CEDAR, CVL,IFN-ENIT, AHDB, ARABASE, CENPARMI-Arabic, Allsra, LMCA, KHATT, QUWI, IAUT-PHCN, IFN-Farsi, CENPARMI-Farsi, HaFt, CENPARMI-Urdu, PE92, HCL2000, CASIA,KHTD, AMHCD, GRUHD,MRG-OHTC
Online handwriting recognition	lAM-onDB, IRONOFF, IBM-UB, ADAB,ARABASE, LMCA, SCUT-COUCH,CASIA
Digit Recognition	NIST, MNIST, CEDAR, CVL, CENPARMI-Arabic, IFN-Farsi, CENPARMI-Farsi
Offline Writer identification/verification	1AM, RIMES, CVL, FireMaker, IFN/ENIT, AHDB, Al-lsra, KHATT, QUWI, AHTID/MW,FHT, HaFT, HCL2000
Online Writer identification/ verification	lAM-onDB, IRONOFF, IBM-UB, ADAB
Word Spotting	1AM, lAM-onDo, lAM-HistDB, CENPARMI-Arabic, CENPARMI-Urdu,CASIA
Handwriting Segmentation	1AM, lAM-HistDB, CEDAR, AHTID-MW, FHT, HaFT,KHTD
Gender Classification	lAM-onDB, QUWI, HCL2000

Table 2 - Usage of Databases

Chapter -4 Pre-Processing (Binarization and Segmentation)

Mostly the work done in handwritten / cursive scripts lacks computational efficiency owing to in-efficient indexing, clustering and poor feature selection. Document is scanned and fed into the database and retrieval is obtained after matching features of query word with all the relevant features of each word in the database. This computational overhead increases with the increase in size of database. Concept of clustering i.e. grouping of each word in separate space is a major challenging area. Features of Query word should only get matched to a cluster and in turn, reference of each character/word is maintained in the document / database.

Recognition of unconstrained cursively handwritten text is still a widely unsolved problem and an active area of research. For large vocabularies and different writing styles in general[70], and for degraded historical manuscripts in particular[71], the accuracy of an automatic transcription is far from being perfect.

Handwritten text/writing is another challenging area, wherein different writing styles of individuals pose search, maintenance and segmentation problems. Problem in handwritten documents enhances not only due to variability between writers, but also due to variation in individual person's writing too. Indexing in handwritten text is a problematic area and requires enhancements in research.

Document image retrieval using word spotting is a popular research topic in document analysis domain after the advent of the digital library concept. Different approaches for word spotting that have been proposed over the years to facilitate information searching are discussed in chapter 2.

In our proposed work, employing a granular approach for word spotting by introducing dynamic matching approach at word level has been undertaken. For that, the first step is to perform the indexing of document

images that includes the segmentation of words and characters, defining features for the characters and storing all this information in individual data files. It is a time consuming process, that's why document image indexing is done beforehand /offline to facilitate users in rapid information search and document image retrieval.

The whole indexing process has been divided into following broad sub-steps:

a. First Step is Binarization. Lots of research has already been undertaken in this field. However, in order to test few images, we did carryout binarization using famous techniques which are explained in subsequent paragraphs.

b. Second step involves the segmentation of words. Details covered in Current Chapter.

c. Third Step is the feature extraction from segmented words and assigning them to clusters. Details covered in Chapter 4.

d. In the last step, the segmented word is matched with the training clusters and is assigned to a particular cluster where indexed file is generated and constantly updated for each incoming word.

4.1 **Document Image Binarization**

Binarization is a technique used in pre-processing stage wherein efforts are made to separate background from foreground. In an ideal state, resultant should be aimed at foreground text in black and background as white. Various methods exists that incorporates different thresholding criteria's, yet there accuracy in giving ideal results for all type of documents has not been achieved [72]. As an example; few algorithms might demonstrate promising results in documents having marks of strain while they might not be suitable for extremely low intensity variations.

Research in Binarization of documents is as old as document image analysis. To this end, number of researchers has proposed multiple Binarization algorithms in the recent past. Careful analyses of these algorithms classify them in two broad categories i.e. global thresholding and local

thresholding. In case of Global thresholding techniques; employment of solitary intensity threshold value i.e. one fixed threshold is used for one particular image. Method proposed by Otsu using global thresh-holding is still most commonly used. In this method, the threshold value is calculated using some heuristics based on global image attributes. This helps in classification of image pixels into foreground (text) or background (non-text) pixels[73]. Global thresholding techniques have a glaring limitation i.e. these methods generally don't perform well in case of uneven illumination and noise. Resultantly it can be inferred that will not perform well on low quality document images.

Local thresholding methods, on the other hand, compute a threshold for each pixel (or group of pixels) in the image on the basis of the content in its neighborhood. As opposed to global thresholding, local methods generally perform better for low quality images, specially classifying pixels near text and object boundaries as either foreground or background. Different binarization methods have been evaluated in [74] and [72] for different types of documents. In[74], authors have presented a multi-step approach with focus on document images using initially a low pass Wiener filter for preprocessing for obtaining image referenced as I. In the second step, Sauvola's algorithm [75] to extract the initial binary image S. Third step is estimation of background to get image B. Final thresh-holding is undertaken by incorporating image I and B. A pixel is taken as text pixel if corresponding distance between I and B exceeds a predefined threshold d. In[76], authors present an evaluation of eleven locally adaptive binarization methods for gray scale images with low contrast, variable background intensity and noise. In that evaluation, Niblack's method [77] was found to be the best of them all. Different improvements have since been made to the original Niblack's method to improve the results. These include Sauvola's algorithm[75], Wolf's work [78] and Feng's method[79].

Effort was made using fixed global thresholding on our document images but the results were not satisfactory. On the contrary, once it was done on some local sliding-window based thresholding methods on our images, results were immediately better. A customized binarization algorithm 'NICK' by improving the original Niblack's formula is developed in [80] . The results achieved using NICK are better than other Niblack inspired methods as later shown in the results. An account of some of the well-known local sliding-window binarization methods which were tested and comparisons drawn are provided in Result Chapter (Chapter 5).

4.2 **Segmentation**

Segmentation is a critical phase of information retrieval process. In this Section discussion on segmentation process will be undertaken in detail including its various problems particularly in relation to handwritten text and Urdu language fonts. Processing / matching is directly proportional to the accuracy of segmentation. Snow ball effect is seen if the errors in segmentation are not done dealt and have pronounced outcome in later phases like extraction of features to obtain feature vector, indexing, matching, retrieval etc. The main purpose of segmentation is to accurately extract a word from scanned image. All out efforts are made to eliminate / reduce errors i.e. variable writing classes, sloping lines, punctuations, commas, overlapped alphabets / characters, erased words etc. Efforts have been made to Segmentise handwritten documents using innovative techniques to make the foundation as accurate as possible [81].

4.2.1 **Problems in Segmentation- Handwritten English Language**

4.2.1.1 **Style of Writing**

Identical script written by changed authors produces totally diverse outcomes. Figure 4-8 provides evidence of this hypothesis

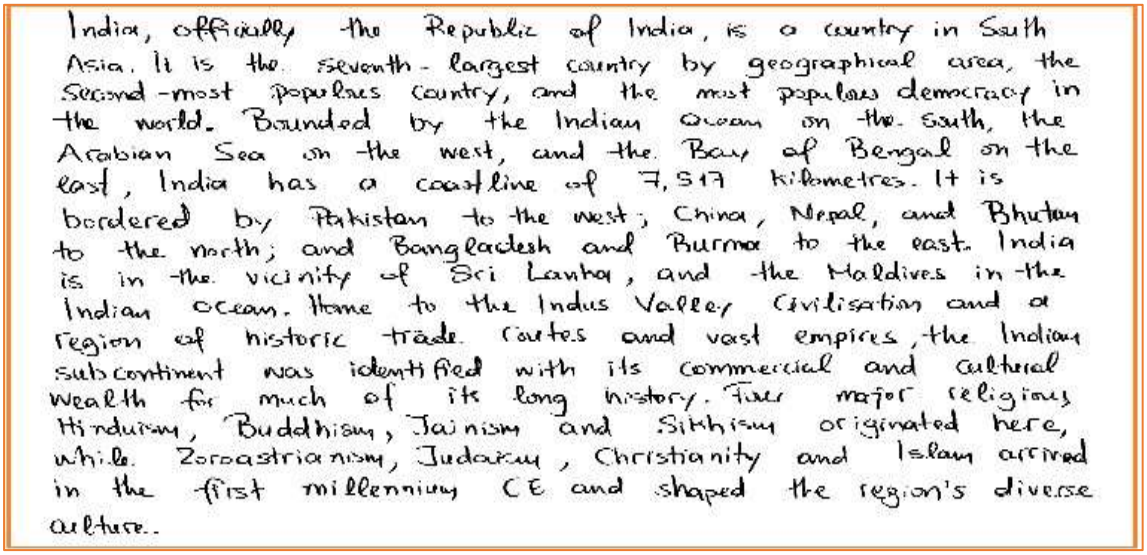

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the South, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture..

Figure 4 - Handwritten Text by 1st Author

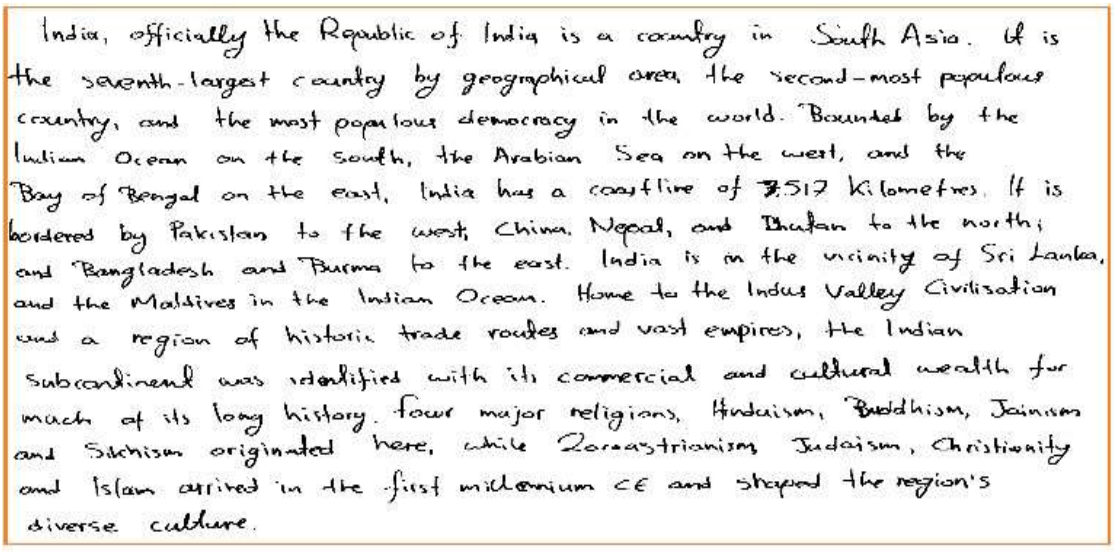

India, officially the Republic of India is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

Figure 5 - Handwritten Text by 2nd Author

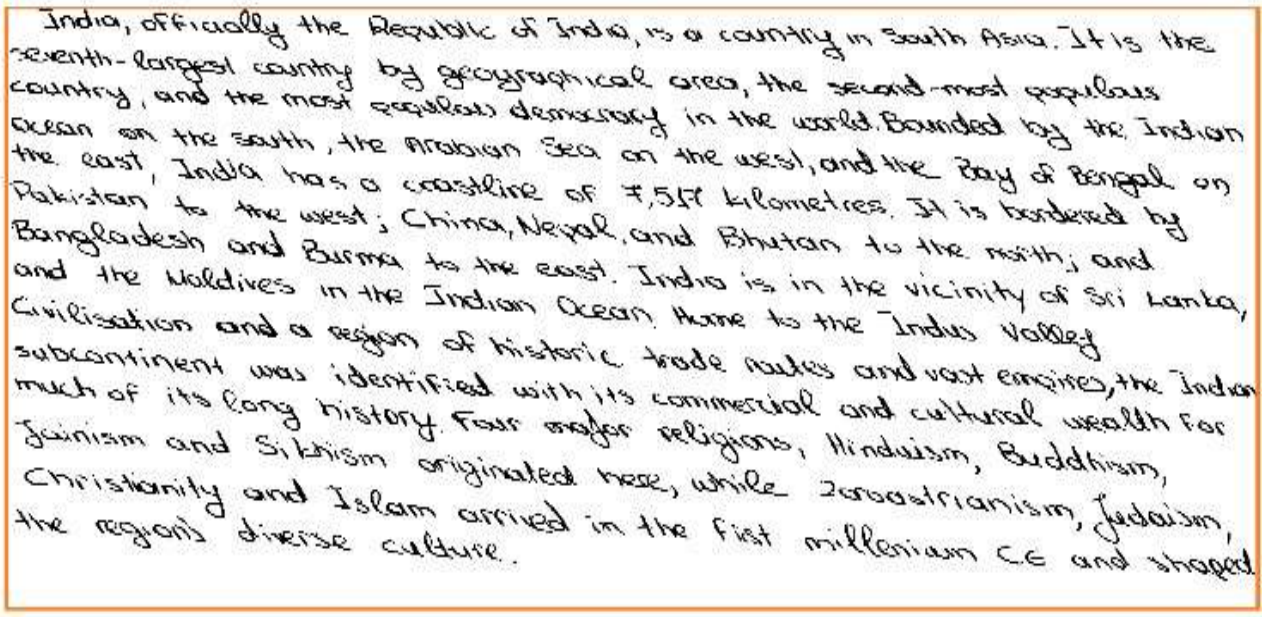

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

Figure 6 - Handwritten Text by 3rd Author

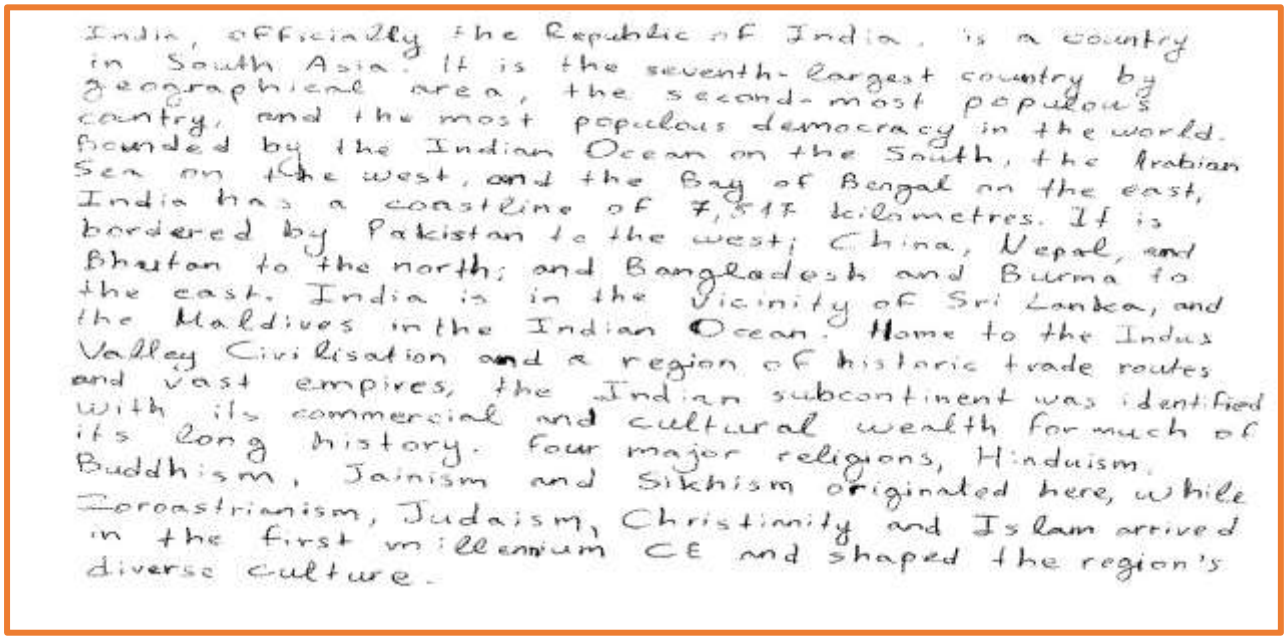

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the South, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

Figure 7 - Handwritten Text by 4th Author

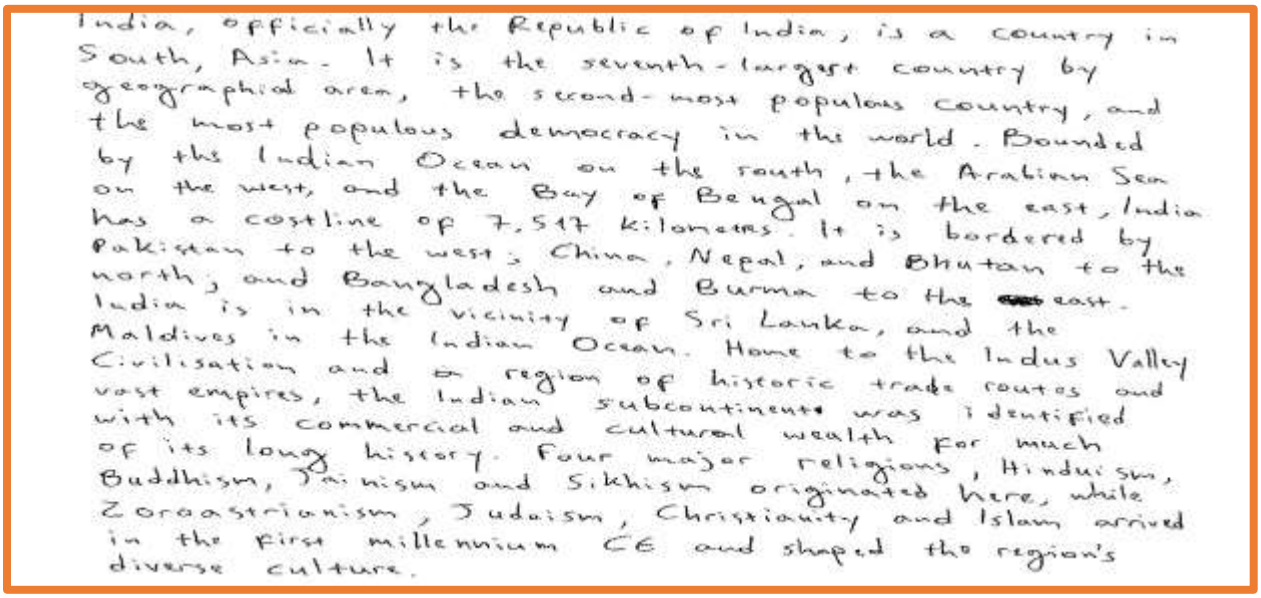

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

Figure 8 - Handwritten Text by 5th Author

4.2.1.2 **Sloping / Slanting Lines**

Hand written scripts may comprise sloping / inclined lines that generates challenging tasks like overlapped scripts / characters between two adjacent lines making connected component analysis extremely tough as shown in Figure 9.

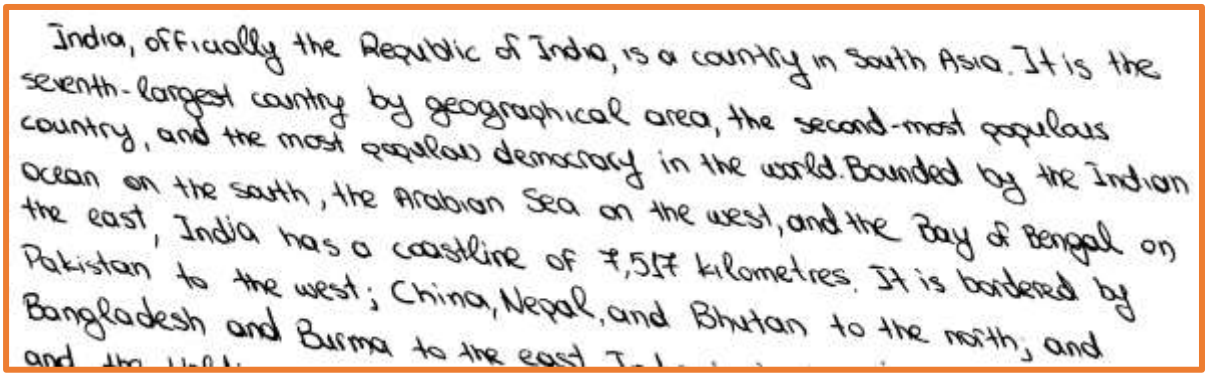

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east

Figure 9 - Slanting Lines have their own challenges

4.2.1.3 **Unevenness / Irregularity in Inter Character / Inter Word Distances**

In case of typed textual scripts, inter word / character distance is maintained in a constant manner. However, in case of hand written scripts, it is totally dependent on human behaviors, environment and discretion. This particular aspect generates challenges by not affording researchers to have a common thresh-holding for variability in diverse blueprints. Effect of Commas, punctuations becomes more noticeable as shown in Figure 10.

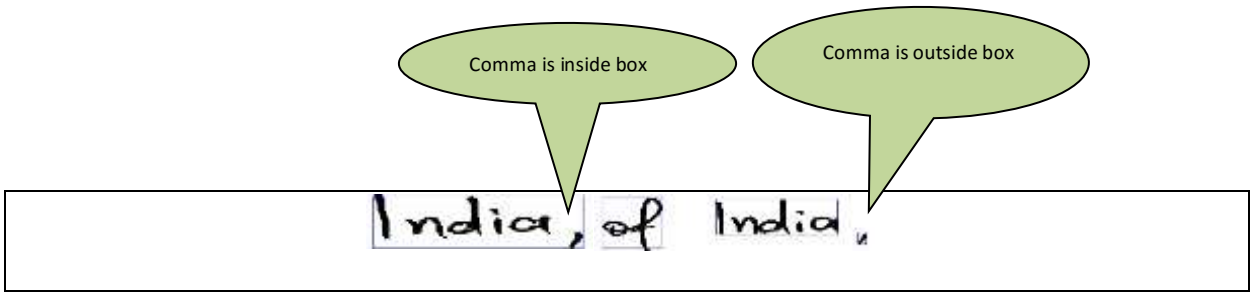


Figure 10 - Challenges faced due to Irregularity in spaces

4.2.1.4 **Miscellaneous Errors**

In addition to above mentioned errors, handwritten texts have multiplicity of faults like interspacing, Erasing / cutting, over-writing etc as shown in Figure 11.

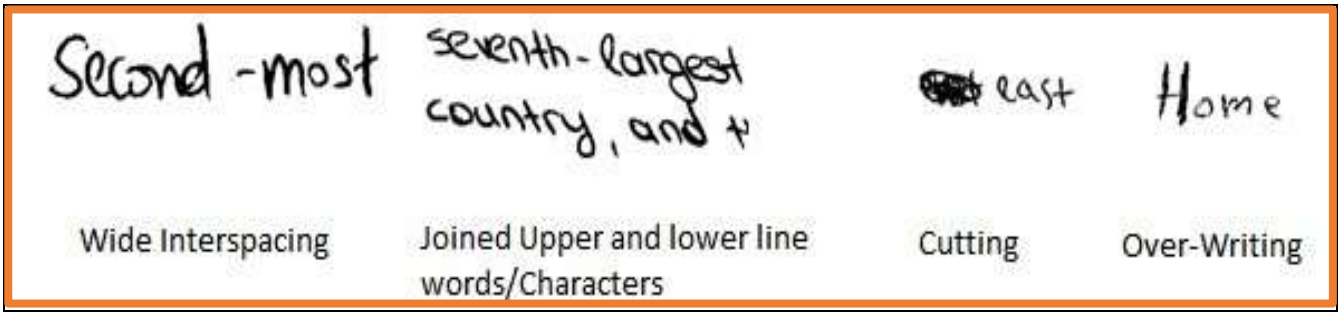


Figure 11 - Punctuations, cutting, Over-writing

4.2.2 **Problems in Segmentation in Urdu Script**

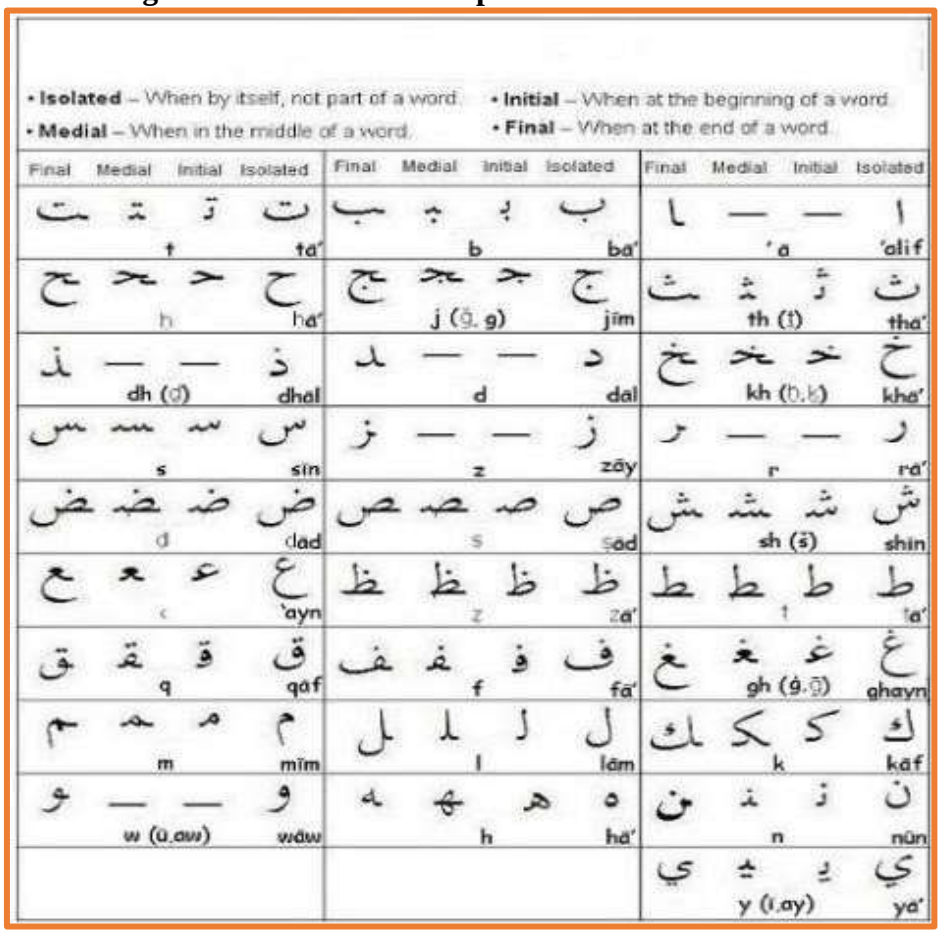


Figure 12 - Different shapes of Characters [67]

The word spotting techniques outlined above have mostly been applied to word or character level segmentation on text based on the Latin alphabet. Proposed method of treating words as images on Urdu printed text also has also been undertaken. The highly cursive nature of Urdu text, however, makes its segmentation into words or characters very challenging. Besides that, Urdu words are written in joining and each letter can take different shapes based on its location in a word, as shown in Figure 12. Irregular inter and intra word / character spacing is also a continuous feature / problem area as depicted in Figure 13.

For Urdu text, segmentation is generally carried out at partial word (ligature) level [68]. The recognition and spotting of Urdu words, however, remains a less explored area. Among significant works on Urdu text, a word recognition system based on SVM is presented in [69] where authors assume that a word comprises a maximum of 4 components. A sliding window of four connected components (CCs) is then used to produce nominee words. The method reported a recognition rate of around 70% on the CENPARMI Urdu database. In previous work [68], a word spotting based indexing and retrieval technique is proposed where a word is segmented into ligatures (partial words) and features are extracted from each ligature. Ligatures from query word are matched with ligatures in database and prediction is performed. However, this system is computationally complex since each and every ligature of all the documents to be indexed is stored in the database. During retrieval, a query ligature has to be checked for similarity against all the ligatures in the reference based and later ligatures are merged into complete words. Extending the ideas presented in [68] by introducing a richer set of features and a clustering step to group similar partial words into clusters has been done. This allows for rapid search as the query ligature is to be matched only against the clusters rather than each and every ligature.

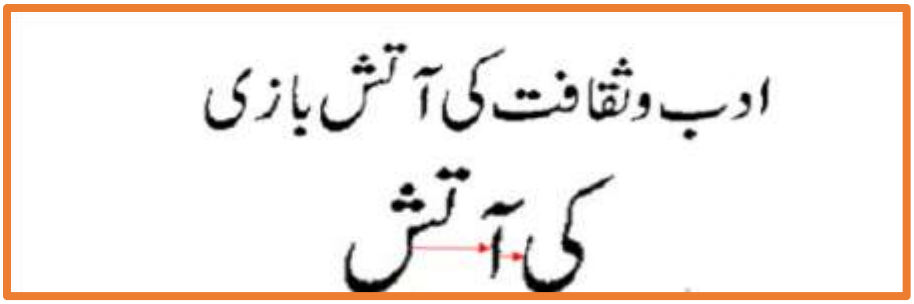


Figure 13 - Non-uniform intra and inter word distance making word segmentation challenging task

4.3 **Problem Resolution Techniques for Segmentation of Words in English Hand Written Text**

Handwritten documents are comparatively challenging to segmented-out in words as compared to typed textual formats due to complications of overlapping, cutting, over-writing of characters and words, varying styles of punctuations marks and variations in writing styles have pronounced effects causing irregular and asymmetric effects and hence, segmentation of words from handwritten documents need to be solved first as a pre-processing step.

These above mentioned problems have been addressed in segmentation phase. Punctuations are removed from the text by applying morphological operations, inter word gaps are adjusted by calculating pixel densities while overlapped characters are detected and then sub-divided by cutting them form point of overlapping, while discarding areas having larger pixel values at one instance. Segmentation is performed using connected component analysis and RLSA is employed.

Figure 14 shows an overview of word segmentation process from documents.

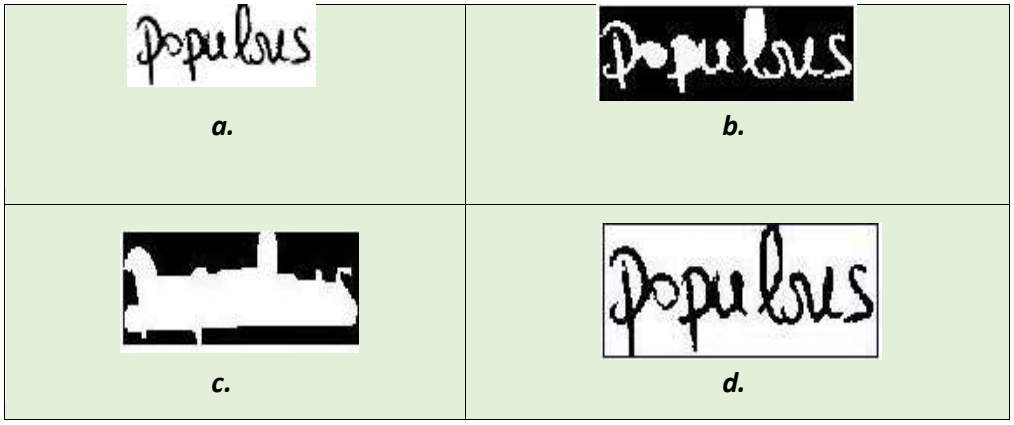


Figure 14 - Overview of Segmentation Process: (a) Original Image (b) Filling of holes (c) RLSA (d) Segmented word

4.4 **Description of Algorithms**

- Firstly image inversion is done and subsequently open spaces likes holes are filled
- Secondly script lines are processed individually. This is undertaken by calculating total number of pixels horizontally and vertically. Pixels are also summed up in gaps that exist between two consecutive lines.
- Threshold is applied and pixels are equated to zero. This creates separation between lines. Thereafter Run Length Smoothing Algorithm (RLSA) is used to dilate and erode the image to obtain connected component. This also gives us aspect ratios which are vital for further processing.
- Final step is conversion of image back to its original shape. In order to extract the word image a rectangular box is used to demarcate the intended image i.e. segmented word. Marking of boundary is done in X and Y directions.

4.5 **Techniques Used for Improvement in Pre-Processing Stage**

4.5.1 **Detecting Overlapping Components between Two Adjacent Rows**

Printed documents are presented using state of the art mechanisms in which gaps, inter / intra charcter spacing is ensured using software techniques. However, writers use their own styles while writing on papers. This causes problems like overlapped words as evident from Figure 15.

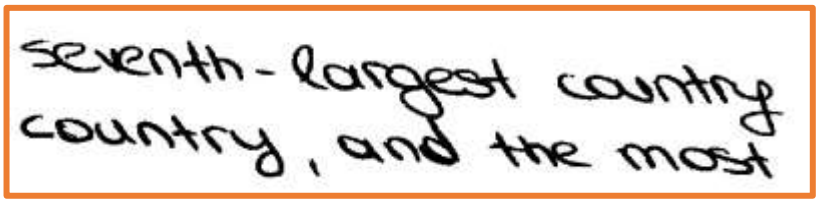


Figure 15 – Overlapping Characters

Above mentioned issue is resolved as under:-

- Firstly for each X Direction (row) height is calculated and average is taken
- Secondly, inter row gap is obtained
- xx = Calculation obtained from 1st step + 2nd step and threshold is taken as 10
- Decision making is done as to, if any character lies between the calculation obtained in 'xx'; that individual character (s) is presumed to be overlapping.

4.5.1.1 **Solution of Resolving Overlapping Characters**

- Phase-1 – Calculate the height of overlapped segments.
- Phase-2 - Inter Line (s) Gap is calculated.
- Phase-3 – Addition is done using calculations obtained in Phase-1 &2.
- Phase-4 – Segmented / Character in focus is divided from middle.

4.5.2 **Segment Filtration**

- Initially information that is not relevant to the word. This is done using thresholding function in which pixels that are below a certain threshold is removed.
- Secondly, in order to remove commas, punctuation marks; it is customary that they exist mostly on extreme right bottom corner of word. Using this fact, algorithm is designed to work dynamically. 4 x Zones are made for each extracted word and if there exists any small fraction of pixels relevant / compared to the actual word, it is assumed to be some irrelevant information.. Once the commas are removed, original image is replaced with the processed image. Algorithm works as under:-
- Step-1. To preserve the original image, a copy is made.
- Step-2. Normalisation of segments is undertaken by calculating the average size of script and dividing it by total segments.
- Step-3. MATLAB function 'Bwareaopen' is applied on copied image.

- Step-4. Original Image is cropped from right bottom corner.
- Step-5. Original image is repced with processed image i.e. without commas, punctuation marks.

4.5.3 **Moving Bounding Boxes**

- Rectangular Boxes obtained as a result of segmentation are critical for any further processing and their accuracy is a must factor. In case some extra information is available with bounding boxes it is dealt as under:-

- Firstly; image is parsed from right to left using keeping center in ficus. Attempt is to find first white pixel.
- In second step; replacement of bounding box is done. New minus old (containing extra information).
- For all subsequent steps, loop back is applied that checks each bounding box and keep replacing the bounding box if any extra information is retrieved.
- Diagrammatic representation is depicted in Figure 16.

Figure 16 - Results of techniques employed in Pre-Processing Stages

4.6 **Segmentation of Words from a Sample Document Image**

After the application of various techniques, one sample image after segmentation is shown in Figure 17.

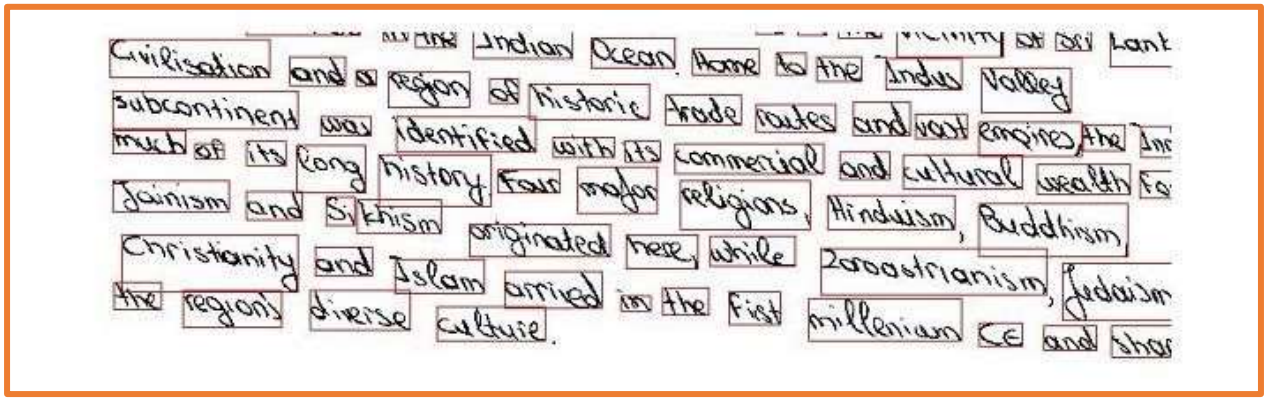


Figure 17 - Sample document image with segmented words

4.7 **Segmentation in Urdu Scripts**

The segmentation techniques outlined above have mostly been applied to word or character level segmentation on text based on the Latin alphabet. The highly cursive nature of Urdu text, however, makes its segmentation into words or characters very challenging. For Urdu text, segmentation is generally carried out at partial word (ligature). Figure 18 shows a sample image and Figure 19 shows extraction of partial words.

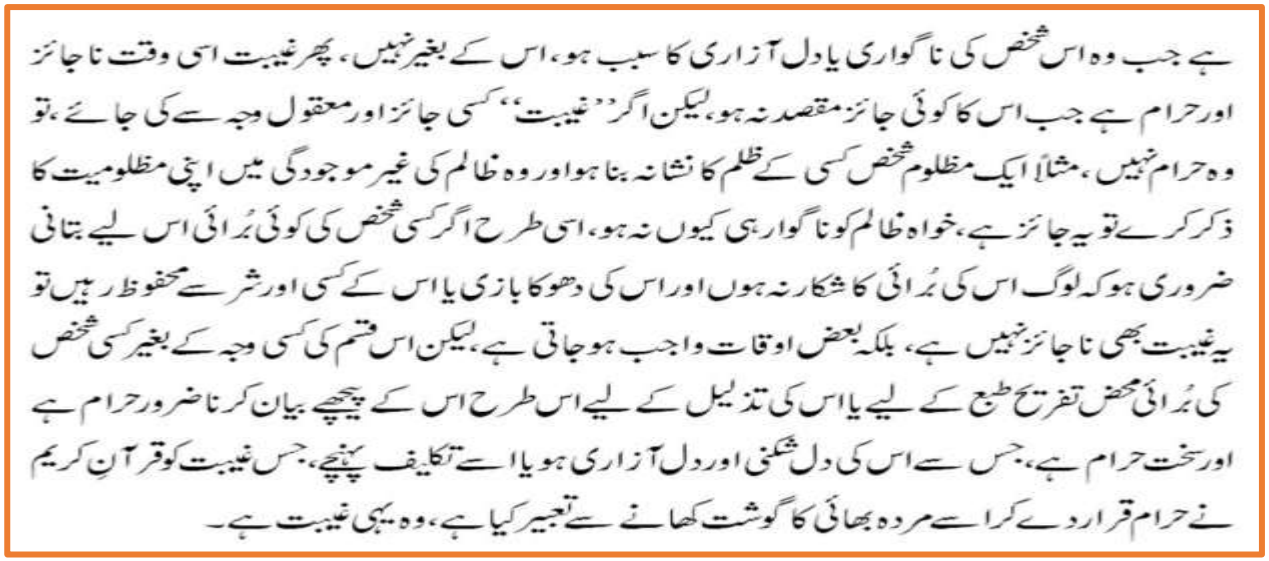


Figure 18 – One Sample Urdu Image Document

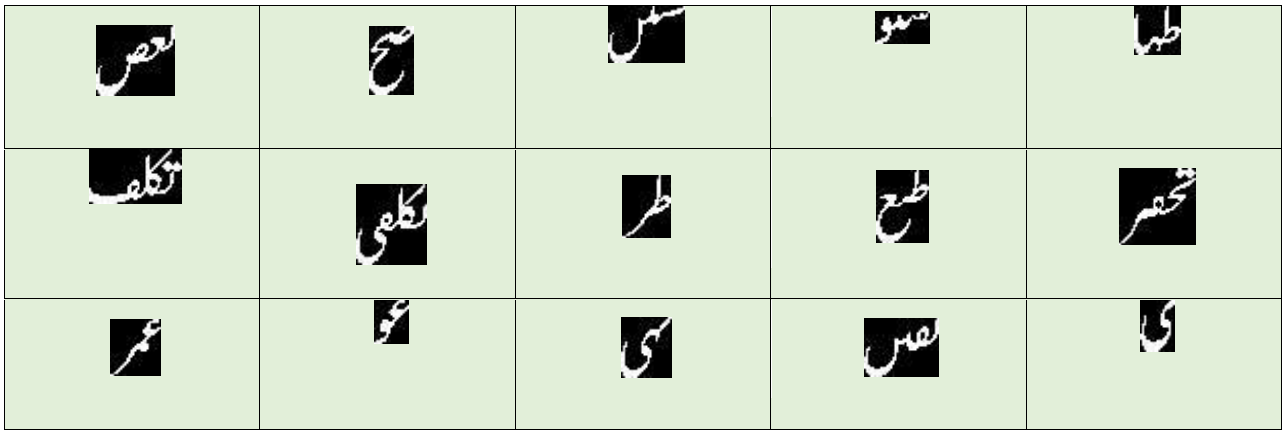


Figure 19 – Examples of partial words ligatures extracted as a result of Segmentation

4.8 **Results of Scanned Images using Abbyy Fine Reader Software**

Comparison of results on handwritten documents obtained through well-known OCR system has also been done and it was found that efficiency is reduced while employing them on handwritten documents due to problems already highlighted in segmentation chapter. Results obtained through OCR using abbyfine reader are presented in Figure 20.

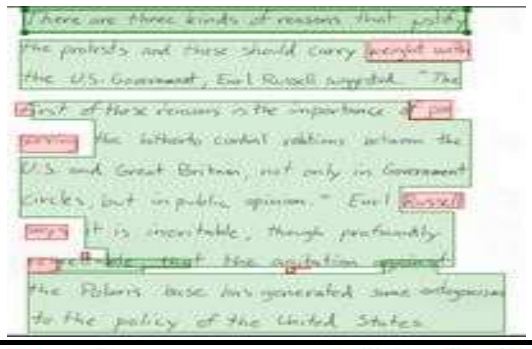
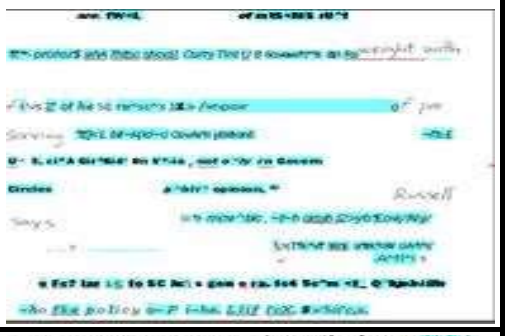
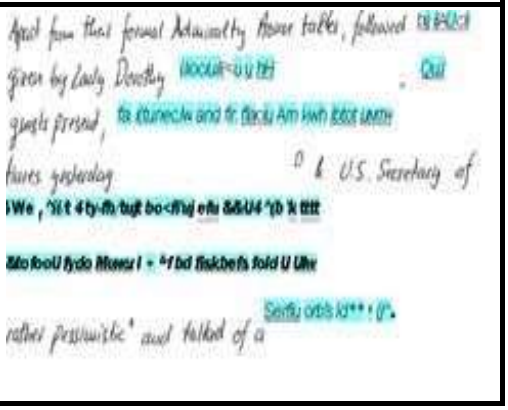
Image	Result Using OCR
	
	

Figure 20 – Results obtained through OCR on handwritten documents

4.9 **Summary**

Segmentation in handwritten and cursive scripts poses a number of challenges and needs to be dealt meticulously as errors at this stage will get pronounced in later stages of information retrieval system. Major issues with RLSA based word segmentation include merging of words across different lines and fusion of punctuation marks with the word. Overlapped character, Commas , punctuation marks are few examples. The problem of overlapped characters is addressed by finding the average row height and average row gap in the document to define a threshold. Segmented words with height greater than the threshold are likely to be the result of a false segmentation and are split into two. The split point is determined by starting at the center of the segmented bounding box and finding the row with minimum number of text pixels.

The removal of punctuations marks (commas, periods etc.) is traditionally carried out by applying a threshold on the area of connected components and filtering the small components. Our proposed technique extends the same idea but to avoid removal of dots associated with characters (‘i’ and ‘j’), applying area based thresholding only on the extreme right components in the lower half of the word bounding box has been done. This allows removal of punctuation marks and the resulting words are extracted from the document for further processing. Use of RLSA and connected component together with dynamic calculation of slant lines thus removing un-necessary punctuations, overlapped characters, inter and intra word spaces, formulation of ligatures have yielded significant results[81].

Chapter -5 Document Indexing (Feature Extraction and Clustering)

Feature extraction is considered the most critical step in a pattern classification problem that aims to find a characteristic representation of patterns under study (words in our case). As discussed earlier, both structural and statistical features have been investigated in classification problems. In our work done, employment of a set of statistical features for which rich classifiers are available has been undertaken. In order to cater the varying writing styles of different writers, our proposed technique focuses on conversion of words into shapes, eliminating the unnecessary intra-word white spaces and smoothing the word boundaries. For each extracted word, the empty columns between different components are eliminated and the baseline (row with maximum number of text pixels) of the word is determined. All the components in a word are then joined together by a horizontal line passing through the baseline. Loops and gaps in characters are filled by applying the morphological region filling algorithm to the word image and finally the contour of the word is extracted hence representing each word as a unique shape. Figure 21 illustrates an example of a word extracted from a document and converted into shape.

Once the words are converted to a binary shape represented by its contour, next step is feature extraction. Set of features capturing the shape information like local orientations, curvatures, geometry and other shape descriptors have been employed. These features are detailed in the following paragraphs. Combinations of existing and new features have pronounced our research and have set new avenues for the researchers around the globe. By converting words into shapes advantages of two best existing techniques i.e. shape matching and word spotting have been accrued. This integration has yielded enormous results.

Original	

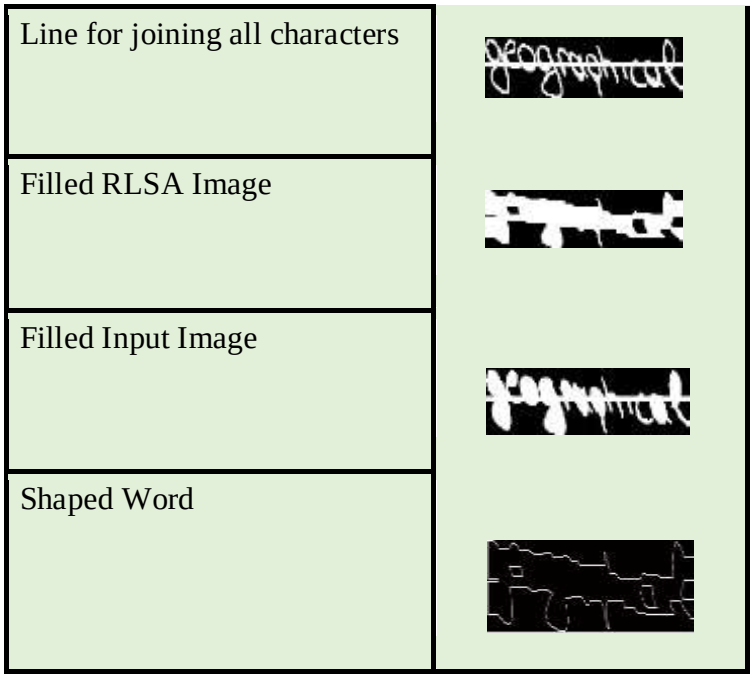


Figure 21 - Overview of extracting shaped word in English Script

5.1 **Feature Extraction from Words**

5.1.1 **Chaincode Features**

Chaincodes have been applied to a number of shape recognition problems with varying degrees of success. In case of document recognition, chaincodes have been applied to problems like writer identification and verification[82, 83], character and word recognition[84-86], classification of writing styles [87] and handwriting based characterization of writer demographics[88].

In our case, exploiting the chaincode representation of the contour of the word shape to extract local orientation and curvature information. The contour is represented as a string of Freeman chaincodes and a histogram of these codes is computed to be used as a feature. The normalized

histograms of chaincodes represent the distribution of local orientations whereas the dominant orientations are reflected as peaks in the histogram.

To capture the local curvature information from the contour, computing a 2D histogram of chaincode pairs where the entry (i,j) of the histogram represents the probability of finding the pair (i,j) in the contour representation of a word. Extending the same idea further, a 3D histogram of chaincode triplets is also computed and is normalized to be used as a feature. Figure 22 illustrates an example word and the corresponding histogram of chaincodes as well as the 2D histogram of chaincode pairs.

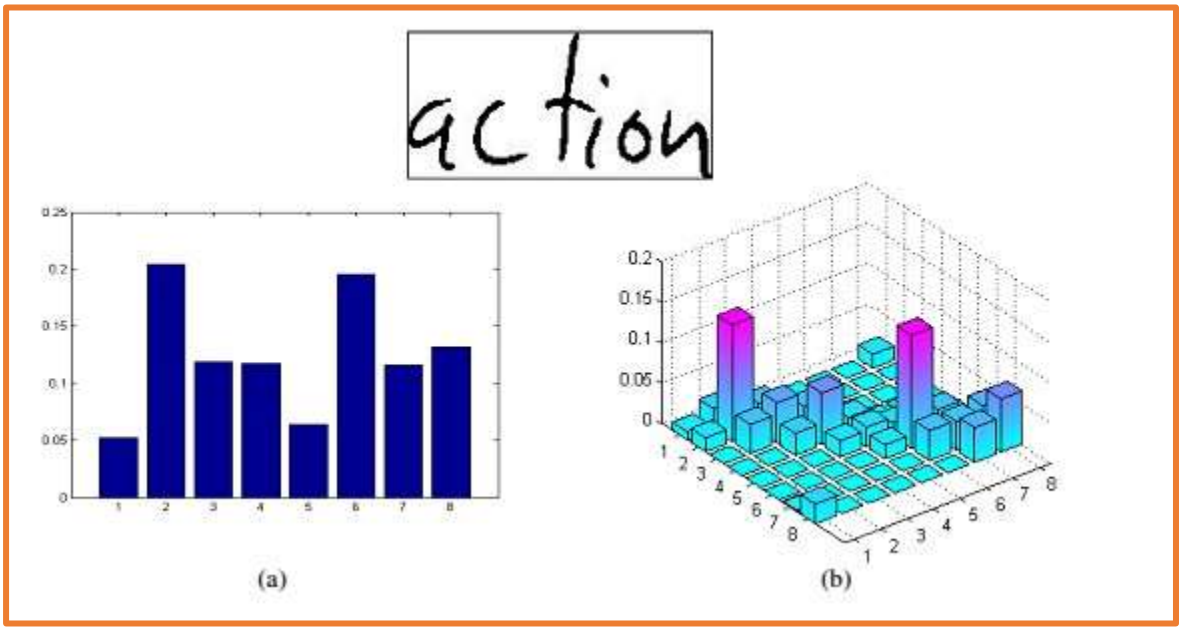


Figure 22 - Example word and (a): Histogram of chaincodes (b): Histogram of chaincode pairs

5.1.2 **Polygon Features**

In addition to the chaincode representation, the orientation and curvature information of a word is also captured by approximating the word shape by a polygon. Polygonization of word contours not only represents a distant scale of observation but the computed features are also more robust to distortions as compared to the chaincode based features. The sequential polygonization

algorithm presented in [89] is applied to the contour of a word to represent it by a set of line segments. The slope of each segment and the curvature (angle) between each pair of neighboring segments are computed and their distributions are used as features. Each of the two distributions is quantized into eight bins. In addition to these distribution, length weighted distribution of the segment slopes (curvatures) is also calculated where each bin of the histogram is incremented by the length(s) of the segment(s) having a particular slope (angle). These distributions are then normalized to have a sum of 1 and are used as features in characterizing a shaped word as shown in Figure 23.

Figure 23 - An example word and its polygonized contours

5.1.3 **Pixel Density Features**

These features capture the pixel density information in different zones of a word. To compute zone based features (*f*7), a word is divided into four sub-images by placing a 2x2 grid on the image (Figure 24). For each zone, the proportion of pixels with respect to the total number of text pixels in the segmented word is computed.

$$D_i = \frac{N_i}{\sum_{i=1}^{4} N_i} \qquad i = 1,2,3,4$$

With N_i being the number of pixels in zone i.

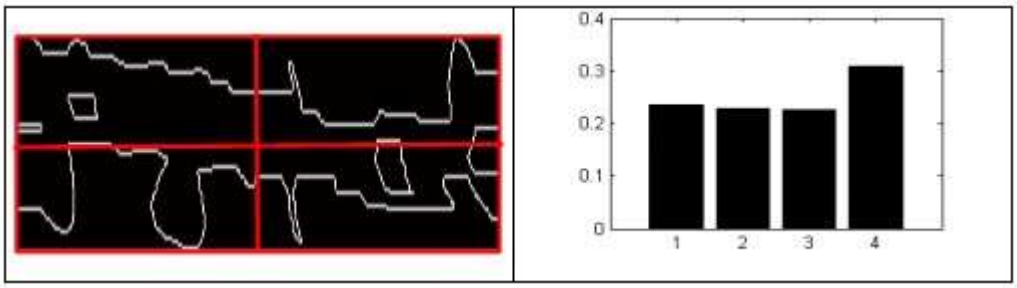


Figure 24 - Shaped word divided into four zones and the pixel density in each zone

5.1.4 **Profile Features**

Profile features have been very effectively applied to a number of recognition problems[90]. Among various profile features, upper and lower profiles have been most widely employed. Upper profile is computed by finding, in each column, the distance of the first text pixel from the top of the bounding box of segmented word. In a similar fashion, the lower profile is calculated by finding the distance of the last text pixel from the top of the bounding box. Both the profiles are normalized by dividing them by the height of the bounding box of the segmented word. The dimension of upper and lower profile is the same as the number of columns (width) of the shaped word. Typically, profiles are matched using the well-known Dynamic Time Warping (DTW) method[44]. In our implementation, however, mean and standard deviation of each profile giving a four dimensional profile feature has been undertaken.

5.1.5 **Projection Features**

Horizontal and vertical projections of a shaped word are computed through summation of total number of text pixel in each row (column) of the shaped word. The projections are then normalized by division with the width (height) of the image. Similar to profile features, we compute the mean and standard deviation of the horizontal and vertical projections and employ them as features.

Similar to the orientation features computed from the polygonzied contours of a word, our technique computes the local orientation information in small zones of the word. The word image is converted into skeleton and is divided into 9 (3 × 3) equal sized windows. Features are then extracted from each zone of the word. These features include the number of horizontal, vertical, left diagonal and right diagonal lines. In addition, our technique also computes the normalized length of all horizontal, vertical, left diagonal and right diagonal lines and the normalized area of the word skeleton.

5.1.6 **Shape Descriptors**

These features capture the shape properties of the word image and include the well-known Hu moments, Euler number, the ratio of total number of ink pixels to the total number of pixels and eccentricity of the shape.

5.1.7 **Delta Features**

These features are aimed at capturing the information on loops, holes and overall structure of the word. Morphological closing is applied on the image to close holes and gaps and the (normalized) difference between the original and the closed word image is computed. This value is relatively high for words with loops and holes and serves to discriminate them from other words. In addition, to capture the general structure of the word, iterative application of morphological dilation on the word image with 4 different structuring elements and the percentage change in area of the word between two successive dilations is used as feature.

$$SE1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$SE2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$SE3 = (1 \quad 1 \quad 1)$$

$$SE4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

The difference in area between two successive iterations of dilation (with one of the structuring elements) on three sample words is illustrated in Figure 25 while the respective percentage area changes for the three words are presented in Figure 26. In all cases, the increase in area is relatively higher for the initial iterations and stabilizes gradually as the number of iterations increase. Depending upon the characters within a word, the change in area of the word is different and can be exploited to characterize the word. In our implementation, 6 iterations of dilation giving a total of 24 features for the four structuring elements have been undertaken.

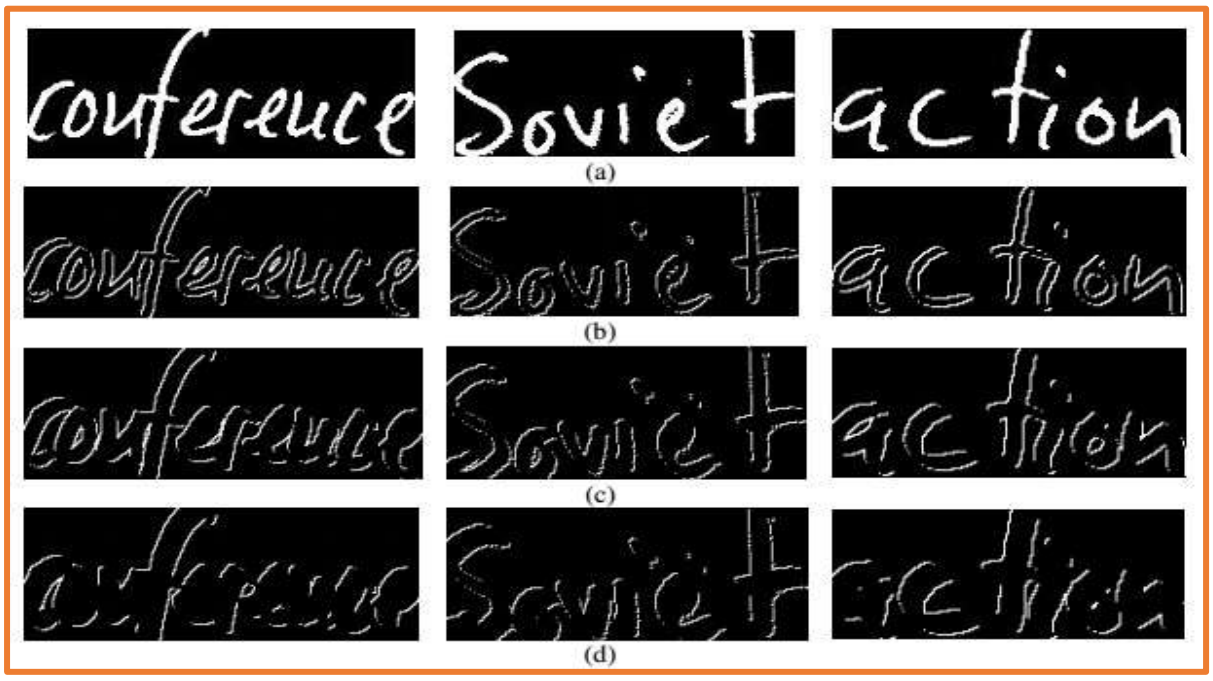


Figure 25 - Change in area of word after dilation with one of the structuring element (a): Original Image (b): Change after iteration 1 (c): Change after iteration 3 (d): Change after iteration 5

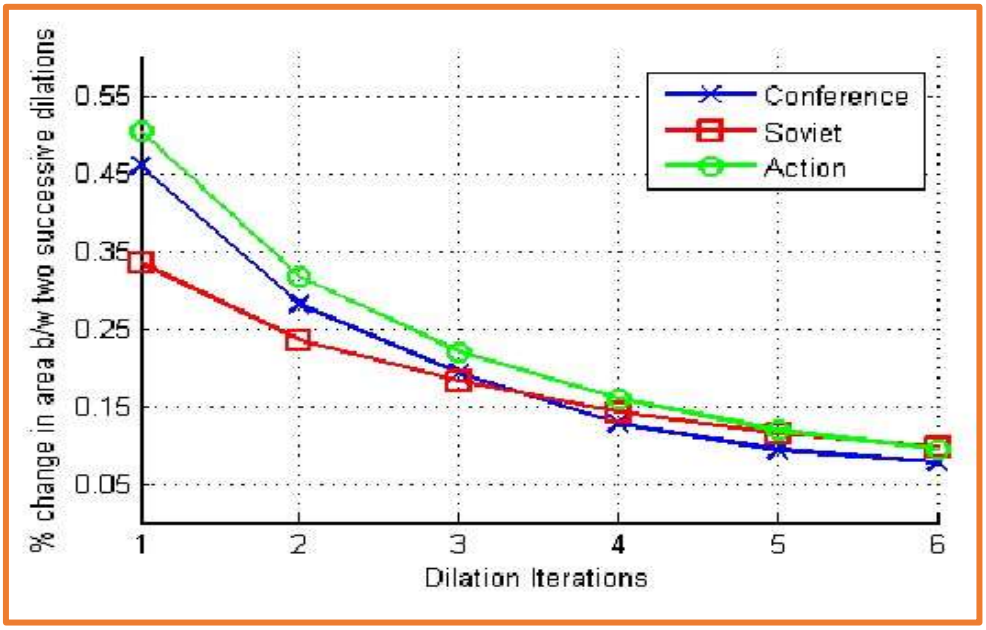


Figure 26 - Percentage change in area of three words as a function of number of dilation iterations

Feature	Description	Computed Form	Dimension
f1	Chaincode based Features	Shaped Word	615
f2	Polygon Features	Shaped Word	42
f3	Projection Features	Shaped Word	4
f4	Profile Features	Original Word	4
F5	Zone based Orientation Features	Skeleton of Word	58
f8	Delta Features	Original Word	9
		Total	732

Table 3 – Summary of Features

5.2 **Feature Extraction in Urdu Script**

Once the ligatures are extracted in Urdu script, our technique computes a set of features to characterize each ligature. Each ligature is treated as a unique shape and a set of shape descriptors is extracted from a ligature for subsequent matching. All the above narrated features have also been applied on Urdu script. These features provide useful statistical information on the ligature under study and allow discriminating different ligature classes. An example of word converting to shape is depicted in Figure 27.

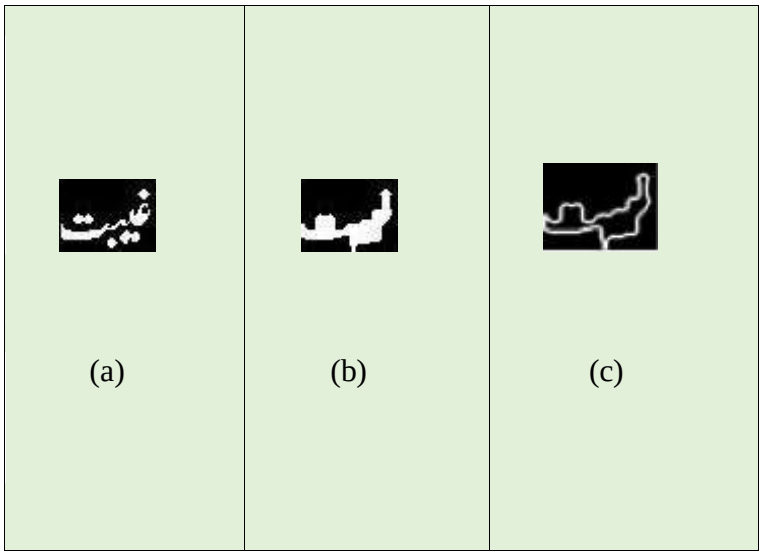


Figure 27 – Overview of extracting word in Urdu Script (a) Original Word (b) Morphological Operations (c) Shaped Word

5.3 **Clustering of Words in English Script**

Prior to indexing of documents, clusters of words need to be generated where each cluster is intended to contain multiple instances of the same word to capture the writer-dependent variations within a word. This clustering is carried out offline, serves the subsequent steps of indexing and retrieval and can be manual, automatic or semi-automatic. In our implementation, semi-automatic

clustering of words where the clusters generated by a sequential clustering algorithm are manually corrected prior to indexing have been used.

To generate word clusters, sample of 20 document images containing more than 1000 words have been used. Each segment image was converted into words and then represented each word by a feature vector as discussed in the previous section. In clustering; our proposed framework has employed sequential clustering algorithm [83, 91]. The main advantage of using this technique is that it does not necessitate / require prior knowledge of number of clusters. In the first step, it randomly picks a word and assumes it to be the center / mean (representative) of first cluster. Subsequently, with every incoming word; this technique computes its distance with the center of each cluster and choose the nearest cluster as a potential candidate. If the distance of the incoming word to the nearest cluster is lower than an empirically determined threshold, the incoming word is assigned to that cluster and the cluster mean is updated. On the contrary; if the distance does not meet the requisite condition, a new cluster is created with the word in question as the mean of the newly generated cluster. This process is repeated until all the words / ligatures have been assigned to a cluster.

The most noteworthy short-coming of this clustering technique is that it is sensitive to the order in which words are accessed by the algorithm. It is worth mentioning here that clustering is an offline job and making of clusters through the use of this technique will give us approximate set of words. These clusters will be manually inspected and corrected prior to indexing. This infers that overall performance of framework is not totally dependent / sensitive on clustering stage. Executing the mentioned clustering algorithm on the sample images, our technique got a total of

136 clusters. These clusters, naturally, contain some errors which are corrected manually.

Similarly, clusters with less than 5 elements are removed. After refinement, total of 88 clusters containing a total of 941 words were obtained. These clusters are then employed to train the

SVM based classifiers which are used to index the given set of documents.

5.4 **Support Vector Machine Training**

Once the clusters of words are generated, the features extracted from these words are used to train a multi-class support vector machine (SVM) to learn to discriminate between different word classes. The SVM is based on one-against-all implementation using the radial basis kernel function while the parameters of SVM (C and gamma) are empirically chosen through cross validation. Features extracted from a word are fed to the trained SVM which assigns a probability of classification to each of the classes (clusters of words). The word is assigned to the cluster for which SVM reports the maximum probability. In case the probability of assignment is less than a pre-defined threshold, the word is assumed to not belong to any of the clusters in the database and is discarded. Example of a cluster with Word 'Agreement' is shown in Figure 28.

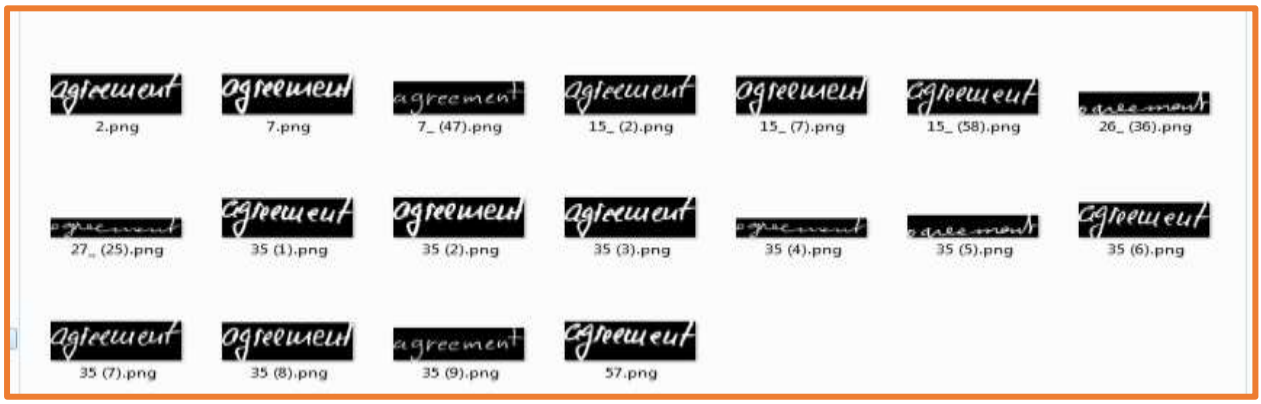


Figure 28 – Cluster of Word 'Agreement' using all possible instances existing in dataset

5.5 **Clustering in Urdu Script**

The main idea of clustering is to group similar patterns (ligatures/partial words in our case) into clusters. In our study, clustering has been done with the objective of reducing the computational complexity of the retrieval phase. Once ligatures in a document are segmented and features are extracted, clustering is carried out to group different instances of the same ligature into groups. The clustering step establishes clusters of similar ligatures which serve as training data for subsequent classification. This allows matching a ligature in question only with the clusters rather than with all the ligatures in the database.

Manually generating these clusters is naturally a monotonous and time consuming task, therefore, employment of a semi-automatic clustering where the ligatures are grouped using a sequential clustering algorithm was undertaken**.** The algorithm starts by randomly picking a ligature and considering it as the center of the first cluster. For each of the subsequent ligatures, the distance with the centroid of each cluster is computed and the ligature is assigned to the nearest cluster if the distance is below a predefined threshold. Otherwise, a new cluster is generated and the ligature under consideration is assumed to be the center of the newly created
cluster.

An important parameter in the aforementioned clustering algorithm is the similarity threshold. A relaxed threshold may result in distinct ligatures to be grouped into same clusters while a tight threshold may result in slightly varying shapes of the same ligature to be distributed into different clusters. The later of the two cases is illustrated in Figure 29 where slightly varying shapes of the same ligatures are attributed to different clusters due to a tight threshold. In our implementation, employing an empirically determined threshold to reduce the clustering errors to as low as possible has been done.

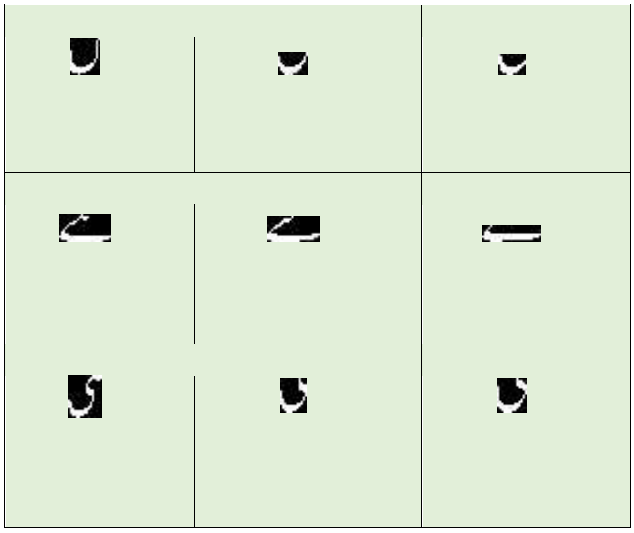


Figure 29 Example ligatures (in rows) which appear same but are assigned to different clusters because of tight threshold

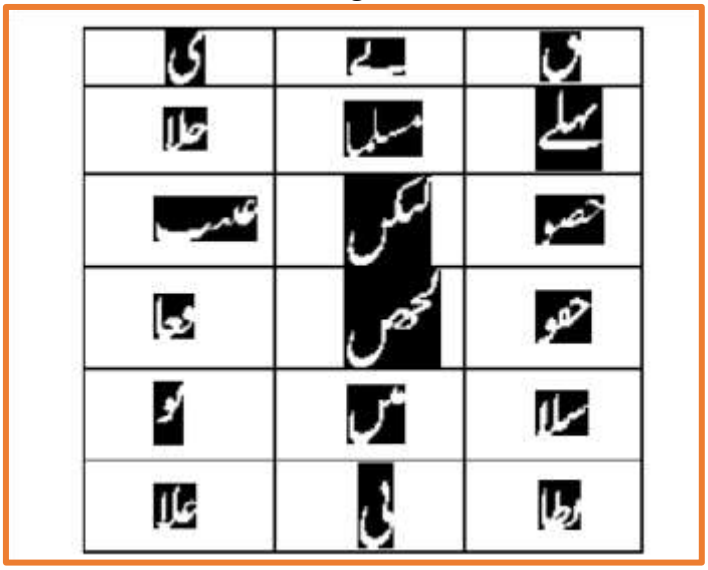


Figure 30- Examples of Ligatures in different Clusters

The sequential clustering process is applied until all the ligatures are assigned to clusters. Once the process completes, the generated clusters are examined to check for any errors. Since the clusters serve as reference base for indexing, the clustering errors are manually corrected so that the training data does not contain any erroneous clusters.

It is worth mentioning that the total number of valid ligatures in Urdu language exceeds twenty thousand but a vast majority of these is rarely used. Most of the frequently employed vocabulary of Urdu can be generated using only few hundred frequently occurring ligatures. Figure 30 shows sample ligatures from different clusters and Figure 31 shows storage of same type of ligatures in a single cluster.

Figure 31 – Cluster of a Ligature (partial word) in Urdu script

Chapter -6 Matching and Retrieval

Once the SVM is trained to learn to discriminate between different word classes, next step is to proceed for matching and retrieval steps. During indexing, the document images presented to the system are segmented into words and are matched against the word clusters and the index file of each cluster is updated. During retrieval, a query word presented to the system is matched with the clusters in the database and the documents containing instances of the query word are presented to the user. Each of these modules is discussed in detail in the following.

6.1 **Training and Generation of Index File**

For indexing, a set of documents is presented to the system is preprocessed, words are segmented and features are extracted from each word as discussed earlier. The features extracted from a word are fed to the trained SVM which outputs the probability scores of the word belonging to each of the clusters (classes). The word is attributed to the class for which the SVM scores the maximum score. It should be noted that our proposed framework has a necessary and sufficient number of word clusters, it is assumed that for each word there is a corresponding cluster available in database. In an eventuality where a word does not find a corresponding cluster, applying a threshold on the confidence score of the SVM reject the words. In case the confidence score of the nearest word cluster is above the threshold, our framework assigns the word to the respective cluster and update the index file of the cluster. The index file contains information on the image ID from database and the coordinates of the word within that image.

The index file is updated every time a new word is added to the cluster. A sample index file of a cluster is illustrated in Figure 32 where the first column represents the file number in database and the rest of the columns represent the information of x-position, y-position, height and width of the respective word within the image.

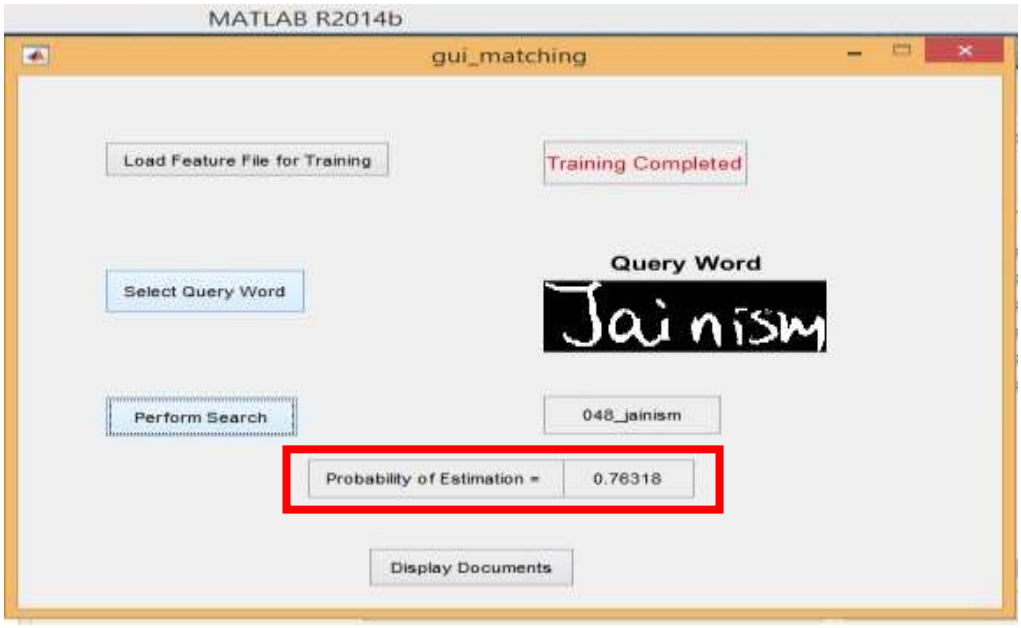
The index of a cluster file allows keeping track of all instances of the respective word within the indexed set of documents.

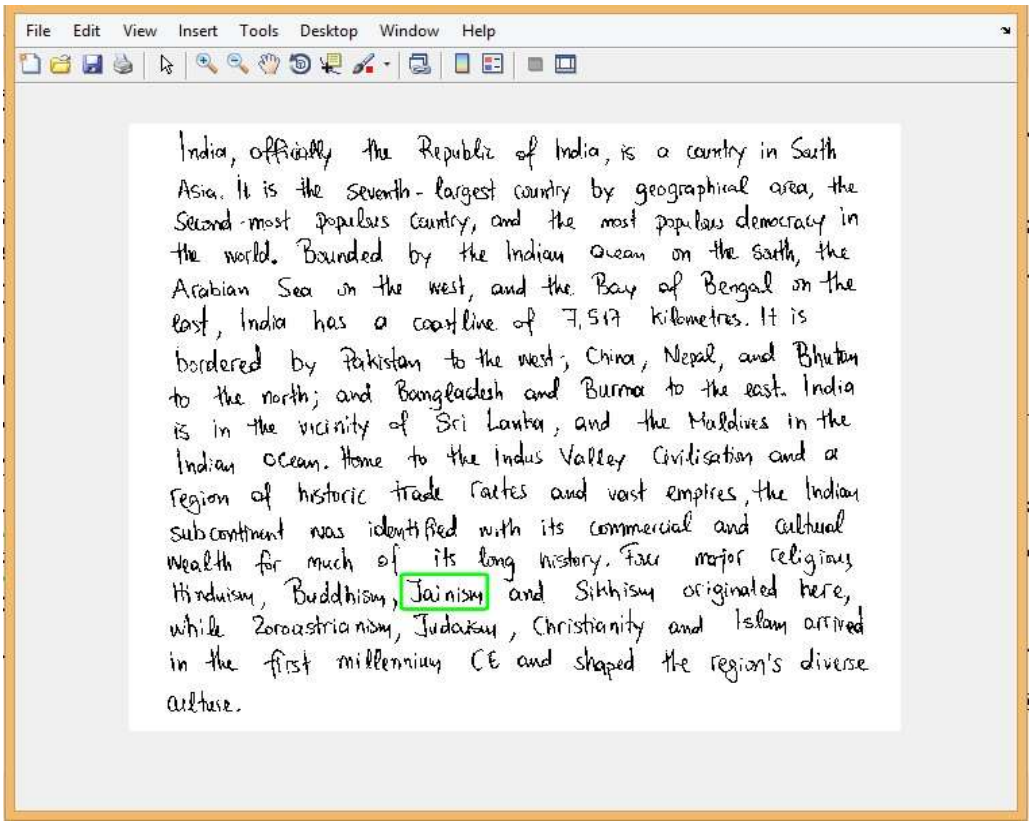
1.png	277	937	90	42
1.png	591	907	63	78
2.png	79	1620	131	75
2.png	1726	742	105	74
3.png	197	368	126	91
3.png	1702	533	72	83
4.png	441	989	134	76
4.png	661	1702	114	76
5.png	44	353	135	68
5.png	512	699	79	94
7.png	639	211	143	85

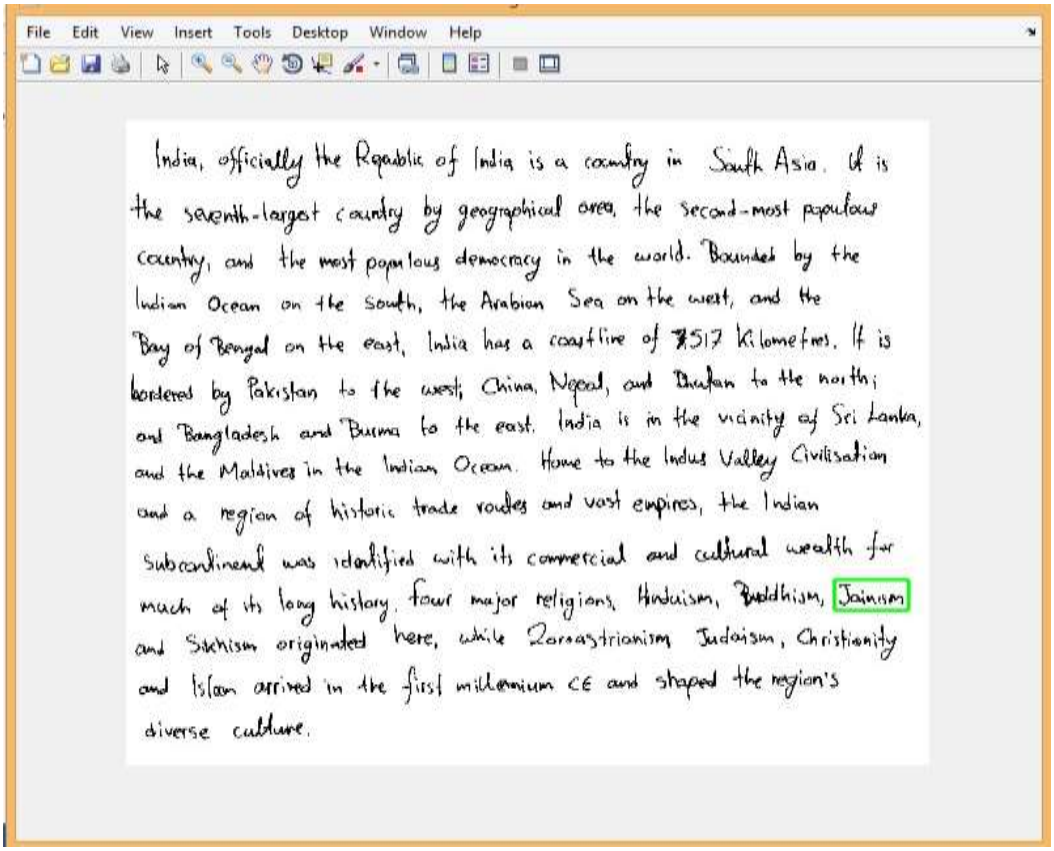
Figure 32 - A sample index file of a cluster showing image number, X,Y coordinates, height and width of bounding box

6.2 **Retrieval in English Handwritten Text**

During retrieval, a query word image is presented to the system and the idea is to retrieve all documents containing occurrences of the provided word. Features are extracted from the query word image and the SVM is used to find the most probable word cluster that matches the query word. Once the cluster is identified, the index file of the respective cluster is parsed to retrieve all documents containing instances of the query word. Each document containing the query word is displayed on screen with the queried word highlighted on the document. Figure 33 illustrates a retrieval session with the system where the query word ‘Britain’ is provided to the system and the documents containing instances of the word are retrieved and presented to the user.


MATLAB R2014b
gui_matching
Load Feature File for Training
Training Completed
Query Word
Select Query Word
Jainism
Perform Search
048_jainism
Probability of Estimation =
0.76318
Display Documents


File Edit View Insert Tools Desktop Window Help
India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.


File Edit View Insert Tools Desktop Window Help
India, officially the Republic of India is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

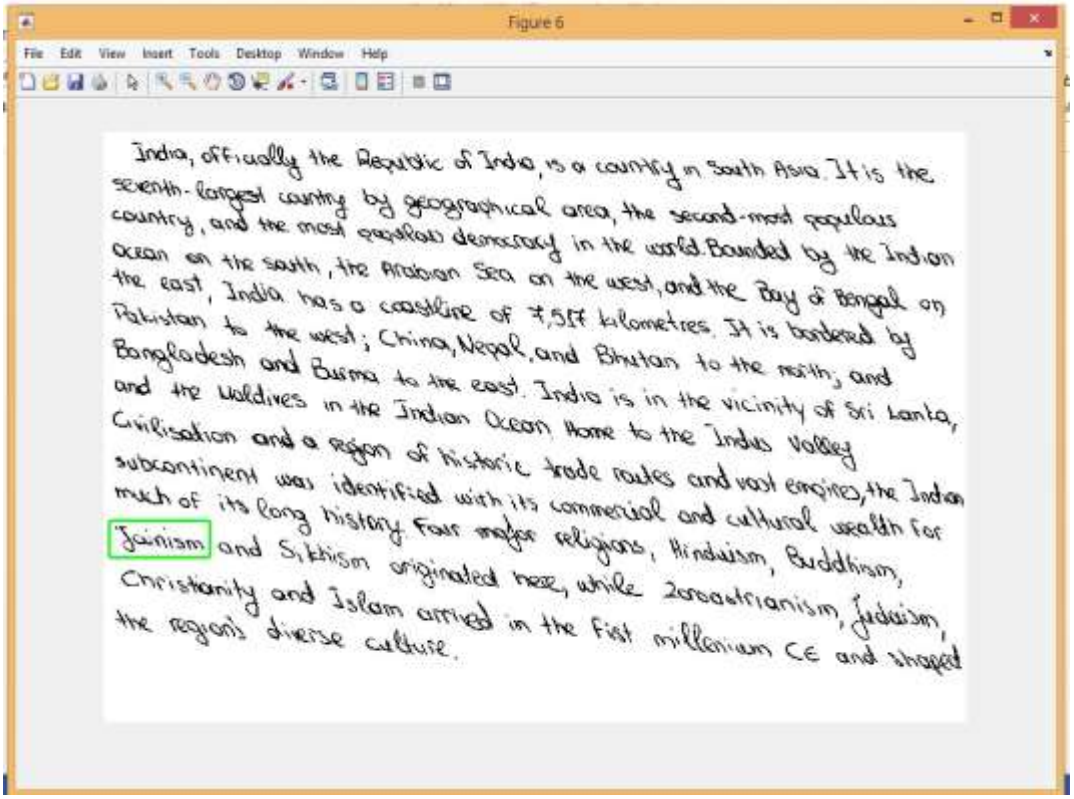


Figure 6
File Edit View Insert Tools Desktop Window Help
India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.


gui_matching
Load Feature File for Training
Training Completed
Query Word
Select Query Word
Asia
Perform Search
007_asia
Probability of Estimation =
1
Display Documents

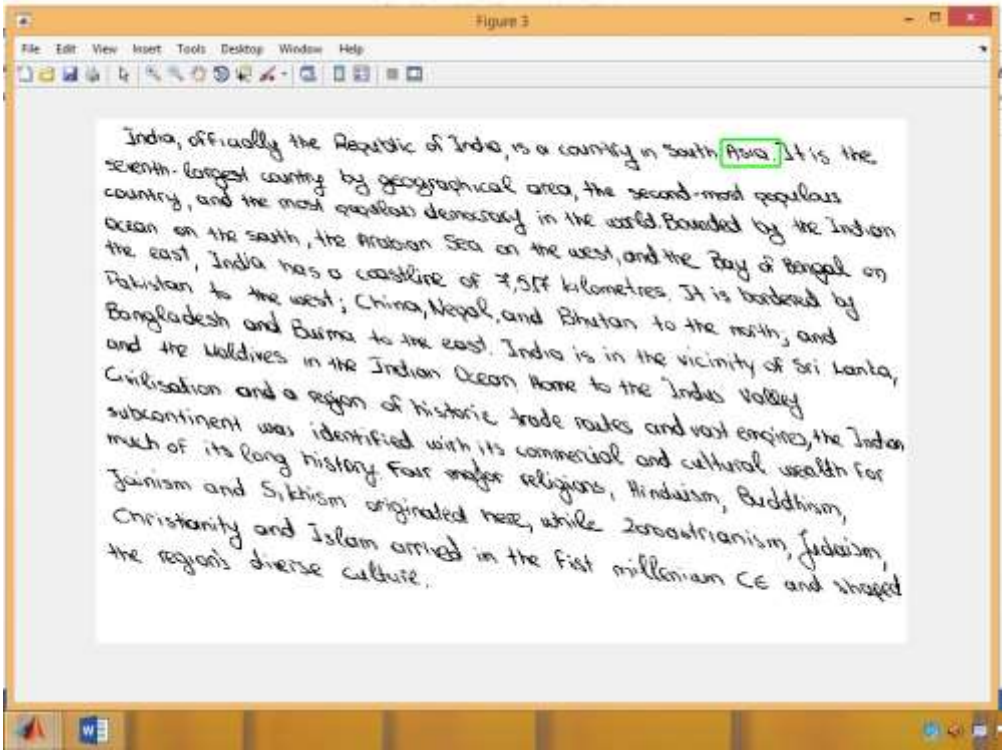
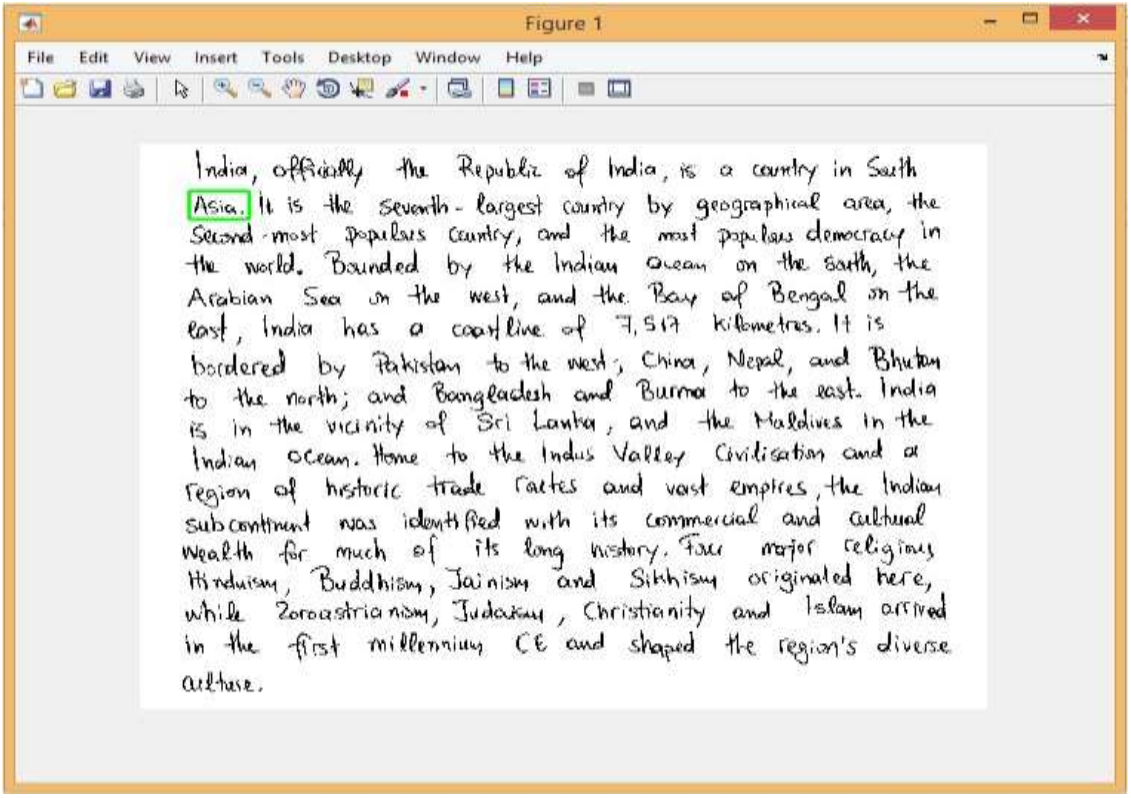


Figure 3
File Edit View Insert Tools Desktop Window Help
India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north, and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the First millennium CE and shaped the region's diverse culture.



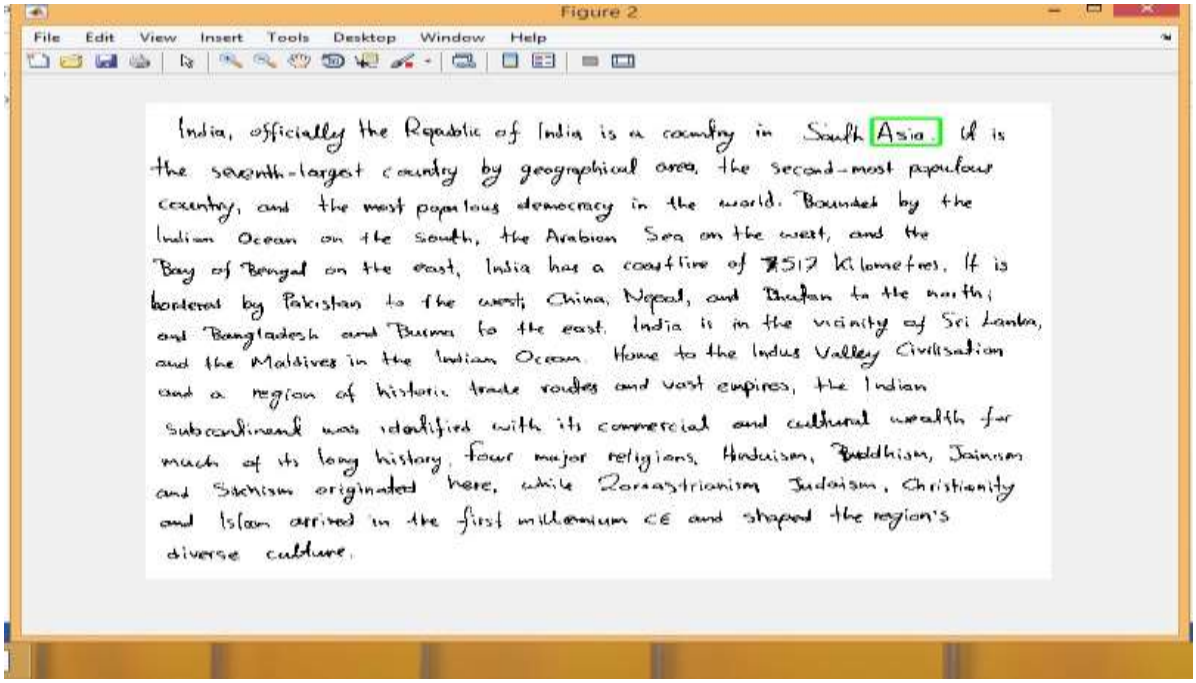


Figure 33 - A retrieval session with the system

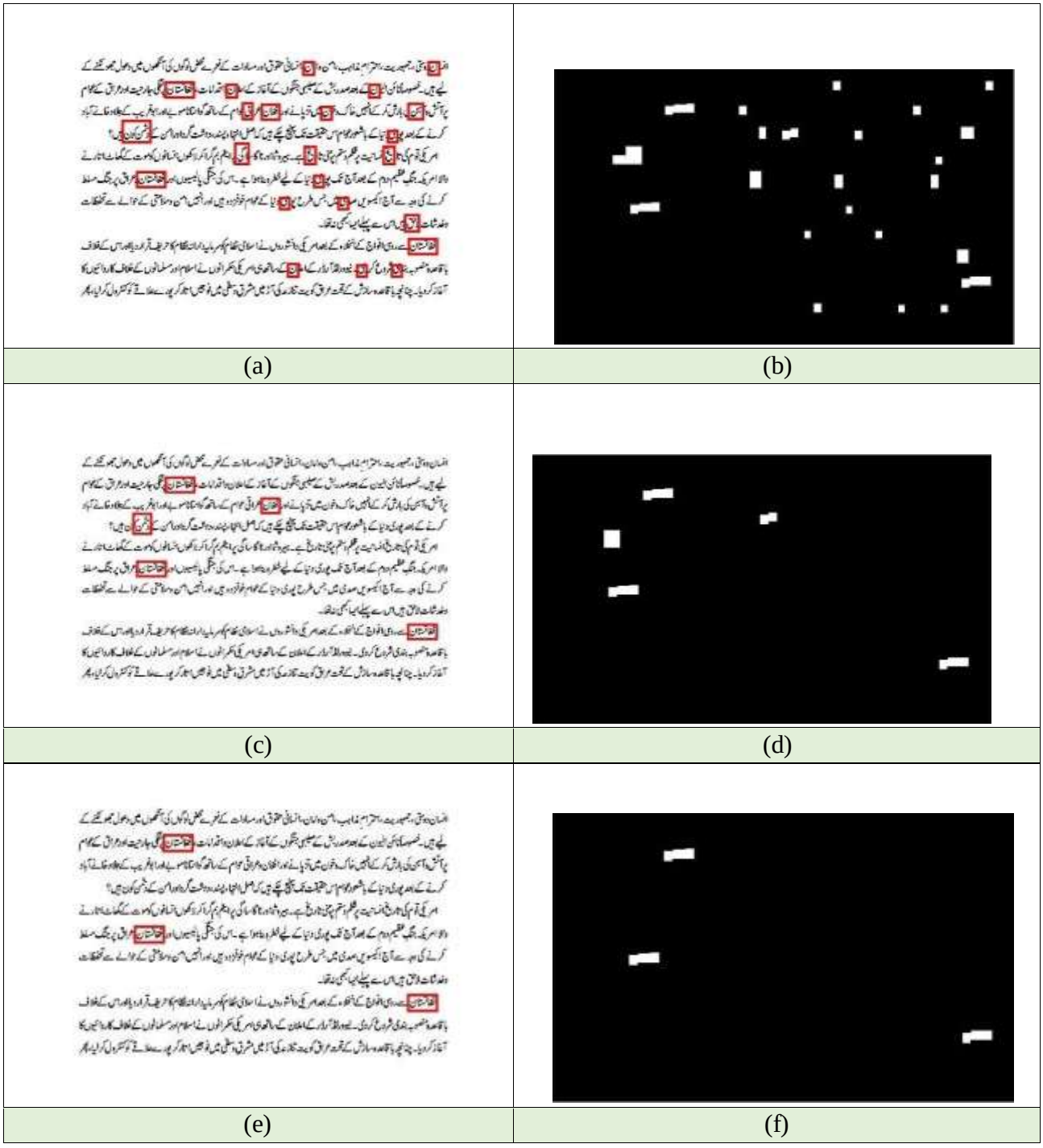
6.3 **Retrieval in Urdu Script**

During the retrieval phase, a query word image is presented to the system. The word is segmented into ligatures and features are extracted from each ligature. A ligature is then compared with the clusters in the reference base and the nearest matching cluster is identified using the nearest neighbor classification. Once the closest cluster is determined, the index file associated with the cluster is parsed to retrieve all the documents containing the occurrences of the query ligature (Figure 34). The process is repeated for all the ligatures in the query word and finally the retrieval results are merged to keep only those documents which contain the complete query word. The retrieval results along with the query words highlighted are presented to the user.

Figure 34 - Ligatures retrieved for a query – many of the ligatures belong to words other than the query word

A major issue while retrieving the documents containing instances of the query word is that the ligatures in the query word may be part of words other than the query word itself. Since the matching is carried out at ligature level, a ligature which is a part of many different words results in a large number of false positives. This problem is illustrated in Figure 35 where a large number of instances are retrieved and many of these do not represent the query word.

Problem of unwanted retrieved ligatures is addressed by merging the retrieval results of all the ligatures in the query word. From the view point of implementation, generation of a binary image where the bounding boxes of the retrieved ligatures represent 1 while the rest of the image is 0 has been done. Morphological closing is then applied so that for words comprising more than one ligature, all the retrieved ligatures of the same words are merged together into a single connected component. Finally, the bounding boxes of the connected components in the closed image are used to find the instances of the spotted words in the indexed documents. The process is summarized in Figure 35 (a – d).



(a) (b)

(c) (d)

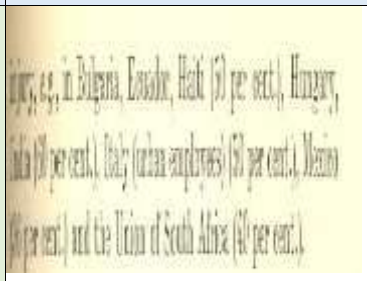
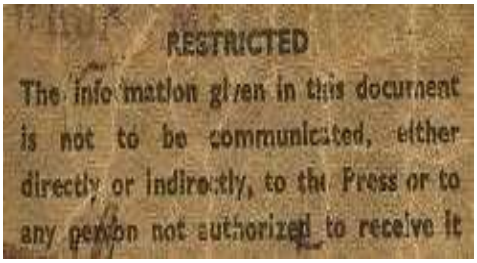
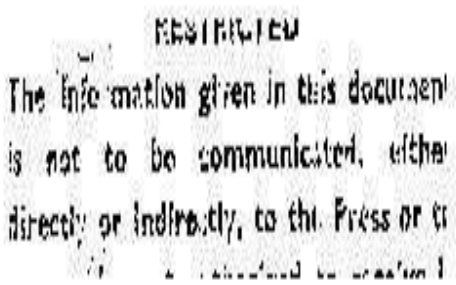
(e) (f)

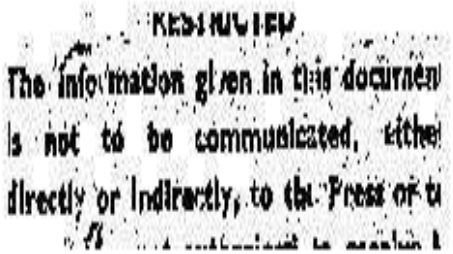
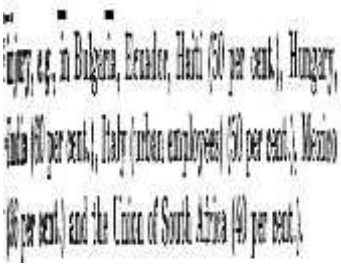
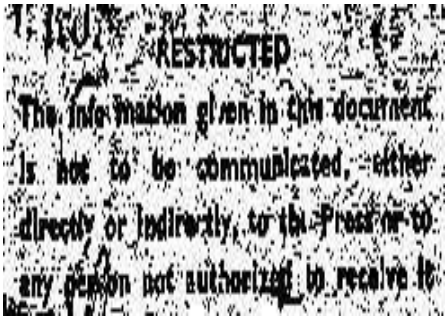
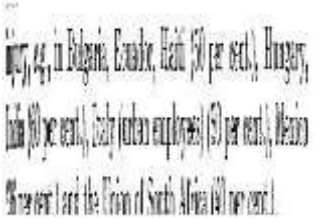
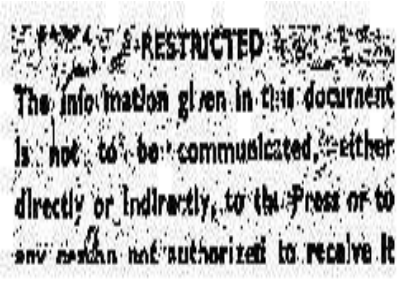
Figure 35 - Removal of false positives through morphological operations. (a) Detection with false positives (b) Corresponding binary image (c) Result after morphological operations (d) Corresponding binary image (e) Result obtained after shape matching (f) Corresponding binary image

Chapter -7 Results & Discussions

7.1 **Results of Binarization**

In this section the results of binarization are presented for certain images from data set of ancient documents images comprising several types of degradations and qualities (pigments, holes, humidity traces, ink degradation and thin lines, transparence effect, presence of folds and tears). The results, after applying the aforementioned algorithms recorded and presented for two images from collection (Data Set). Graphical results are shown below in the Table 4.

Technique Used	**Image-1**	**Image-2**
Original Image		
Sauvola Algorithm		

Wolf Algorithm		
Yosef Algorithm		Image Washed Out
Feng Algorithm		
Niblack Algorithm		

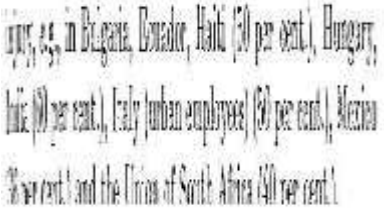
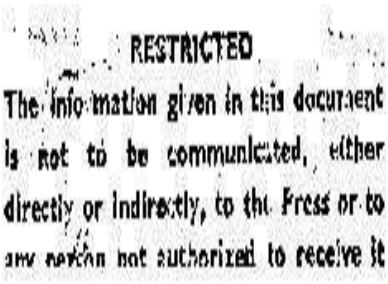
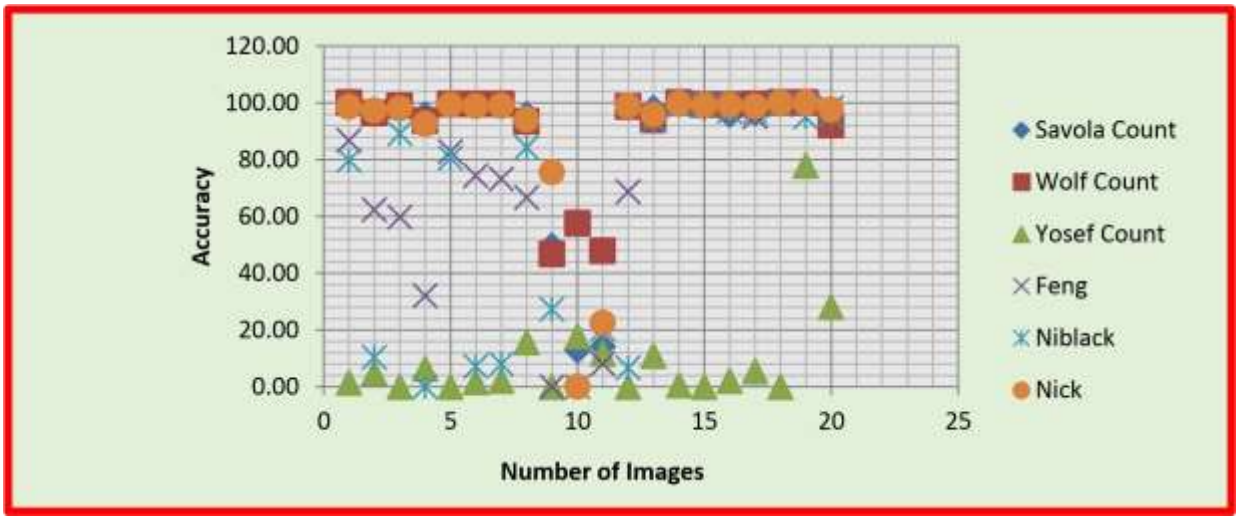
Nick Algorithm		

Table 4 – Results of Binarization Techniques

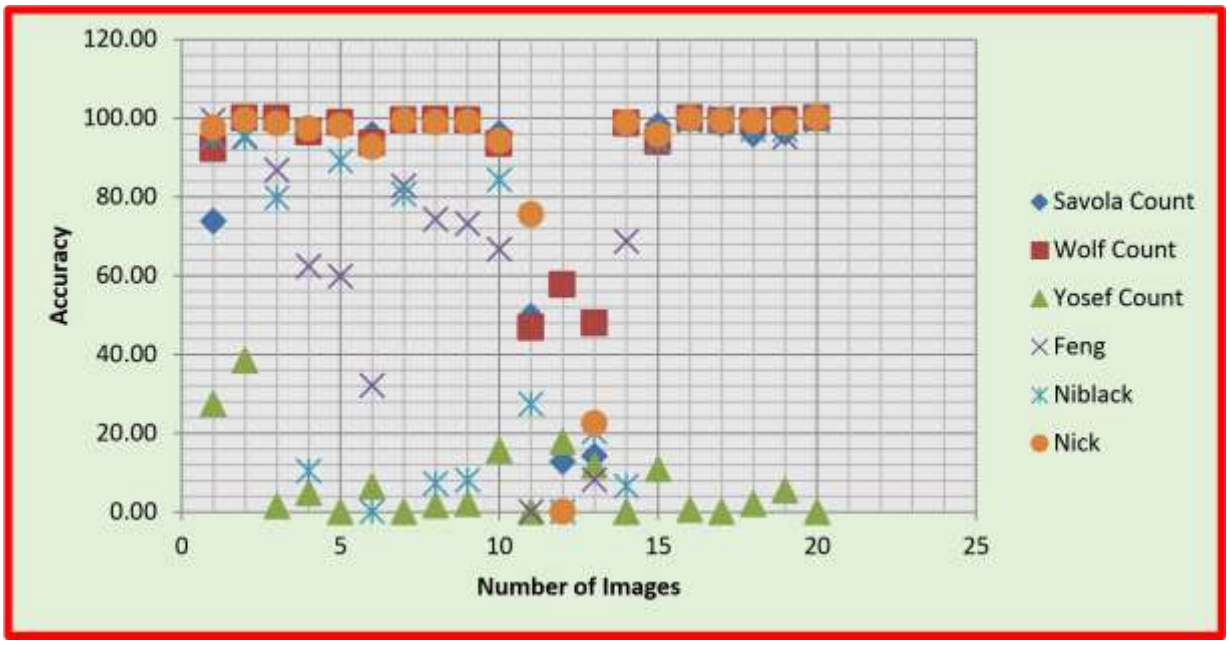
Data set of around 50 images were taken with varying texture, quality and appearance containing averagely 20 thousands image characters. Binarization was done using window size of 7 and 11 for aforementioned techniques. Window size plays a major role in binarization of images. NICK achieved the best accuracy rate of 94% while the Wolf[10] achieved 93% as shown in Figure 36.

Sr. No.	Algorithm	Accuracy (%)
1	Sauvola Count	90
2	Wolf Count	93
3	Yosef Count	6
4	Feng	74
5	Niblack	66
6	NICK	94

Table 5 - Recall rates of various algorithms using Abbyy Fine Reader



a.



b.

Figure 36 - Comparison of the algorithms (a) Window size 11*11 (b) Windows Size 7*7

7.2 **Results of Segmentation**

Results of our employed techniques are displayed in Figure 37 & 38. In Figure 37 characters in blue bounding box are the initial segmented characters which may contain some errors. Characters

in red bounding box are the results of post processing steps employed on segmented characters.

Figure 38 shows the results with accuracy percentage.

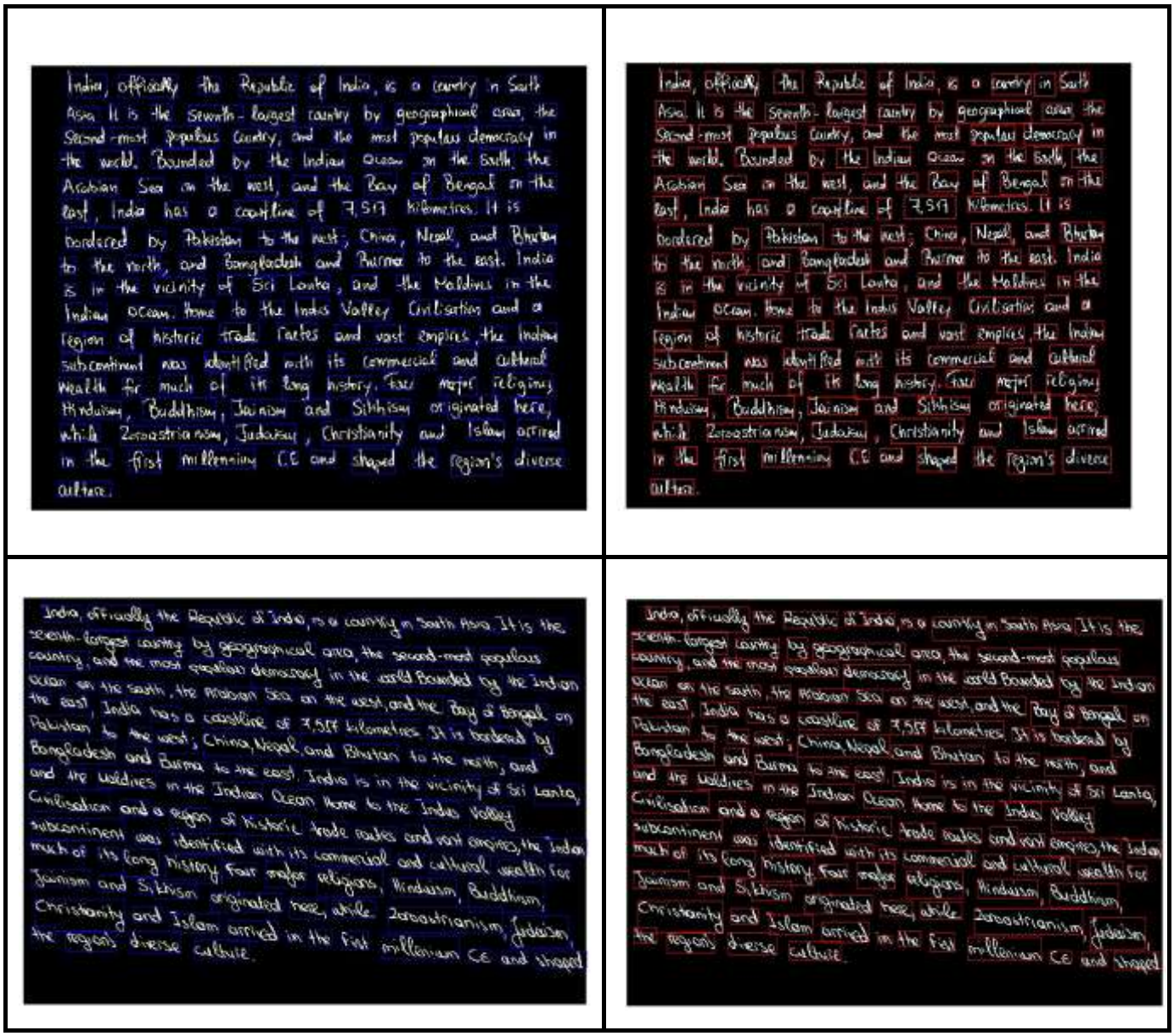


Figure 37- Initial Segmented Characters (Left) Post Processing of Segmented characters (Right)

Sample Image	Total Words	Hits	Misses	Accuracy
1.	161	158	3	98.1%

2.	161	160	1	99.3%
3.	161	159	2	98.75%
4.	161	157	4	97.5%
5.	161	160	1	99.3%

Figure 38 – Results of Segmentation based on same text but different writer on sample images

7.3 **Results of Indexing and Retrieval**

The experimental study of the proposed system was carried out on the IAM handwriting database[69]. For indexing our framework has employed a set of 50 document images containing more than eighty words in each document . The words in these documents were segmented and compared with the clusters in the database generating index files for each of the clusters. For retrieval, a total of 150 query words were presented to the system and the performance was recorded in terms of precision, recall and f-measure. It should be noted that only those words were presented as query for which the corresponding clusters exist in the database. In order to study the effectiveness of the proposed conversion of word image into a contoured shape, evaluations were carried out 'with and without conversion'. The results of these evaluations are presented in Table 6. With the proposed conversion scheme, an overall precision of 84% and recall of 89% were achieved. Using the original word image to extract features, the precision and recall read 72% and 76% respectively. This shows the effectiveness of the proposed conversion of word into a contour shape before extraction of features.

Feature	***Precision***	***Recall***	***F-Measure***
With Shaped Word	0.84	0.89	0.86

Without Shaped Word	0.72	0.76	0.74

Table 6 – Retrieval Results with and without using shaped feature

7.3.1 **Results of Performance Analysis**

In order to study the performance sensitivity to different parameters of the system, a series of experiments were carried out to study the performance evolution as a function of number of clusters in the database. The number of query words in each case is fixed to 35. The results of these evaluations are summarized in Figure 39. Naturally, the precision and recall are higher for a smaller number of clusters and show a gradual but not dramatic decrease as the number of clusters increases.

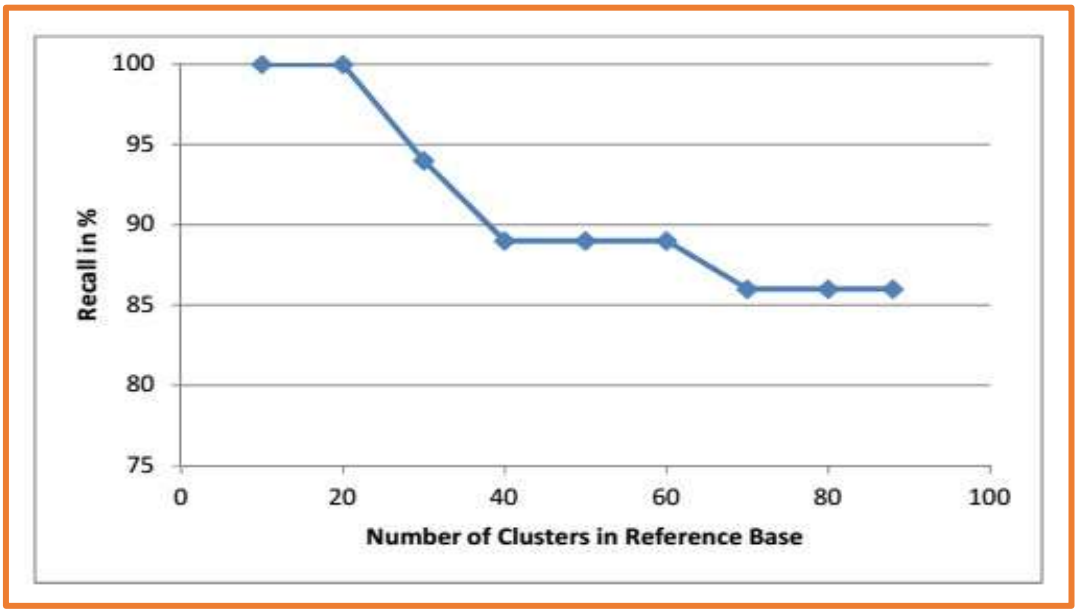


Figure 39 - Recall rates on 35 query words as a function of number of clusters in the reference base

7.3.2 **Results using Principal Component Analysis**

Study has also been carried out to analyze the contribution and significance of different features employed in our study. Principal Component Analysis (PCA) is applied to reduce the dimensionality of the feature vector and keep the most relevant features only. Unlike traditional PCA which transforms the feature space to a new representation space of reduced dimensions, identifying the features was carried out, wherein, showing high variance and hence possess discriminatory properties, a key to good classification. Proposal is to sort the features with respect the corresponding Eigen values of the covariance matrix and keep the top K features. In our experiments, value of $K = 200$ was kept to pick the 200 most discriminative features and study the precision and recall using this subset of features. The retrieval results using the selected subset of features are summarized in Table 7. It can be seen that using a small set of features, the system reports approximately the same precision and recall.

Feature Set Dimension	Precision	Recall	F-Measure
200	0.82	0.88	0.85

Table 7- Retrieval results on a subset of features

7.3.3 **Determining Region of Convergence**

Calculating Region of convergence curves for the data in two different ways has also been undertaken. Firstly, while having a large number of those words as well which are not part of training clusters. Such words cause increase in false positives. Secondly, only those words for testing which have associated words present in training clusters was used. ROC curves in Figures 40, 41 & 42 shows that our proposed system works very well when training set for all possible words are present and segmentation is done perfectly. Table 8 shows the results using various thresholds.

Threshold	True Positives	False Positives	False Negatives	FP %	FN %	FP	TP
0.99	147	0	428	0	74.434	0.000	0.256

0.95	300	0	275	0	47.826	0.000	0.522
0.9	479	0	96	0	16.695	0.000	0.833
0.85	502	1	72	0.173	12.521	0.001	0.873
0.75	508	3	64	0.521	11.130	0.002	0.883
0.65	515	6	54	1.043	9.3913	0.005	0.896
0.55	525	8	42	1.391	7.3043	0.006	0.913
0.45	525	9	41	1.565	7.1304	0.007	0.913
0.35	525	11	39	1.913	6.7826	0.008	0.913
0.25	525	17	33	2.956	5.739	0.013	0.913
0.15	531	24	20	4.173	3.4782	0.018	0.923
0.1	532	30	13	5.217	2.260	0.023	0.925
0.05	538	32	5	5.565	0.869	0.024	0.936
0	539	36	0	6.260	0	0.027	0.937

Table 8 – ROC Table Data

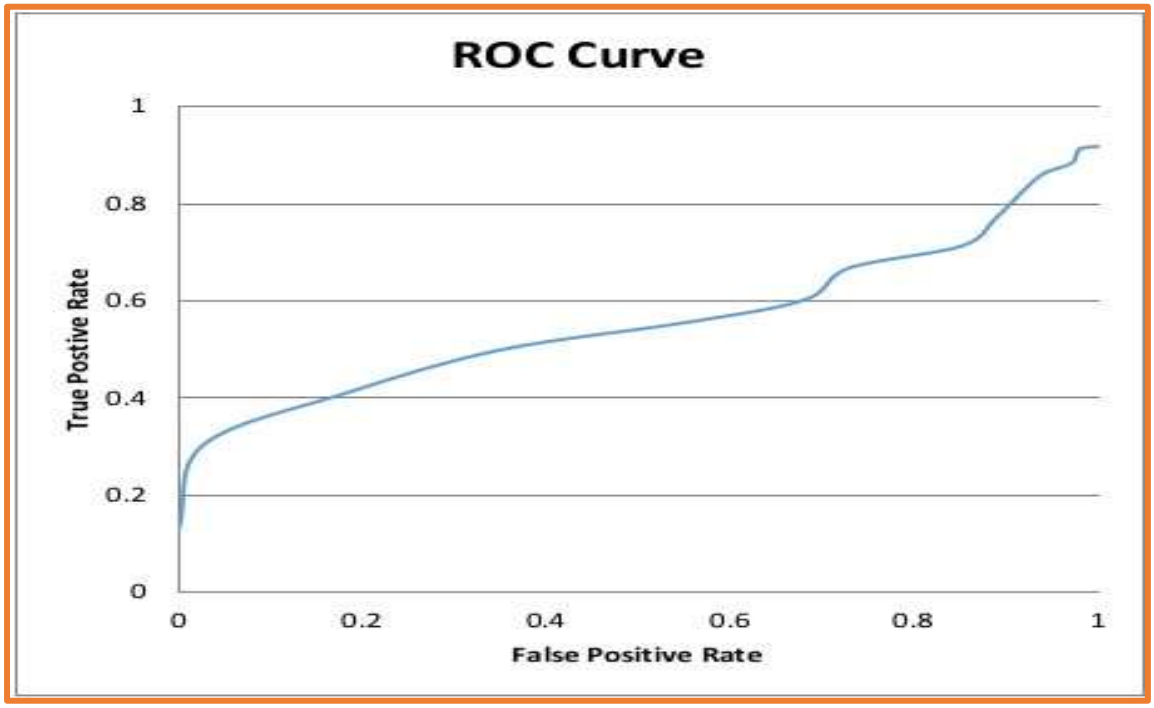


Figure 40 - ROC curve in presence of large number of outliers

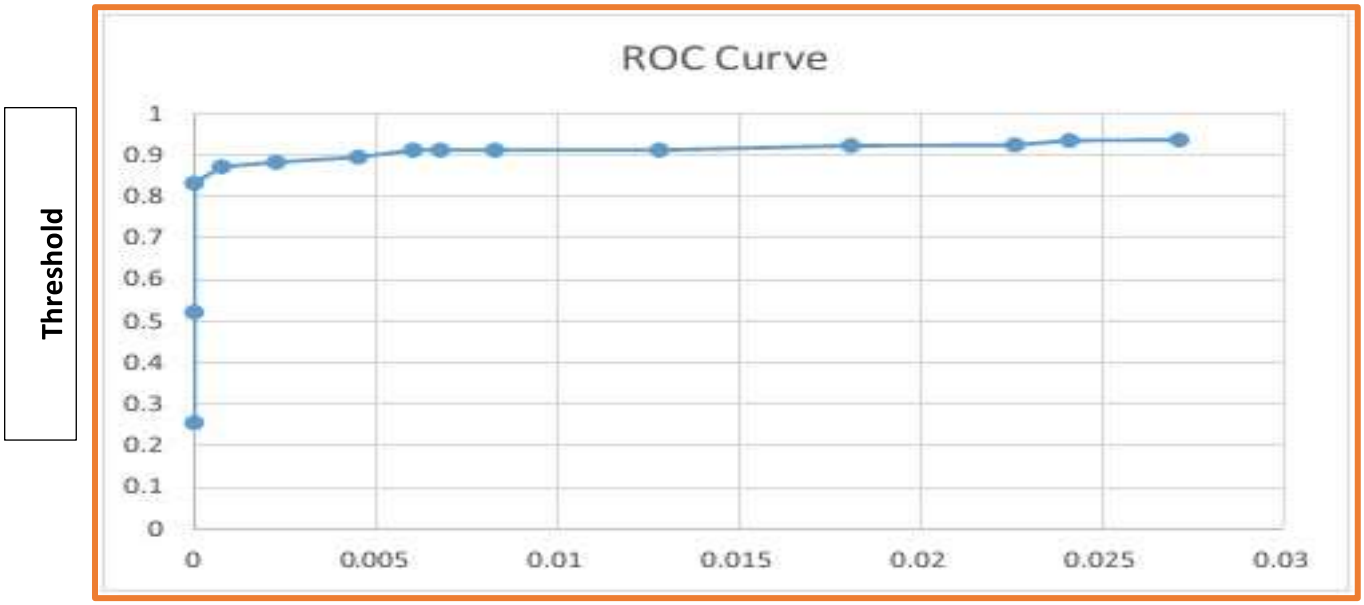


Figure 41 - ROC Curve in absence of outliers

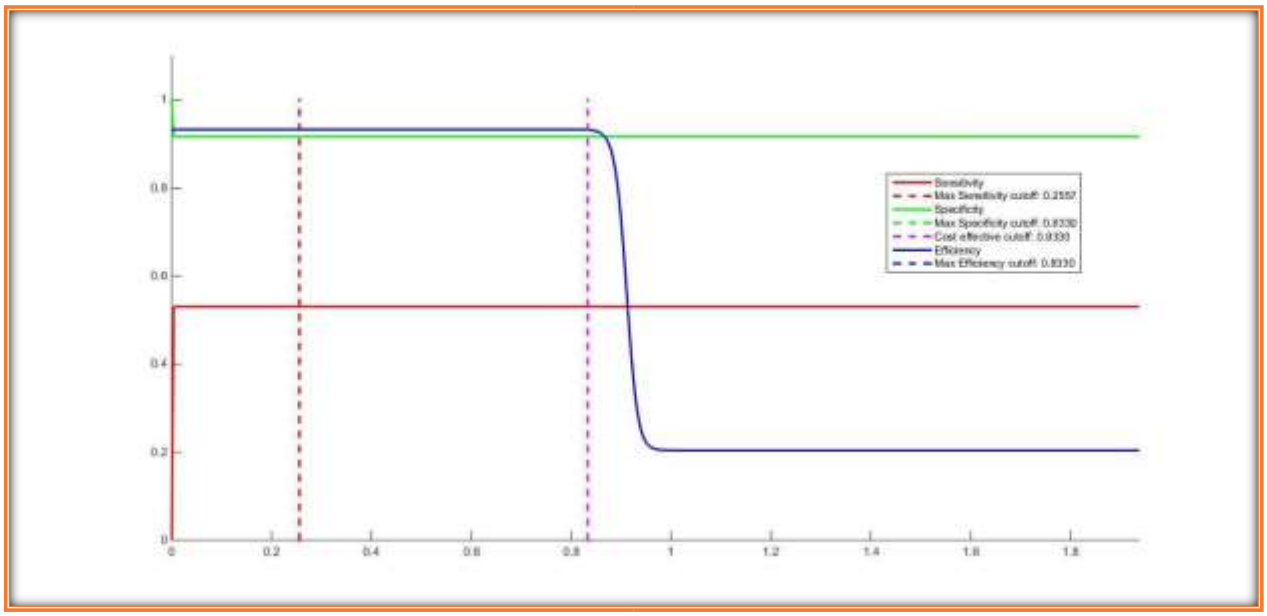


Figure 42 – ROC Curve

Representing the results in form of confusion matrix has also been obtained. In Table 9, results are shown when features are extracted without shape feature while in Table 10, results obtained after

applying the “shaped word” feature is shown. It can be clearly seen that by using our proposed feature, efficiency is system is significantly improved. Table 11 shows result obtained by applying the word-spotting system on 24 words after generation of index files from first 10 images in the IAM database. Table 12 shows the matched frequency of each word that was present in each document. It can be seen that once a complete word was given as a query, results are extremely good. As an example in Table 13, once Word ‘ Conference’ was given it perfectly matched with corresponding words ‘britain’ but also matched with lesser probability ‘action’ as members also contain almost similar characters and is closer to eight characters of conference.

These words are discarded once shaped word feature was applied as shown in Table 11.

	britian	common	conference	government	home	kennedy	members	minister	president	soviet	today
britian	**1**										
common		**0**									
conference			**3**				**1**				
government	**1**			**3**							
home					**0**						
kennedy						**4**					
members							**0**				
minister	**1**							**3**			
president	**1**		**1**					**1**	**1**		
soviet									**1**	**2**	
today											**0**
Ground Truth	**6**	**5**	**6**	**12**	**5**	**8**	**3**	**7**	**8**	**7**	**4**
Total Detections	**4**	**0**	**4**	**3**	**0**	**4**	**1**	**4**	**2**	**2**	**0**
True Detections	**1**	**0**	**3**	**3**	**0**	**4**	**0**	**3**	**1**	**2**	**0**

Table 9 - Confusion Matrix for results without using “shaped word” feature

	britian	common	conference	government	home	kennedy	members	minister	president	soviet	today
britian	5										
common		5									
conference			5	1							
government				9							
home					5						
kennedy						5					
members							2				
minister	1							6			
president				1					8		
soviet								1		5	
today											4
Ground Truth	**6**	**5**	**6**	**12**	**5**	**8**	**3**	**7**	**8**	**7**	**4**
Total Detections	**6**	**5**	**5**	**11**	**5**	**5**	**2**	**7**	**8**	**5**	**4**

True Detections	5	5	5	9	5	5	2	6	8	5	4

Table 10 – Confusion Matrix for results using "Shaped Word" feature

IAM Images/ Words	britian	common	government	labour	main	minister	more	not	one	over	should	than	that	these	they	too	west	year
Image-1				1				1				1	1			1	1	
Image-2	1		2								1		1	1				
Image-3	1									1								
Image-4		1									1		1					
Image-5	1														1			
Image-6			1			1	1	1					2					
Image-7	1						1						1				2	
Image-8			1		1				1				1	1				1
Image-9													1		1			
Image-10						1												
Ground Truth	4	1	4	1	1	2	2	2	1	1	2	1	8	2	2	1	3	1
True Detections	4	1	3	1	1	2	1	1	1	1	1	1	4	1	1	1	2	1

Table 11 – Detection of words in documents using "shaped word" feature

	about	action	africans	agreement	and	are	because	been	britain	but	common	conference	country
about	**2**				1								
action		**3**											
africans			**1**										
agreement				**3**									
and					**46**								
are						**5**							
because			1				**2**						
been								**3**					
britain		1							**5**				
but					1					**6**			
common											**5**		
conference												**5**	
country													**4**

Table 12- Matched Frequency of Sample Words

7.4 **Results of Urdu Script**

The proposed system was evaluated on contemporary printed Urdu documents. Prior to indexing, clusters of ligatures are generated. In our study, a total of 221 clusters were produced which were manually corrected to remove any errors and a final set of 196 clusters of frequently occurring Urdu ligatures was used as the reference base. For indexing, ligatures of 20 printed Urdu documents were indexed and the retrieval was performed using 30 query words having a total of 172 instances in the indexed documents. Out of these 164 instances were correctly retrieved realizing a recall rate of 95%. The results of these experiments are summarized in Table 13 while a retrieval session with query words is illustrated in Figure 43. To evaluate the effectiveness of the validation step, we also computed the precision of the system with the final validation step. It can be seen from Table 14 that the precision of the system improves by using the final validation step which reduces the false positives.

Query Word	**Retrieved Documents**	
انفسیاتی	کراچی یونیورسٹی سے ایک پروفیسر صاحب نے اپنے ایک خط میں مجھے لکھا ہے: "غیبت کے متعلق حضور اکرم صلی اللہ علیہ وسلم کے ارشادات اور قرآن مجید، سب سے میں متفق ہوں کہ غیبت ایسی چیز ہے جیسے بھائی اپنے بھائی کا گوشت کھائے، لیکن میری الجھن انفسیاتی ہے (میں فلسفیات، فلسفہ اور نفسیات کا طالب علم ہوں) انسان اگر غیبت سے اپنے آپ کو روکے رکھے تو یہ گویا تقویٰ ہے، لیکن عام زندگی میں ہم جب ایک دوسرے کا اس کی غیر حاضری میں ذکر کرتے ہیں تو ہمیں اس کا احساس نہیں ہوتا، عورتیں اس معاملے میں بہت آگے ہیں، کسی دعوت سے آنے کے بعد تنقید کا سلسلہ شروع ہو جاتا ہے، کھانے، کپڑے، سب پر تنقید ہوتی ہے، سوال یہ ہے کہ اگر ہم دوسروں کے متعلق بات نہ کریں تو پھر کیا کریں؟ خاموشی یقیناً سب سے بہتر ہے، لیکن وہ کسی ولی اللہ یا بزرگ کو زیب دیتی ہے، ہم کو نہیں، اگر دوسروں کے ذکر کو نکال دیا جائے تو ہماری روزانہ کی گفتگو میں کچھ نہ رہے گا، ہم تمام	وقت خاموش بیٹھے رہیں گے، بنظر اغیبت ایک بہت بڑی انفسیاتی الجھن ہے، ہم تقویٰ اختیار کریں تو نہ کسی کی برائی کریں اور نہ کسی کی برائی سنیں، ایسا کرنے کے لیے ہمیں بہت جدوجہد کرنی ہوگی جو عام زندگی میں ممکن نہیں ہے، غیبت کے بغیر ہماری زندگی ایسی ہوگی، جیسے ساز کے بغیر موسیقی، اس موضوع پر اگر آپ جنگ ہی میں لکھ دیں تو شاید میری طرح بہت سے لوگوں کی الجھن دور ہو سکے۔" پروفیسر صاحب نے جو سوال اٹھایا ہے اس کے جواب کے لیے پہلے یہ سمجھنا ضروری ہے کہ "غیبت" کیا چیز ہے؟ اسے سمجھنے کے لیے کہیں دور جانے کی ضرورت نہیں، خود حضور صلی اللہ علیہ وسلم نے
شخص	ہے جب وہ اس شخص کی ناگواری یا دل آزاری کا سبب ہو، اس کے بغیر نہیں، پھر غیبت اس وقت ناجائز اور حرام ہے جب اس کا کوئی جائز مقصد نہ ہو، لیکن اگر "غیبت" کسی جائز اور معقول وجہ سے کی جائے، تو وہ حرام نہیں، مثلاً ایک مظلوم شخص کسی کے ظلم کا نشانہ بنا ہو اور وہ ظالم کی غیر موجودگی میں اپنی مظلومیت کا ذکر کرے تو یہ جائز ہے، خواہ ظالم کو ناگواری کیوں نہ ہو، اسی طرح اگر کوئی شخص کسی کو برائی اس لیے بتائی ضروری ہو کہ لوگ اس کی برائی کا انکار نہ ہوں اور اس کی دھوکا بازی یا اس کے کسی اور شر سے محفوظ رہیں تو یہ غیبت بھی ناجائز نہیں ہے، بلکہ بعض اوقات واجب ہو جاتی ہے، لیکن اس قسم کی کسی وجہ کے بغیر کسی شخص کی برائی محض تفریح طبع کے لیے یا اس کی تذلیل کے لیے اس طرح اس کے پیچھے بیان کرنا ضرور حرام ہے اور سخت حرام ہے، جس سے اس کی دل شکنی اور دل آزاری ہو یا اسے تکلیف پہنچے، جس غیبت کو قرآن کریم نے حرام قرار دے کر اسے مردہ بھائی کا گوشت کھانے سے تعبیر کیا ہے، وہ یہی غیبت ہے۔	بڑے مختصر اور جامع لفظوں میں "غیبت" کی پوری حقیقت بیان فرما دی ہے۔ آپ صلی اللہ علیہ وسلم نے فرمایا کہ: "غیبت یہ ہے کہ تم اپنے بھائی کا تذکرہ (اس کی غیر حاضری میں) اس انداز سے کرو کہ (اگر اسے پتہ چلے تو) اسے ناگوار ہو۔" "غیبت" کی اس تعریف میں بنیادی اہمیت اس بات کو حاصل ہے کہ کسی کا تذکرہ اس طرح کیا جائے کہ وہ اس کے لیے ناگواری کا موجب ہو، اگر اس بات کا یقین ہے کہ اس تذکرے سے اسے ناگواری نہیں ہوگی تو وہ غیبت نہیں ہے، خواہ وہ اس کی کسی برائی ہی کا بیان ہو، لہٰذا اگر کچھ دوست آپس میں بے تکلف ہیں اور ان کے درمیان اسی مذاق اس طرح چلتا رہتا ہے کہ اس میں کسی شخص کی واقعی برائی کا بیان اسے ناگوار نہیں گزرتا اور ایسی صورت میں وہ اپنے کسی غیر حاضر دوست کا تذکرہ اسی بے تکلفی کے ماحول

Figure 43 - Retrieval session with the system

An inspection of the retrieval errors reveals that few of the errors are caused by ligature classes which are visually very close to one another. Some of the closely resembling ligatures are summarized in Figure 44 where it can be seen that it is very challenging to discriminate between these classes. Besides, some words only differ by very few ligatures. For example افغان and افغانستان differ only by one ligature نستا so when searching for the later, the first word may also be retrieved.

Query Instances	True Positives	False Positives	False Negatives	Recall	Precision
172	164	18	8	95.34%	90.10%

Table 13 - Summary of Results without Shaped Feature

Query Instances	True Positives	False Positives	False Negatives	Recall	Precision
172	164	6	8	95.34%	96.47%

Table 14 – Summary of Results with the application of Shaped Feature

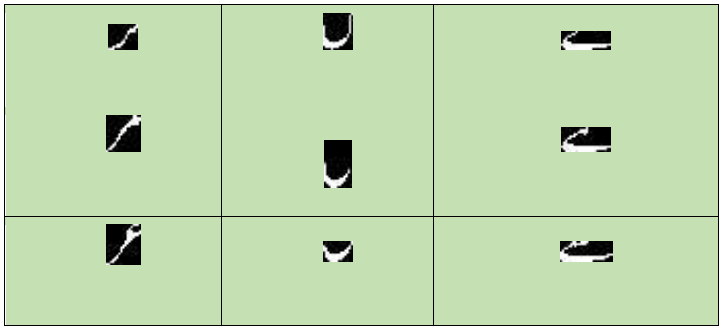


Figure 44 - Examples of visually resembling ligatures (in columns)

The experimental evaluations of the system were carried out on an HP core i7 2.40 GHz machine with 8 GB RAM. On the average, our system took less than 5 seconds to process the query word and display the documents to user. The retrieval time, naturally, is a function of number of ligatures in the query word. Table 15 summarizes the retrieval times for sample query words.

Query Word	Documents Retrieved		Detected Words	Time (secs)
	Initial	After Validation		
غیبت	5	4	19	1.96

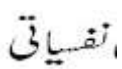
	15	2	2	6.92
	2	2	5	1.39
	2	2	3	1.43
	7	2	2	4.00

Table 15 - Retrieval times of sample query words

Two publications have been successfully completed where promising results have been acknowledged [92, 93].

7.5 **Comparative Analysis of Own Vs Others**

A comparison of well-known word spotting systems is given in Table 16. Our Proposed framework of script independent using shaped word feature is a new innovation and has not been used / experimented by researchers around the globe. However, the results calculated using this framework has been compared with similar work and our proposed methodology has achieved significant success over the work already carried out by other researchers.

Study	**Type of Approach**	**Features**	**Classifier**	**Database**	**Results**
Safwan et al. [42]	Line based	Gradient features	Hidden Markov Model	IAM (English), AMA (Arabic) and LAW (Devanagari)	Average Precision = 60%
Frinken et al. [43]	Line based	CTC Token Passing algorithm	Neural Networks	IAM, GW and PARZIVAL	Precision IAM=76% GW=71% PARZIVAL=92%

Serrano et al. [44]	Word level segmentation	Means of the Gaussians	Hidden Markov Model	GW and IFN/ENIT	Average Precision = 91%
Fischer et al [45]	Line based	Geometrical features	Hidden Markov Model	IAM, GW	Average Precision IAM= 55% GW = 74%
Kumar et al. [46]	Line based	Gradient, Structural Concavity and Intensity features	Bayesian logistic regression classifier	IAM (English), AMA (Arabic) and LAW (Devanagari)	Average Precision IAM = 49% AMA = 54% LAW = 51%
Ranjan et al [47]	Word level segmentation	Profile features	Support Vector Machine	Custom English documents	Average Precision = 81%
Almazan et al. [48]	Word level segmentation	SIFT	Support Vector Machine	IAM, GW and IIIT5K	Average Precision IAM = 55.73% GW = 92.90% IIIT5K = 72.28%
Own Proposed Method / Framework	Word Level segmentation	Shape Descriptors	Support Vector Machine	IAM and Urdu	Precision= 84% Recall = 89%

Table 16 - Comparison with few Existing Word Spotting Systems

Chapter -8 Conclusion and Future Work

8.1 **Conclusion**

In this work, a word spotting based solution to document indexing and retrieval is presented. Words extracted from document images are represented by a set of features are grouped into clusters. These clusters are used to train the multi-class support vector machine (SVM). The documents to be indexed are segmented into words and the nearest match cluster for each word is determined using the SVM. The index file of the respective cluster is updated to keep information about the document and the location of the word in the document. During retrieval a query word

presented to the system is matched against the clusters in the database and the index file of the matched cluster is used to retrieve the documents containing instances of the query word. The proposed scheme evaluated on the handwritten images of the IAM database realized promising precision and recall rates. Proposed framework was also successfully tested on Urdu images.

Challenges in handwritten documents and cursive fonts are enormous, however, through innovative techniques and dynamic approach of problem solving , these aspects can be resolved. One of the major limitation of our work was non-availability of an authentic database of Urdu language. Reputed organizations / journals do held competitions in handwritten documents languages. Competitions may be organized at National / International / University level based on Urdu Language for better utility and research.

The system has been tested using MATLAB with an i7 workstation with 8 GB RAM.

8.2 **Future Work**

The present system has been evaluated on contemporary handwritten images. With some added preprocessing steps, the system may also be adapted for retrieval of ancient historical documents where the traditional OCR systems cannot work. The system can also be extended further by introducing a fully automated clustering technique which allows creation of new clusters as unseen words are added to the system.

As regards to Urdu language, its cursive nature, difficult word formation approach, appearance of characters within words, overlapping of partial words, dots and diacritic marks, segmentation and varying styles poses enormous challenges and our proposed system will be able to perform well after the segmentation issues are addressed. We are planning to further extend our research by utilizing / experimenting on varying styles of scripts.

The number of training clusters can also be increased to address any new incoming word in documents during the indexing phase. Our idea to address the problem of outliers is to collect them in separate cluster and then classify them based on already described features. In this way, new clusters will be obtained which will be added to the system through clustering. However during development of such system, involvement of manual inspection needs to be reduced as much possible and that's the challenging task which can be performed as an improvement in our proposed system.

Feasibility on socio-economical need of proposed framework will also be initiated; supported with some industrial statistics and requirements.

References

[1] K. Khurshid, C. Faure, and N. Vincent, "Fusion of Word Spotting and Spatial Information for Figure Caption Retrieval in Historical Document Images," in *ICDAR*, Spain, 2009.

[2] A. Vinciarelli, "A Survey On Off-Line Cursive Word Recognition," *Pattern Recognition,* vol. 35, pp. 1433–1446, 2002.

[3] R. P. a. S. N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 63-84, 2000.

[4] A. A. a. A. C. Downton, "Special issue on the analysis of historical documents," *International Journal on Document Analysis and Recognition,* vol. 9, pp. 75-77, 2007.

[5] B. a. P. I. Gatos, "Segmentation-free word spotting in historical printed documents," in *10th International Conference on Document Analysis and Recognition*, 2009.

[6] R. Bertolami, Gutmann, C., and Bunke, H., "Shape code based lexicon reduction for offline handwrittenword recognition," in *The Eighth IAPR Workshop on Document Analysis Systems*, 2008.

[7] Y. L. F. a. E and H. Leydier, "Textual indexation of ancient documents," presented at the ACM symposium on Document engineering, 2005.

[8] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis, "A complete optical character recognition methodology for historical documents " in *The Eighth IAPR Workshop on Document Analysis Systems*, 2008.

[9] R. F. Moghaddam and M. Cheriet, "Application of multi-level classifiers and clustering for automatic word spotting in historical document images," in *10th International Conference on Document Analysis and Recognition*, 2009.

[10] K. Terasawa, H. Imura, and Y. Tanaka, "Automatic evaluation framework for word spotting," in *10th International Conference on Document Analysis and Recognition*, 2009.

[11] A. F. Volkmar Frinken, Horst Bunke, R. Manmatha, "Adapting BLSTM Neural Network based Keyword Spotting trained on Modern Data to Historical Documents," in *12th International Conference on Frontiers in Handwriting Recognition*, 2010.

[12] A. F. Volkmar Frinken, Horst Bunke, R. Manmatha, "A Novel Word Spotting Method Based on Recurrent Neural Networks," *IEEE transactions on pattern analysis and machine intelligence,* vol. 34, 2012.

[13] M. X. a. L. C. Yuchen Liu, "Improved Keyword Spotting System by Optimizing Posterior Confidence Measure Vector Using Feed-forward Neural Network," in *International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, 2014.

[14] A. Tarafdar, U. Pal, P. P. Roy, N. Ragot, and J.-Y. Ramel, "A Two-Stage Approach for Word Spotting in Graphical Documents," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 319-323.

[15] S. Impedovo, F. M. Mangini, G. Pirlo, D. Barbuzzi, and D. Impedovo, "Voronoi Tessellation for Effective and Efficient Handwritten Digit Classification," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 435-439.

[16] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM Based Word Spotting in Handwritten Documents Using Subword Models," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3416-3419.

[17] A. F. V. Frinken, and H. Bunke, "A Novel Word Spotting Algorithm Using Bidirectional Long ShortTerm Memory Neural Networks," in *4th Workshop on Artificial Neural Networks in Pattern Recognition*, 2010.

[18] V. G. Huaigu Cao, Anurag Bhardwaj, "Unconstrained handwritten document retrieval," *International Journal on Document Analysis and Recognition* vol. 14, pp. 145-157, 2011.

[19] B. Moysset and C. Kermorvant, "On the Evaluation of Handwritten Text Line Detection Algorithms," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 185-189.

[20] K. C. S. Mallikarjun Hangarge, Rajmohan Pardeshi, "Directional Discrete Cosine Transform for Handwritten Script Identification," in *12th International Conference on Cocuments Analysis and Recognition*, 2013.

[21] Sebastian Sudholt and G. A. Fink, "HOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents," in *ICFHR*, 2016.

[22] J. Almaz, A. Gordo, A. Fornes, and E. Valveny, "Word Spotting and Recognition with Embedded Attributes," presented at the Transactions on Pattern Analysis and Machine Intelligence, 2014.

[23] D. Aldavert, R. Toledo, and J. Llados, "Towards Query-by-Speech Handwritten Keyword Spotting," in *International Conference on Document Analysis and Recognition*, 2015.

[24] L. Rothacker and G. A. Fink, "Segmentation-free query-by-string word spotting with bag-offeatures HMMs," in *International Conference on Document Analysis and Recognition*, 2015.

[25] H. S. Baird, "Difficult and urgent open problems in document image analysis for libraries," in *1st International workshop on Document Image Analysis for Libraries*, 2004.

[26] A. Antonacopoulos and D. Karatzas, "Semantics-based content extraction in typewritten historical documents," in *8th International Conference on Document Analysis and Recognition*, 2005.

[27] C. L. Tan and Z. Zhang, "Text block segmentation using pyramid structure," in *Proceedings of SPIE, the International Society for Optical Engineering*, 2001.

[28] T. M. Rath, *Retrieval of handwritten historical document images*: PhD thesis, Graduate School of the University of Massachusetts Amherst., 2005.

[29] O. Okun, D. Doermann, and M. Pietikainen, *Page segmentation and zone classification: The state of the art.*: Technical report, University of Maryland, 1999.

[30] J. Duong, M. Ct, Emptoz, H., and C. Suen, "Extraction of text areas in printed document images," in *ACM Symposium on Document Engineering DocEng'01*, Atlanta (USA), 2001.

[31] N. Journet, R. Mullot, V. Eglin, and R. J.-Y. Ramel, "Analysed'images de documents anciens:categorisation de contenus par approche texture," in *CIFED, Colloque International sur l'Ecrit et le Document*, 2006.

[32] Z. Shi and V. Govindaraju, "Multi-scale techniques for document page segmentation," in *Eighth International Conference on Document Analysis and Recognition (ICDAR)*, 2005.

[33] N. Journet, V. Eglin, J.-Y. Ramel, and R. Mullot, "Ancient printed documents indexation: a new approach. In Pattern Recognition and Data Mining," *Lecture Notes in Computer ScienceLectures Notes in Computer Science,* pp. 513-522, 2005.

[34] C. Faure and N. Vincent, "Simultaneous detection of vertical and horizontal text lines based on perceptual organization," in *16th Document Recognition and Retrieval Conference, DRR*, 2009.

[35] S. Bukhari, F. S., and T. Breuel, "Segmentation of Curled Textlines Using Active Contours," presented at the Eighth IAPR Workshop on Document Analysis System, 2008.

[36] K. Y. Wong, R. G. Casey, and F. M. Wahi, "Document analysis Syatem," *IBM Journal of Research Development,* vol. 26, pp. 647 – 656, 1982.

[37] Y. Y. Tang, C. D. Yan, M. Cheriet, a. Suen, and C. Y., *Automatic analysis and understanding of documents, Handbook of pattern recognition & computer vision*: World Scientific Publishing, 1993.

[38] K.-H. Lee, Y.-C. Choy, and S.-B. Cho, "Geometric structure analysis of document images: A knowledge-based approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 1224 – 1240, 2000.

[39] A. K. Jain, *Fundamentals of digital image processing*: Prentice Hall, 1989.

[40] T. Pavladis and J. Zhou, "Page segmentation by white streams," in *International conference on document analysis and retrieval*, 1991.

[41] I. Ar and M. E. Karsligil, "Text area detection in digital documents images using textural features," *Computer Analysis of Images and Patterns (CAIP), Lecture Notes in Computer Science* vol. 4673, pp. 555–562, 2007.

[42] T. Randen and J. H. Husøy, "Segmentation of text/image documents using texture approaches," in *Proc. NOBIM-Konferansen-94*, Norway, 1994.

[43] B. Gatos, Ntirogiannis, K., and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *10th International Conference on Document Analysis and Recognition*, 2009.

[44] T. Adamek, N. E. O'Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents.," *IJDAR,* vol. 9, pp. 153 – 165, 2007.

[45] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis, "Keywordguided word spotting in historical printed documents using synthetic data and user feedback," *IJDAR,* vol. 9, pp. 167 – 177, 2007.

[46] T. M. Rath, S. Kane, A. Lehman, E. Partridge, and R. Manmatha, "Indexing for a digital library of geaorge washington's manuscripts: A study of word matching techniques," University of Massachusetts Amherst2002.

[47] S. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 149 – 164, 2001.

[48] T. Steinherz, E. Rivlin, and N. Intrator, "Offline cursive script word recognition - a survey," *International Journal on Document Analysis and Recognition,* pp. 90-110, 1999.

[49] J. L. Rothfeder, S. Feng, and T. M. Rath, "Using corner features correspondences to rank word images by similarity," in *Conference on Computer vision and pattern recognition*, 2003.

[50] J. P. Lewis, "Fast template matching," in *Vision Interface*, 1995.

[51] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR,* vol. 9, pp. 139 – 152, 2007.

[52] G. M. Reicher, "Perceptual recognition as a function of meaningfulness of stimulus material," *Journal of Experimental Psychology,* pp. 275–280, 1969.

[53] J. Li, Y. Fan, and N. Le, "Document image retrieval with local feature sequences," in *0th International Conference on Document Analysis and Recognition*, 2009.

[54] A. Andreev and N. Kirov, "Word image matching based on hausdorff distances," in *10th International Conference on Document Analysis and Recognition*, 2009.

[55] S. Marinai, S. Faini, E. Marino, and G. Soda, "Efficient word retrieval by means of som clustering and pca," in *Workshop on Document Analysis Systems VII Lecture Notes in Computer Science*, 2006.

[56] S. Marinai, E. Marino, and G. Soda, "Exploring digital libraries with document image retrieval," in *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, 2007.

[57] K. Zagoris, N. Papamarkos, and C. Chamzas, "Web document image retrieval system based on word spotting," in *IEEE International Conference on Image Processing*, 2006.

[58] M. Rusinol and J. Llados, "Word and symbol spotting using spatial organization of local descriptors," in *The Eighth IAPR Workshop on Document Analysis Systems*, 2008.

[59] Shunyi Yao, Y. Wen, and Y. Lu., " HoG based Two-Directional Dynamic Time Wraping for Handwritten Word Spotting," presented at the ICDAR 2015.

[60] S. Bai, L. Li, and C. Tan, "Keyword Spotting in Document Images through Word Shape Coding," in *10th International Conference on Document Analysis and Recognition, Barcelona*, 2009.

[61] R. Bertolami, C. Gutmann, and H. Bunke, "Shape code based analysis for libraries," in *1st International workshop on Document Image Analysis for Libraries*, 2004.

[62] V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos, "Word-based adaptive ocr for historical books," in *10th International Conference on Document Analysis and Recognition*, 2009.

[63] P. E. Mitchell and H. Yan, "Newspaper document analysis featuring connected line segmentation," in *Sixth International Conference on Document nalysis and Recognition*, 2001.

[64] Y. Lu and M. Shridhar, "Character segmentation in handwritten words," *Pattern Recognition,* pp. 77-96, 1996.

[65] Y. Leydier, LeBourgeois, F., and Emptoz, H., "Textual indexation of ancient documents," in *ACM symposium on Document engineering*, 2005, pp. 111-117.

[66] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 690 – 706, 1996.

[67] K. Khurshid, C. Faure, and N. Vincent, "Feature based word spotting in ancient printed documents," in *8th edition of PRIS in 10th Int'l conference on Enterprise Information Systems, ICEIS* Spain, 2008.

[68] R. Hussain, A. Raza, I. Siddiqi, K. Khurshid, and C. Djeddi, "A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation," *Eurasip Journal of Image and Video Processing,* 2015.

[69] U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition.," *International Journal on Document Analysis and Recognition,* vol. 5, pp. 39-46, 2002.

[70] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1996.

[71] R. Manmatha, C. Han, E. M. Riseman, and W. B. Croft, "Indexing handwriting using word matching," in *1st ACM Internationall Conference on Digital Libraries*, 1996.

[72] G. Leedham, C. Yan, Takru, K., J. H. N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in *Seventh International Conference on Document Analysis and Recognition (ICDAR)*, 2003.

[73] N. Otsu, "A threshold selection method from grey level histogram," *IEEE* vol. 9, pp. 62-66, 1969 Transactions on Systems, Man, and Cybernetics.

[74] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition,* vol. 39, pp. 317 – 327, 2006.

[75] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition,* vol. 33, pp. 225–236, 2000.

[76] Ø. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," vol. 17, pp. 312 – 315, 1995.

[77] W. Niblack, *An Introduction to Digital Image Processing*: Prentice Hall, 1986.

[78] C. Wolf and J.-M. Jolion, "Extraction and recognition of artificial text in multimedia documents," *Pattern Analysis and Applications,* pp. 309 – 326, 2003.

[79] M.-L. Feng and Y.-P. Tan, "Contrast adaptive binarization of low quality document images," *IEICE Electron. Express,* vol. 1, pp. 501–506, 2004.

[80] K. Khurshid, I. Siddiq, C. Faure, and N. Vincent, "Comparison of niblack inspired binarization methods for ancient documents," in *16th Document Recognition and Retrieval Conference,* , USA, 2009.

[81] R. Hussain and A. Masood, "Word Segmentation of Hand Written English Text for Improvement of Word Spotting Results," presented at the Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016), UK, 2016.

[82] I. Siddiqi and N. Vincent, "A set of chain code based features for writer recognition," in *Proc. of 10th International Conference on Document Analysis and Recognition*, 2009.

[83] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," *Pattern Recognition,* vol. 43, pp. 3853–3865, 2010.

[84] H. Y. Y. Nakano, "Cursive handwritten word recognition using multiple segmentation determined by contour analysis," *IEICE Transactions on Information and Systems,* vol. E79, pp. 464–470, 1996.

[85] F. Kimura, N. Kayahara, Y. Miyake, and M. Shridhar, "Machine and human recognition of segmented characters from handwritten words," in *Proc. of the 4th International Conference on Document Analysis and Recognition*, 1997.

[86] M. Blumenstein, B. Verma, and H. Basli, "A novel feature extraction technique for the recognition of segmented handwritten characters," in *Proc. of the Seventh International Conference on Document Analysis and Recognition*, 2003.

[87] M.E.Dehkordi, N.Sherkat, and T.Allen, "Handwriting style classification. ," *International Journal of Document Analysis and Recognition,* vol. 6, pp. 55–74, 2003.

[88] I. Siddiqi, C. Djeddi, Raza, A., and L. Souici-meslati, "Automatic analysis of handwriting for gender classification," *Pattern Analysis and Applications,* 2014.

[89] K. Wall and P.-E. Danielsson, "A fast sequential method for polygonal approximation of digitized curves. ," *Computer Vision, Graphics, and Image Processing,* vol. 28, 1984.

[90] K. Khurshid, C. Faure, and C. Faure, "Word spotting in historical printed documents using shape and sequence comparisons," presented at the Pattern Recognition 45(7), 2012.

[91] A. Bensefia, T. Paquet, and L. Heutte, "A writer identification and verification system," *Pattern Recognition Letters,* vol. 26, pp. 2080–2092, 2005.

[92] Muhammad Rashid Hussain, Haris Ahmad Khan, Imran Siddiqi, Khurram Khurshid, "Language Independent Keyword Based Information Retrieval System of Handwritten Documents using SVM Classifier and Converting Words into Shapes," *Pakistan Journal of Engg. & Applied Sciences,* vol. 19, pp. 63-76, 2016.

[93] Raashid Hussain, Imran Siddiqi, Khurram Khurshid, Asif Masood, "Keyword based Information Retrieval System for Urdu Document Images," presented at the 11th International Conference on Signal-Image Technology & Internet-Based Systems, Thailand, 2015.

Appendix-A

Sample Images

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the fist millenium CE and shaped the region's diverse culture.

India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the South, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

India, officially the Republic of India, is a country in South, Asia. It is the seventh-largest country by geographical area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the west, and the Bay of Bengal on the east, India has a coastline of 7,517 kilometres. It is bordered by Pakistan to the west; China, Nepal, and Bhutan to the north; and Bangladesh and Burma to the east. India is in the vicinity of Sri Lanka, and the Maldives in the Indian Ocean. Home to the Indus Valley Civilisation and a region of historic trade routes and vast empires, the Indian subcontinent was identified with its commercial and cultural wealth for much of its long history. Four major religions, Hinduism, Buddhism, Jainism and Sikhism originated here, while Zoroastrianism, Judaism, Christianity and Islam arrived in the first millennium CE and shaped the region's diverse culture.

That argument did not apply to the Polaris submarine. So long as the Soviet Union had nuclear weapons, the West, somewhere, must have them too. It was far better for a weapon used for retaliatory purposes to be under the sea rather than on land. This was why the Labour Party did not think it right to oppose the Polaris depot ship.

There are three kinds of reasons that justify the protests and these should carry weight with the U.S. Government, Earl Russell suggested. "The first of these reasons is the importance of preserving the hitherto cordial relations between the U.S. and Great Britain, not only in Government circles, but in public opinion." Earl Russell says it is inevitable, though profoundly regrettable, that the agitation against the Polaris base has generated some antagonism to the policy of the United States.

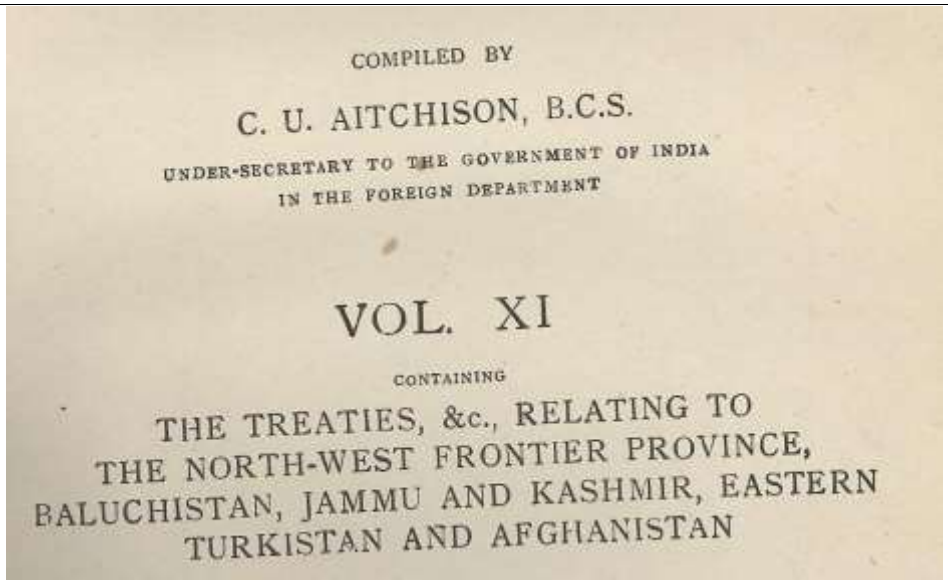
THE PATHAN REVOLT

IN

NORTH-WEST INDIA.

Being a complete narrative of the Maizar Outrage and Punitive Expedition in the Tochi Valley, the Siege and Relief of the Malakand and Chakdara Fort, the Battle of Shabkadar, the Mohmand Rising, the Afridi Seizure of the Khyber Pass, the Orakzai Attacks on the Samana Outposts, the Fall of Saragarhi, the Defence of Gulistan, General Jeffreys' Hot Fighting with the Mamunds, together with an account of the Punitive operations by Brigadier-General Sir Bindon Blood, K.C.B., and Brigadier-General E. R. Elles, C.B., in the Swat Valley and the Mohmand Country.

By H. WOOSNAM MILLS,
(Of the *Civil & Military Gazette*, Lahore).



COMPILED BY

C. U. AITCHISON, B.C.S.

UNDER-SECRETARY TO THE GOVERNMENT OF INDIA
IN THE FOREIGN DEPARTMENT

VOL. XI

CONTAINING

THE TREATIES, &c., RELATING TO
THE NORTH-WEST FRONTIER PROVINCE,
BALUCHISTAN, JAMMU AND KASHMIR, EASTERN
TURKISTAN AND AFGHANISTAN

کراچی یونی ورسٹی سے ایک پروفیسر صاحب نے اپنے ایک خط میں مجھے لکھا ہے:

''غیبت کے متعلق حضور اکرم صلی اللہ علیہ وسلم کے ارشادات اور قرآن مجید، سب سے میں متفق ہوں کہ غیبت ایسی چیز ہے جیسے بھائی اپنے بھائی کا گوشت کھائے، لیکن میری الجھن نفسیاتی ہے (میں نفسیات، فلسفہ اور عمرانیات کا طالب علم ہوں) انسان اگر غیبت سے اپنے آپ کو روکے رکھے تو یہ گویا تقویٰ ہے، لیکن عام زندگی میں ہم جب ایک دوسرے کا اس کی غیر حاضری میں ذکر کرتے ہیں تو ہمیں اس کا احساس نہیں ہوتا، عورتیں اس معاملے میں بہت آگے ہیں، کسی دعوت سے آنے کے بعد تنقید کا سلسلہ شروع ہو جاتا ہے، کھانے، کپڑے، سب پر تنقید ہوتی ہے، سوال یہ ہے کہ اگر ہم دوسروں کے متعلق بات نہ کریں تو پھر کیا کریں؟ خاموشی یقیناً سب سے بہتر ہے، لیکن وہ کسی ولی اللہ یا بزرگ کو زیب دیتی ہے، ہم کو نہیں، اگر دوسروں کے ذکر کو نکال دیا جائے تو ہماری روزانہ کی گفتگو میں کچھ نہ رہے گا، ہم تمام

وقت خاموش بیٹھے رہیں گے، مختصراً غیبت ایک بہت بڑی نفسیاتی الجھن ہے، ہم تقویٰ اختیار کریں تو نہ کسی کی بُرائی کریں اور نہ کسی کی بُرائی سنیں، ایسا کرنے کے لیے ہمیں بہت جدوجہد کرنی ہوگی جو عام زندگی میں ممکن نہیں ہے، غیبت کے بغیر ہماری زندگی ایسی ہوگی، جیسے ساز کے بغیر موسیقی، اس موضوع پر اگر آپ جنگ ہی میں لکھ دیں تو شاید میری طرح بہت سے لوگوں کی الجھن دور ہو سکے۔''

پروفیسر صاحب نے جو سوال اٹھایا ہے اس کے جواب کے لیے پہلے یہ سمجھنا ضروری ہے کہ ''غیبت'' کیا چیز ہے؟ اسے سمجھنے کے لیے کہیں دور جانے کی ضرورت نہیں، خود حضور صلی اللہ علیہ وسلم نے

بڑے مختصر اور جامع لفظوں میں ''غیبت'' کی نپی تلی حقیقت بیان فرما دی ہے، آپ صلی اللہ علیہ وسلم نے فرمایا کہ:

''غیبت یہ ہے کہ تم اپنے بھائی کا تذکرہ (اس کی غیر حاضری میں) اس انداز سے کرو کہ (اگر اسے پتہ چلے تو) اسے ناگوار ہو۔''

''غیبت'' کی اس تعریف میں بنیادی اہمیت اس بات کو حاصل ہے کہ کسی کا تذکرہ اس طرح کیا جائے کہ وہ اس کے لیے ناگواری کا موجب ہو، اگر اس بات کا یقین ہے کہ اس تذکرے سے اسے ناگواری نہیں ہوگی تو وہ غیبت نہیں ہے، خواہ وہ اس کی کسی بُرائی ہی کا بیان ہو، لہٰذا اگر کچھ دوست آپس میں بے تکلف ہیں اور ان کے درمیان ہنسی مذاق اس طرح چلتا رہتا ہے کہ اس میں کسی شخص کی واقعی بُرائی کا بیان اسے ناگوار نہیں گزرتا اور ایسی صورت میں وہ اپنے کسی غیر حاضر دوست کا تذکرہ اسی بے تکلفی کے ماحول

ہے جب وہ اس شخص کی ناگواری یا دل آزاری کا سبب ہو، اس کے بغیر نہیں، پھر غیبت اسی وقت ناجائز اور حرام ہے جب اس کا کوئی جائز مقصد نہ ہو، لیکن اگر ''غیبت'' کسی جائز اور معقول وجہ سے کی جائے، تو وہ حرام نہیں، مثلاً ایک مظلوم شخص کسی کے ظلم کا نشانہ بنا ہو اور وہ ظالم کی غیر موجودگی میں اپنی مظلومیت کا ذکر کرے تو یہ جائز ہے، خواہ ظالم کو ناگوار ہی کیوں نہ ہو، اسی طرح اگر کسی شخص کی کوئی بُرائی اس لیے بتانی ضروری ہو کہ لوگ اس کی بُرائی کا شکار نہ ہوں اور اس کی دھوکا بازی یا اس کے کسی اور شر سے محفوظ رہیں تو یہ غیبت بھی ناجائز نہیں ہے، بلکہ بعض اوقات واجب ہو جاتی ہے، لیکن اس قسم کی کسی وجہ کے بغیر کسی شخص کی بُرائی محض تفریح طبع کے لیے یا اس کی تذلیل کے لیے اس طرح اس کے پیچھے بیان کرنا ضرور حرام ہے اور سخت حرام ہے، جس سے اس کی دل شکنی اور دل آزاری ہو یا اسے تکلیف پہنچے، جس غیبت کو قرآنِ کریم نے حرام قرار دے کر اسے مردہ بھائی کا گوشت کھانے سے تعبیر کیا ہے، وہ یہی غیبت ہے۔