

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 014



Thesis On

**DOCUMENT IMAGE ANALYSIS AND
RETRIEVAL SYSTEM**

Submitted in Partial Fulfillment of the Requirements of the award of the Degree of

Doctor of Philosophy

in

Computer Science and Engineering

Submitted by

Mr. UMESH

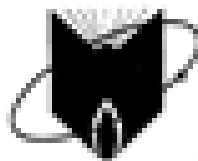
USN: 2BL13PEN01

Under the Guidance/Supervision of

Dr. M. S. Shirdhonkar

Professor, Department of Computer Science and Engineering

Research Centre



Department of Computer Science and Engineering,

**B. L. D. E. A's Vachana Pitamaha Dr. P. G. Halakatti College of
Engineering and Technology, Vijayapur-586 103, Karnataka, India.**

DECLARATION

This is to declare that the thesis work entitled “**Document Image Analysis and Retrieval System**” is carried out by me, **Mr. UMESH**, USN: **2BL13PEN01** a bonafide student of the Research Centre, Department of Computer Science and Engineering, **B.L.D.E.A’s V. P. Dr. P. G. Halakatti College of Engineering & Technology, Vijayapur, Karnataka, India** in partial fulfillment for the award of **DOCTOR OF PHILOSOPHY** in **Computer Science and Engineering** of **Visvesvaraya Technological University, Belagavi**, during the year **2014-19**. To the best of my knowledge and belief, the work reported in the thesis has not been submitted by me for the award of the any degree and is not the repetition of the work carried out by others.

Place: Vijayapur

Date: 13/9/2019



Mr. UMESH

Research Scholar,

Dept. of Computer Science and Engineering,
B.L.D.E.A’s V. P. Dr. P. G. Halakatti College of
Engineering & Technology, Vijayapur, Karnataka,
India.

B.L.D.E.A's



Vachana Pitamaha Dr. P. G. Halakatti College of Engineering and Technology,
Vijayapur-586 103, Karnataka, India.

CERTIFICATE

This is to certify that the thesis work entitled “**Document Image Analysis and Retrieval System**” carried out by, **Mr. UMESH (USN: 2BL13PEN01)** a bonafide student of Research Centre, Department of Computer Science and Engineering, **B.L.D.E.A's V. P. Dr. P. G. Halakatti College of Engineering & Technology, Vijayapur, Karnataka, India** in partial fulfillment for the award of **DOCTOR OF PHILOSOPHY in Computer Science and Engineering** of the **Visvesvaraya Technological University, Belagavi** during the year **2014-19**. To the best of my knowledge, the work reported in the thesis has not been submitted for the award of any degree or diploma and is not the repetition of the work carried out by others. In my opinion, this thesis is of the standard required for the award of the degree of Ph.D.

Dr. M. S. Shirdhonkar

Research Guide/Supervisor,

Professor, Department of Computer Science and Engineering,
B.L.D.E.A's V.P.Dr.P.G.Halakatti College of Engineering & Technology,
Vijayapur – 586103, Karnataka, India

Forwarded by:

Dr. (Smt.) Pushpa B. Patil
Head of Department,
Computer Science & Engg
B.L.D.E.A's V.P.Dr. P.G.H.C.E.T.
VIJAYAPUR

DR. V. P. HUGGI
PRINCIPAL
B.L.D.E.A's V.P. Dr.P.G.H
College of Engg. & Tech.
Bijapur,

ACKNOWLEDGEMENTS

It's my great pleasure to acknowledge all the people who made this Ph.D. thesis possible. I shall begin with almighty God. His Blessings are with me throughout my life and ever more in this research work.

I express my deep sense of gratitude to my research supervisor, **Dr. M. S. Shirdhonkar** Professor, Department of Computer Science and Engineering, B.L.D.E.A's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur, for his kind support and guidance. His dedication, encouragement and meticulous approach to research motivated me to put in the best of my efforts. He has been the pillar of strength, ensuring that I was never alone in my efforts.

I would like to express my sincere thanks to **Dr. (Smt.) Pushpa B. Patil**, Professor and Head, Department of Computer Science and Engineering, for providing great support and encouragement. My special thanks to **Dr. Prakash H. Unki** and **Dr. Sanjeevkumar M. Hatture** for their valuable suggestions. I sincerely acknowledge the entire faculty and staff of Computer Science and Engineering department for their support and help.

I express my sincere gratitude to the **BLDEA Management** for providing moral support and encouragement to carry out this research work. I take this opportunity to thank **Dr. V. P. Huggi**, Principal, B.L.D.E.A's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur, for extending all the facilities needed to carry out my research work. My thanks are due to **Dr. Prakash K. Gonnager** and **Dr. Pradeep V. Malaji** Vice-Principals, B.L.D.E.A's V. P. Dr. P. G. Halakatti College of Engineering and Technology, Vijayapur, for their continuous support and encouragement in this work. I also thank **Dr. R. S. Malladi**, Research coordinator of the institute for his suggestions and encouragement.

I sincerely thank **Prof. S. R. Purohit**, Head of Electronics and Communication Engineering department, of B.L.D.E.A's V. P. Dr. P. G. H. College of Engineering & Technology, for his motivation and support. I also thank all my colleagues for their endeavor support.

I feel a deep sense of gratitude for my parents **Shri. Dattatreya Dixit** and **Smt. Shobhabai Dixit** for their tremendous support, constant encouragement and great guidance in building

my life. I am very much grateful to my wife **Smt. Ashwini Dixit**, for her love, patience and constant support during the Ph.D work. I am thankful to my children **Shridatta** and **Akshay** for their love and support. I thank my entire family for moral support and encouragement.

Since it is not possible to bring in all the names, though I wish to do so, I place on record my sincere gratitude to each one of them who helped directly or indirectly in completing this thesis work.

Mr. UMESH

ABSTRACT

The drastic development in information technology has a lead for the digitization of documents in each and every field. During the digitization process, many of the existing and newly generated documents will be stored in the form of images known as document images. This has created an opportunity for the researchers to develop techniques and algorithms for analysis of document images for implementation of expert systems. Huge database of document images also require the techniques for accessing, searching and browsing of the documents based on certain criteria. The main objective of the proposed research work is to develop new techniques and algorithms for document image analysis and retrieval. The proposed research work is motivated by the Google search engine which allows text-based searching of information from the World Wide Web. The thesis provides new algorithms, techniques and feature extraction schemes for implementation of document image analysis and retrieval system. Five document retrieval techniques based on the logo, signature, face/photo, fingerprint and language are addressed in the thesis.

The thesis first provides an introduction to the document image analysis and retrieval system, its applications and a detailed literature survey. The literature survey discusses about the current state of the art and research trends. An efficient automatic logo-based retrieval technique is proposed by using mathematical tool Singular Value Decomposition (SVD). New feature extraction techniques based on singular value decomposition are used for logo-based document image retrieval. The proposed method is tested on the document images of Tobacco 800 database [41]. The experimental results are better compared to the earlier approach [47].

The thesis also provides signature-based document image retrieval method using multi-level Discrete Wavelet Transform (DWT). To investigate the influence of similarity metrics on retrieval performance, experiments are carried out using seven distance metrics namely Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard. The city block distance provided a highest precision of 80% using multi-level DWT features.

Recently many documents such as identity cards, passports, driving license etc. are embedded with face/photo of a person. For retrieval of such documents, the thesis proposes face/photo based document image retrieval using Gray Level Cooccurrence Matrix based features. In this approach, the cooccurrence matrices for Red, Green and Blue components of the face image

are computed and their diagonal elements are used to construct the feature vector. This helps in reducing the number of features and computational complexity. Experiments are carried out on a database of 810 document images. The proposed method provided a mean average precision of 82.66%.

To provide high security fingerprint impression is being used in many important documents such as property registration, banking transactions, insurance related documents etc. This motivated us to propose a fingerprint-based document image retrieval technique using multi-resolution Local Binary Pattern (LBP) features in the thesis. The proposed method is tested on 1200 document images having fingerprint impression. A mean average precision of 73.08% is obtained for retrieval of top 1, top 5, top 8, top 15 and top 20 document images. The experimental results are encouraging compared to the existing feature extraction techniques [38] and [39].

This thesis also presents a new approach for classification and retrieval of document images based on the language. The multi-resolution Histogram of Oriented Gradients (HOG) features are proposed for the implementation. The system is tested for printed document images of Kannada, Marathi, Telugu, Hindi and English. Proposed system provided classification accuracy of 87.02% for 1006 document images. The retrieval performance obtained using proposed feature extraction scheme is very promising and encouraging. The thesis also provides future directions for the research to elevate current state-of-the-art.

CONTENTS

Declaration	ii
Certificate	iii
Acknowledgements	iv
Abstract	vi
List of Figures	xii
List of Tables	xiv
List of Abbreviations and Symbols	xv
1. Introduction	
1.1 Introduction	1
1.2 Motivation	2
1.3 Architecture of Document Image Analysis	3
1.3.1 Preprocessing	3
1.3.2 Feature Extraction	4
1.3.3 Text level Analysis and Recognition	4
1.3.4 Graphics level Analysis and Recognition	4
1.3.5 Document Description	4
1.4 Architecture of Document Image Retrieval	5
1.4.1 Preprocessing	5
1.4.2 Feature Extraction	6
1.4.3 Similarity Matching	6
1.4.4 Ranking of the Documents	6
1.5 Classification of Document Image Retrieval	6
1.6 Literature Review	7
1.7 Problem Definition	10
1.8 Challenges in Design and Implementation	10
1.9 Evaluation Strategies	11
1.10 Summary of Contributions	12
1.11 Organization of Thesis	14
1.12 Summary and Conclusion	14
2. Automatic Logo-based Document Image Retrieval	
2.1 Introduction	16

2.2	Problem Statement	17
2.3	Related Work	17
2.4	Proposed Logo Detection Method	19
2.4.1	Preprocessing	20
2.4.2	Area-based Thresholding	20
2.4.3	Singular Value Decomposition	20
2.4.4	Logo Detection and Extraction	21
2.5	Proposed Logo-based Document Image Retrieval	21
2.5.1	Feature Computation of Detected Logo	21
2.5.2	Logo Matching	23
2.5.3	Retrieval of Documents	23
2.6	Experimental Results	24
2.6.1	Image Database	24
2.6.2	Performance of Proposed Logo Detection Method	24
2.6.3	Performance of Logo-based Document Retrieval	26
2.7	Summary and Conclusion	29

3. Signature-based Document Image Retrieval

3.1	Introduction	30
3.2	Problem Statement	31
3.3	Related Work	31
3.4	Proposed Methodology	33
3.5	Signature Detection and Extraction	33
3.5.1	Preprocessing	34
3.5.2	Finding Probable Signature Candidates	34
3.5.3	Feature Extraction	34
3.5.4	Signature Detection and Extraction	38
3.6	Signature-based Document Retrieval	39
3.6.1	Feature Matching	39
3.6.2	Ranking of the Documents	40
3.7	Experimental Results	40
3.7.1	Image Database	40
3.7.2	Performance of Signature Detection and Extraction	41
3.7.3	Performance of Signature-based Retrieval	41
3.8	Summary and Conclusion	45

4. Face-based Document Image Retrieval	
4.1 Introduction	46
4.2 Problem Statement	47
4.3 Related Work	47
4.4 Proposed Face/Photo based Document Retrieval System	48
4.4.1 Face Image Detection	49
4.4.2 SVD based Feature Extraction	50
4.4.3 GLCM based Feature Extraction	52
4.4.4 Document Retrieval	54
4.5 Experimental Results	55
4.5.1 Image Database	55
4.5.2 Performance of Face-based Document Retrieval	55
4.6 Summary and Conclusion	58
5. Fingerprint-based Document Image Retrieval	
5.1 Introduction	59
5.2 Problem Statement	60
5.3 Related Work	60
5.4 Proposed Method for Fingerprint-based Document Retrieval	62
5.4.1 Fingerprint Detection	63
5.4.2 Computing Feature Vector for Document Retrieval	65
5.4.3 Fingerprint Matching and Document Retrieval	68
5.5 Experimental Results	69
5.5.1 Image Database	70
5.5.2 Performance of Signature Detection and Extraction	70
5.5.3 Performance of Signature-based Retrieval	71
5.6 Summary and Conclusion	73
6. Language-based Document Image Classification and Retrieval	
6.1 Introduction	75
6.2 Problem Statement	79
6.3 Related Work	79
6.4 Proposed Language-based Classification System	80
6.4.1 Preprocessing	81
6.4.2 SWT based Multi-resolution HOG Feature Extraction	82
6.4.3 SVM Classifier	85

6.5	Experimental Results of Language-based Classification	85
6.5.1	Image Database	85
6.5.2	Performance of Language-based Classification System	85
6.6	Proposed Language-based Document Image Retrieval System	88
6.6.1	Feature Extraction	88
6.6.2	Similarity Matching	90
6.7	Experimental Results of Language-based Document Retrieval	91
6.7.1	Image Database	91
6.7.2	Performance of Language-based Retrieval System	91
6.8	Summary and Conclusion	94
7.	Conclusions and Future Dircetions	
7.1	Conclusions	95
7.2	Future Directions	97
	Author's Publications	98
	Bibliography	100

LIST OF FIGURES

Figure Number	Figure Caption	Page Number
Fig. 1.1	Sample Document Images	2
Fig. 1.2	Architecture of Document Image Analysis	3
Fig. 1.3	Architecture of Document Image Retrieval	5
Fig. 1.4	Classification of DIRS	7
Fig. 1.5	Pictorial Representation of Proposed Research Work	12
Fig. 2.1	Graphical, Text and Mixed logo examples	16
Fig. 2.2	Architecture of the Proposed Logo-based Document Retrieval System	21
Fig. 2.3	Sample Logo Detection Result	25
Fig. 2.4	Comparison of Logo Detection Results	26
Fig. 2.5	Sample Result of Logo-based Document Retrieval	28
Fig. 2.6	Graphical Comparison of Logo-based Document Retrieval Results	29
Fig. 3.1	Plot of a Mother Wavelet	35
Fig. 3.2	DWT Decomposition of an Image	36
Fig. 3.3	Two level DWT	37
Fig. 3.4	Signature Extraction Results	41
Fig. 3.5	Sample Result of Signature-based Retrieval	42
Fig. 3.6	Graphical Comparison of Results	44
Fig. 4.1	Proposed Face/Photo-based Document Retrieval System	49
Fig. 4.2	Sample Result of Face/Photo Extraction	50
Fig. 4.3	GLCM of an Image 'I'	52
Fig. 4.4	Sample Document Retrieval Result	56
Fig. 4.5	Graphical Comparison of Average Precision	57
Fig. 4.6	Comparison of F-measure	58
Fig. 5.1	Proposed Architecture of Fingerprint-based Document Retrieval	63
Fig. 5.2	Fingerprint Detection System	63
Fig. 5.3	Classification Example using SVM	65
Fig. 5.4	Proposed Feature Extraction Scheme(s)	66
Fig. 5.5	Sample Results of Fingerprint Detection	70
Fig. 5.6	Sample Result of Fingerprint-based Document Retrieval	72

Fig. 5.7	Comparison of Average Recall	73
Fig. 6.1	Taxonomy of Language/Script Identification	76
Fig. 6.2	Sample Documents of Kannada, Marathi, Telugu, Hindi and English	76
Fig. 6.3	Vowels and Consonants of Kannada, Telugu, Marathi, Hindi and English	78
Fig. 6.4	Proposed Language-based Document Classification System	81
Fig. 6.5	Un-sharp Mask	82
Fig. 6.6	SWT based Multi-resolution HOG Features of a Sample Document Image	84
Fig. 6.7	Graphical Comparison of Results	87
Fig. 6.8	Building Blocks of Proposed Language-based Document Retrieval System	88
Fig. 6.9	Proposed DWT based Multi-resolution HOG Feature Extraction	89
Fig. 6.10	Sample Result of Kannada Document Retrieval	92
Fig. 6.11	Graphical Comparison of Results for Dataset1	93
Fig. 6.12	Graphical Comparison of Results for Dataset2	93
Fig. 6.13	Graphical Comparison of Results for Dataset3	94

LIST OF TABLES

Table Number	Table Caption	Page Number
Table 1.1	Different Features used for DIA	4
Table 2.1	Logo Detection Results	25
Table 2.2	Experimental Results	28
Table 2.3	Results with N/2 and N/4 Singular Values	29
Table 3.1	Different Distance Metrics	40
Table 3.2	Recall and Precision results using Single-level DWT	43
Table 3.3	Recall and Precision results using Multi-level DWT	43
Table 4.1	Average Precision and Average Recall	56
Table 4.2	Comparison of F-measure	57
Table 5.1	Average Precision and Recall	72
Table 6.1	Details of the Database	85
Table 6.2	Details of Feature Extraction Schemes	86
Table 6.3	Comparison of Results using K-NN Classifier	86
Table 6.4	Comparison of Results using SVM Classifier	87
Table 6.5	Details of 3 Datasets	91
Table 6.6	Average Precision for Dataset1	92
Table 6.7	Average Precision for Dataset2	93
Table 6.8	Average Precision for Dataset3	93

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviation/ Symbol	Description
DIA	Document Image Analysis
DIARS	Document Image Analysis and Retrieval System
DIR	Document Image Retrieval
OCR	Optical Character Recognition
SVM	Support Vector Machine
GSC	Gradient, Structural and Concavity
CRF	Conditional Random Field
DTW	Dynamic Text Warping
DWT	Discrete Wavelet Transform
GLCM	Gray Level Co-occurrence Matrix
LBP	Local Binary Pattern
HOG	Histogram of Oriented Gradients
DIRS	Document Image Retrieval System
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
RCWF	Rotated Complex Wavelet Filters
DT-CWT	Dual-Tree Complex Wavelet Transform
QPM	Query Point Movement
BoW	Bag of words
P	Precision
R	Recall
AP	Average Precision
AR	Average Recall
MAP	Mean Average Precision
SVD	Singular Value Decomposition
KNN	K Nearest Neighbor
LBDIR	Logo-based Document Image Retrieval
E	Energy
SD	Standard Deviation

FV	Feature Vector
FVQ	Query Feature Vector
FDB	Feature Data Base
U and S	Unitary Orthogonal Matrices
S	Singular Matrix
GHT	Generalized Hough Transform
TPS	Thin-Plate Spline
CA	Approximated coefficients
CH	Horizontal coefficients
CV	Vertical coefficients
CD	Diagonal coefficients
FHLA	Fuzzy-based Hybrid Learning Algorithm
RBFNN	Radial Basis Function Neural Network
DCT	Discrete Cosine Transform
DWDPA	Dynamic Weighted Discrimination Power Analysis
PAN	Permanent Account Number
SWT	Stationary Wavelet Transform
FFM	Fuzzy Feature Match
LDP	Local Directional Pattern
LTDF-MLDN	Local Texture Description Framework based Modified Local Directional Pattern
ELM	Extreme Learning Machine
FDCT	Fast Discrete Curvelet Transform
PNN	Probabilistic Neural Network
HMM	Hidden Markov Model
MLP	Multi-Layer Perception

Chapter 1

Introduction

Abstract of the Chapter: Document image analysis and retrieval is a hot area of research due to its wide applications in various fields. This chapter provides a brief introduction to document image analysis and retrieval system. It describes the generic steps practiced for analysis and retrieval of document images. This chapter also discusses the state of art techniques, challenges and the major contributions of the proposed research work.

1.1 Introduction

The developments in computing technologies lead the digitization of a huge number of documents. Most of the documents during digitization are stored in the form of images. To reduce the effort, cost and time involved in understanding these documents, a new field of research called document image processing was evolved during the 1990s with the aim of automizing the document image processing. The document image analysis and retrieval is a sub-field of document image processing developed with the goal to provide a solution towards a paperless office. It has drawn the attention of many scholars in the current decade due to rapid development in technology.

Generally, the scanned documents consisting of printed or handwritten text, symbols, tables, logos and other graphical components are referred to as document images. The scanner, digital camera and mobile camera are the sources of document images used to digitize the documents. Categorically, the document images may include certificates, marks-cards, historical documents, individual pages of the books, legal documents, official memorandums, notices and many more. The digitization of documents is being found in each and every organization due to the equipment available at a low cost. A huge number of document images, generated in day to day life, requires a systematic analysis for document image understanding. The document image analysis deals with the detection of textual and graphical components of the document [1]. However, document image retrieval allows accessing, browsing and searching of the document images based on certain attributes from the huge database [2]. Fig. 1.1 shows sample document images.

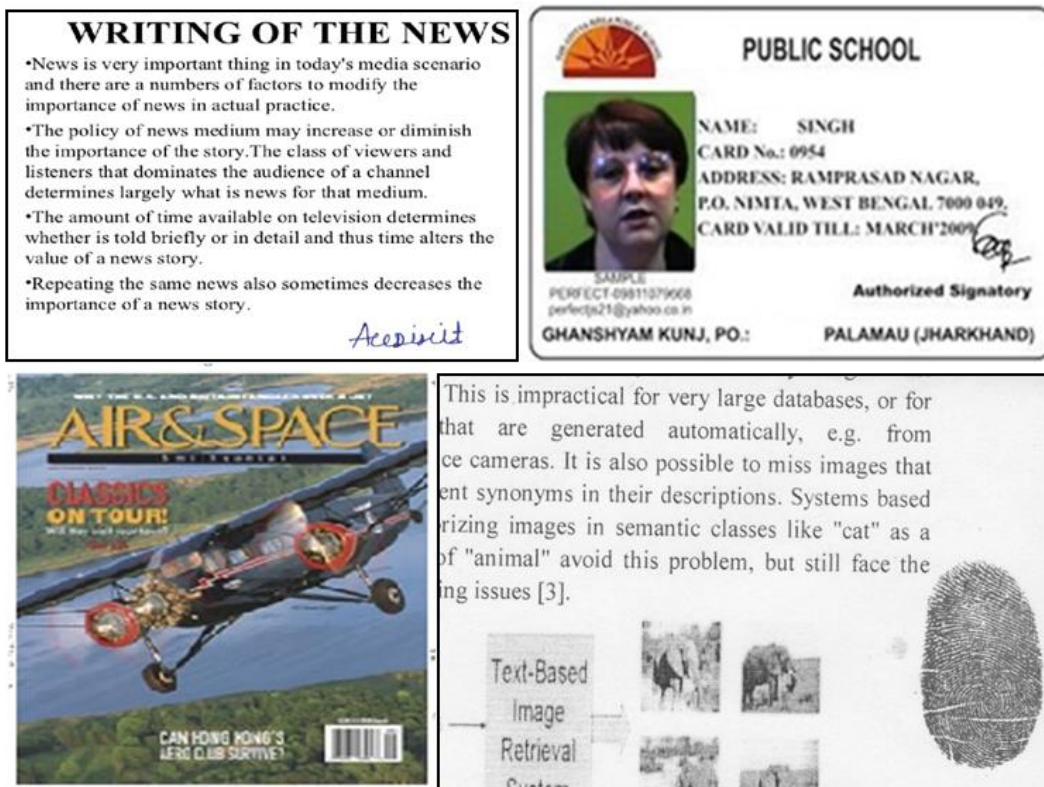


Fig. 1.1 Sample Document Images

Some of the important applications of document image analysis and retrieval system are:

1. Extraction of the Name, Address, Location and Pin code from the emails and postal document images. Automatic script identification from both handwritten and printed text document images.
2. Recognition of vehicle number from number plates for implementation of computerized registration and tracking of vehicles.
3. Searching of the documents based on the title for automation of digital library and reading of specific topics from e-books.
4. Signature-based retrieval of document images in the sectors like banking, insurance and business offices.
5. Logo based document retrieval in small offices and organizations for quick accessing of the document images.

1.2 Motivation

A drastic increase in document images due to cheaper technology has created a challenge for proper analysis and quick retrieval of documents with accurate results. The motivation behind this research work is the Google search engine which is being used for many years. Google provides text-based searching of data from the World Wide Web. A similar

kind of search engine to retrieve the document images based on certain attributes or components of the query is highly required in many organizations such as business offices, banking, insurance, digital library, crime branches, etc. This has motivated our research work to improve the performance of the existing document image analysis and retrieval techniques along with proposing some new retrieval techniques that suite the recent requirements in the domain.

1.3 Architecture of Document Image Analysis

The Fig.1.2 shows the architecture of Document Image Analysis (DIA). Preprocessing, feature extraction, analysis of text and graphical components are the building blocks of the architecture. The document description is an outcome of the analysis. These blocks are briefly discussed in the following sub-sections.

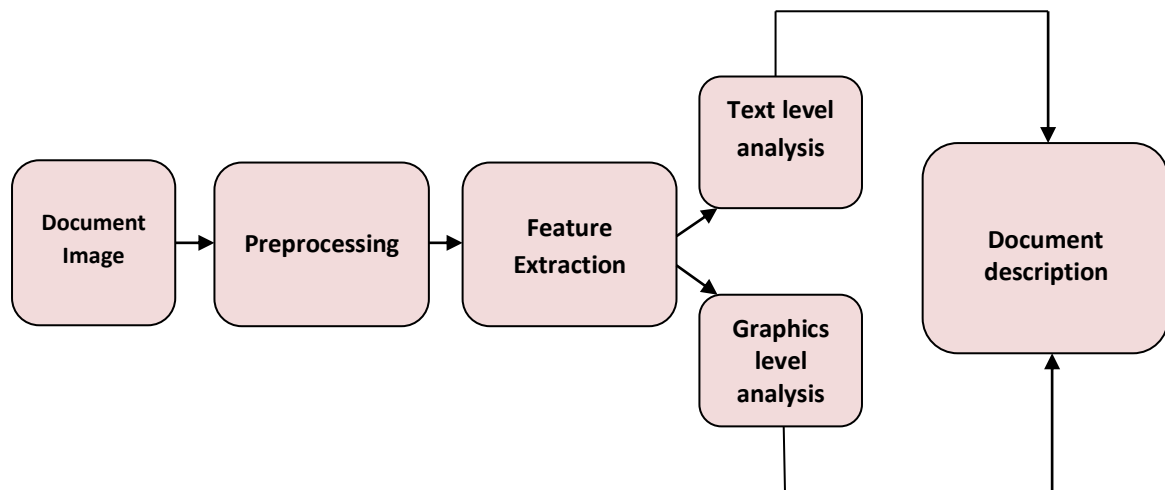


Fig. 1.2 Architecture of Document Image Analysis

1.3.1 Preprocessing

The document images in their original form cannot be used directly for analysis. The preprocessing step prepares the document for extracting suitable features depending on the type of the application. The preprocessing steps may include color to gray-scale conversion, binarization, noise removal, segmentation, etc. Most of the analysis techniques use the gray-scale image or binary image to reduce the number of features. Documents suffer from different types of noises such as impulse noise, duty noise, clutter noise, etc. The noise present in the documents may lead to inaccurate results during document image analysis. Hence suitable filters, techniques or algorithms are used in the preprocessing step for removal of noise. Depending on the application, document

segmentation is also used as part of the preprocessing. In the proposed research work, color to gray-scale, gray to binary conversions along with noise removal algorithms such as median filter are used.

1.3.2 Feature Extraction

This step is used to gather meaningful information from the document images. The local or global features may be extracted during this step. The different features used in the existing system are listed in Table 1.1.

Table 1.1 Different Features used for DIA

Image features	Structural features	Textual features
1.Connected components.	1. Physical Layout.	1. Page/Text layout features.
2.Gaps between row/Columns.	2. Logical structures	2. Textual features from
3.Location and size of cells.	3. Results of functional labeling	OCR results analysis
4.Text histogram.	4. Spatial relations	

The proposed research work employs the properties of the connected components, texture features and the visual features for analysis of document images.

1.3.3 Text Level Analysis and Recognition

Two main types of analysis such as Optical Character Recognition (OCR) and page-layout analysis are applied to the text. The purpose of OCR is to recognize the text for deriving the meaning of characters or words. However, page-layout analysis is used to detect and recognize functional blocks such as words, titles, subtitles, bodies of text, footnotes, etc. The proposed research uses hybrid texture features for text recognition to classify the documents based on the language.

1.3.4 Graphics Level Analysis and Recognition

Graphics level analysis is intended to recognize graphical components of the document image such as lines, curves, logos, photos, signatures, etc. Graphical analysis demands the use of structural, shape and geometrical features. The proposed work mainly employs the connected component analysis for detection of graphical components and their features with Support Vector Machine (SVM) classifier for recognition.

1.3.5 Document Description

A document description is an outcome of the document image analysis. The document description includes the details of both textual and graphical components of the document

image. The description may consist of the number of characters, words, lines, paragraphs, logos, signatures, etc.

1.4 Architecture of Document Image Retrieval

The Document Image Retrieval (DIR) is used to access, browse or search the documents based on certain attributes such as the title, logo, signature, layout, etc. Fig. 1.3 depicts the architecture of DIR with the important building blocks such as preprocessing, feature extraction, similarity matching, ranking and retrieval. These blocks are described in the below sections.

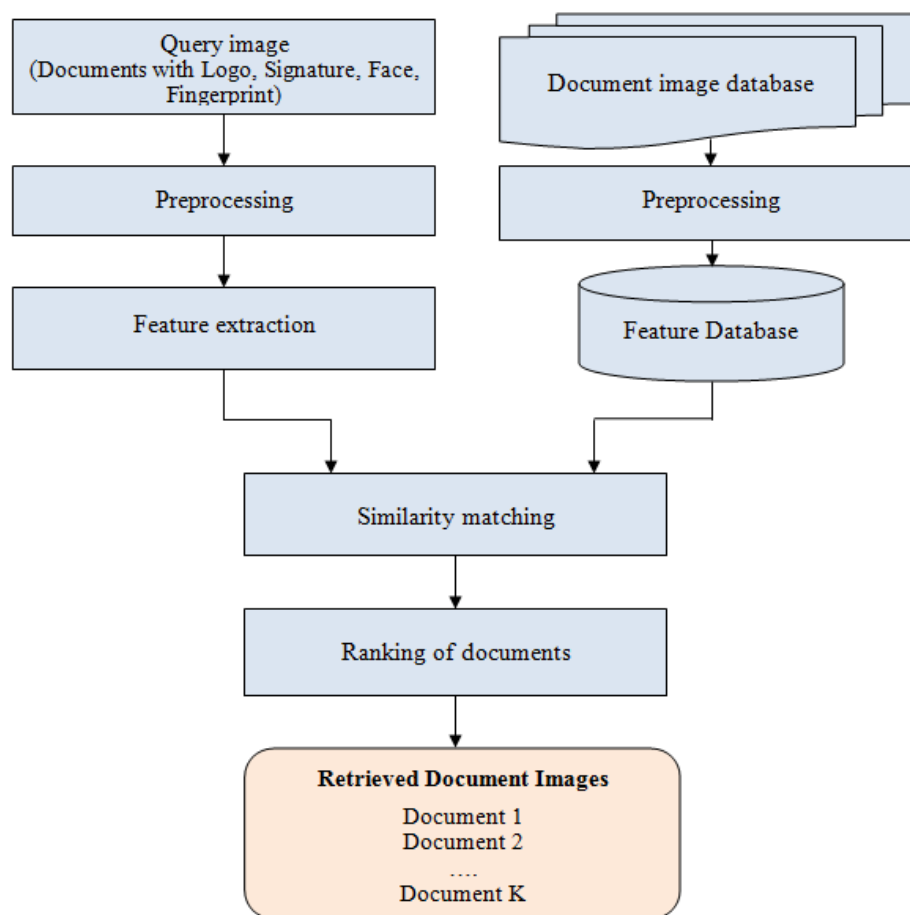


Fig. 1.3 Architecture of Document Image Retrieval

1.4.1 Preprocessing

The preprocessing step in document image retrieval process is much similar to that of preprocessing used in document image analysis. But it may also include segmentation, detection or extraction of a particular component from the query, depending on which retrieval is to be carried out. For example, in logo-based document retrieval, the logo segmentation may be part of preprocessing.

1.4.2 Feature Extraction

Meaningful information is extracted from the query image at this stage. The features from both spatial-domain and frequency-domain can be used for retrieval. The popular feature extraction schemes include visual, geometric, shape and texture-based features or their combination. Various techniques are used to extract the features such as Gradient, Structural and Concavity (GSC) features [3-4], that measure the image characteristics at local, intermediate and large scale. Density distribution and key block features [5], Fisher classifier [2], Conditional Random Field [6], Dynamic Text Warping (DTW) [7] are used in the literature. In the proposed research work, a new set of features by extending the concept of DWT, Gray Level Co-occurrence Matrix (GLCM), Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) are used.

1.4.3 Similarity Matching

This step is used to find the extent of similarity between the query document image and the document images stored in the database. For similarity measure, the query image feature vector and database image feature vectors are compared using the distance metrics [8]. Various distance metrics such as Euclidean, city block, Manhattan, correlation, hamming, etc., can be used. The comparison of different metrics is provided in [9]. In the proposed research, standard Euclidean distance, Canberra distance and Mahalanobis are used for matching of the documents. We also provided a comparison of results using different similarity metrics.

1.4.4 Ranking of the Documents

This step is used to identify the closest document to the query and other documents nearest to the query. The distance measures used in the similarity matching provide a metric having value between 0 and 1. The smallest distance value indicates closest match and vice versa. Ranking of documents includes ordering of the documents to be retrieved based on the distance score. Thus the result of this step is ranked set of retrieved documents.

1.5 Classification of Document Image Retrieval System

Reza Tavoli [10] classified the document image retrieval methods into two categories, viz. traditional methods and new methods. Fig. 1.4 shows a detailed classification of Document Image Retrieval System (DIRS). Traditional methods include searching for documents based on keywords and titles. It also aims at the indexing of the documents

based on Optical Character Recognition (OCR). The traditional method of indexing finds its application in the text to speech conversion, libraries and small organizations. The new methods of document retrieval are evolved as the outcome of research carried out in the last decade. This method includes retrieval of documents based on the signature, logo and layout. The proposed research work has added new techniques such as face-based, fingerprint-based and language-based document retrieval. These techniques are shown with shaded blocks in Fig. 1.4.

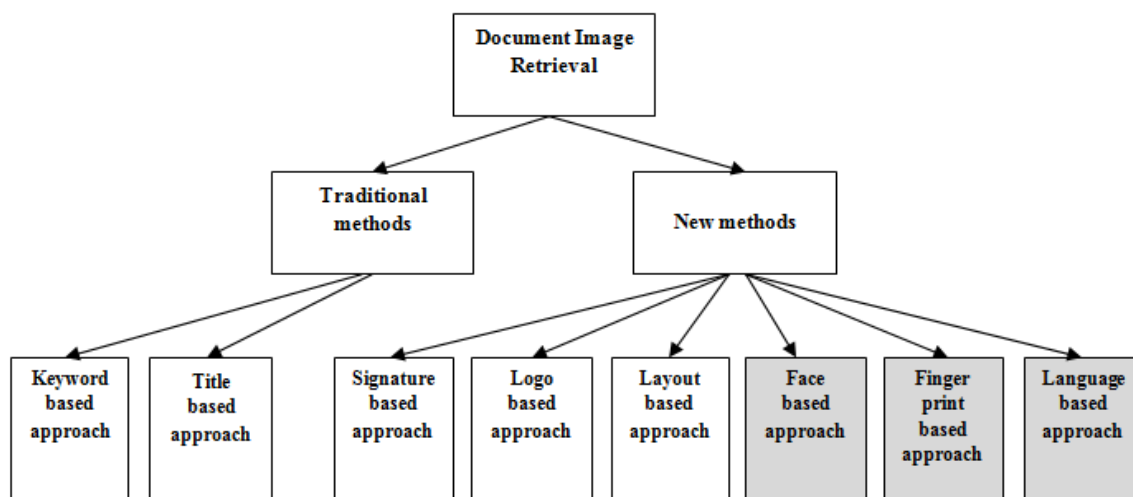


Fig. 1.4 Classification of DIRS

1.6 State of the art Techniques

A lot of methods and techniques has been proposed in the area of document image analysis and retrieval during last two decades. This section briefly describes and highlights the work done in this area by the researchers.

Acquiring knowledge or information from millions of documents generated requires manpower, time and money. To provide a solution to this, Tang et al. [11] developed an automatic knowledge acquisition system. They used geometric and logical structure analysis of documents in the proposed technique. The geometric structure was derived through entropy analysis and is mapped to logical structure for document image understanding. To reduce the effort of information retrieval from digital library Niyogi and Srihari [12] presented a method that employed layout analysis and document structure. The proposed system enabled users to refine their queries from retrieved documents to obtain more relevant information. Liu and Jain [13] proposed an image-based form retrieval using similarity measure. The proposed similarity measure was robust to variation in image attributes. Document image retrieval for Chinese documents

is presented in [14] using the stroke density of characters. The method helps for fast retrieving of Chinese documents with a limitation of few fonts. Lu and Tan [15] developed a technique for information retrieval from digital libraries without using OCR. They used different word shape coding techniques to convert a word image into a shape code. However, the technique was limited to linguistic knowledge.

Liu et al. [16] employed density distribution and key block features for retrieval of document images. The key block features in their presented method helped for improved retrieval performance. This method was suitable for large scale document images with different languages. Nakai et al. [17] presented a technique for document images acquired from a digital camera. Geometric invariant based indexing and voting with hash tables are used in their method to improve accuracy. Jawahar et al. [8] developed an architecture to retrieve relevant documents from the huge database. They used Dynamic Time Wrapping (DTW) based features for word-level matching. Schomaker [18] presented a system to retrieve handwritten lines of text from the historical documents. They used images of line strips for matching with the help of cross-correlation instead of image features. Joutel et al. [19] proposed a classification technique for identifying writers of ancient document images. They used curvelet-based features with two discriminative properties curvature and orientation in the proposed scheme. The proposed technique has an advantage of language independence.

Lu et al. [20] used word-shape coding to retrieve document images. The word-shape coding was obtained by capturing each word image, which is annotated by a group of topological shape features that included character ascenders/descenders, the hole between the characters and also character water reservoirs. Their method was fast and was able to handle various types of document degradation. Vikram et al. [21] presented an algorithm for retrieving person-specific document images based on the face. To achieve this, the documents are normalized in size and tagged with the average of face images. The proposed Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) subspace-based method for recognition.

Hassan et al. [22] presented a method for indexing the documents using shape descriptors with hierarchical distance-based hashing. Relatively organized points on the boundary of an object are used to construct the shape descriptor and a novel hierarchical distance-based hashing is proposed for indexing of the documents. Li et al. [23] used a sequence of local features to retrieve the documents independent of language. The local features

that represent intrinsic and layout characteristics are proposed in their work. They computed word length by using a number of pixels and the sequence of the length of words in the document are used for matching and retrieval of documents. Latin based documents are used to evaluate the performance of their method.

Shirdhonkar and Kokare [24] presented a document image retrieval technique where a signature is used as the query. Rotated Complex Wavelet Filters (RCWF) and Dual-Tree Complex Wavelet Transform (DT-CWT) together are used for feature extraction. The Canberra distance is employed for matching the features. Further to improve the relevance and performance they proposed Query Point Movement (QPM) in their work. An algorithm for writer based document retrieval is proposed by Shirdhonkar and Kokare [25]. They compared the retrieval results using two distance measures Canberra and Euclidean. Canberra distance provided good results compared to Euclidean in their work.

Keyvanpour and Tavoli [26] developed a feature weighted technique to improve the performance of document image retrieval. “Feature weighting is a technique used to approximate the optimal degree of influence of individual features”. This method weights the feature using the coefficient of multiple correlations. Pirlo et al [27] used dynamic time warping to implement layout based document retrieval. Grid-based structural elements are extracted by using morphological operations in their proposed technique. The Random transform was employed to obtain the description of layouts and dynamic time warping for indexing of the documents. Shekhar and Jawahar [28] motivated by sparsity in signal representation and developed word retrieving algorithm using document specific sparse coding. The document images are matched with the help of Bag of words (BoW) technique. Vaizadeh and Kabir [29] proposed a method for document image binarization using an adaptive water flow model. The method was intended for degraded document images. Each blob in the document was classified as a textual part or non-textual part by employing multilayer perceptron, which preserved the connectivity between the strokes.

Cote and Albu [30] presented a method to classify whether each pixel of the document image belongs to one of the four categories: text, image, graphics and background. They used SVM for classification of pixels distributed across the document. An interactive method for transcription of handwritten text is developed by Serrano et al. [31] to reduce the effort of users. Sankar et al. [32] proposed word annotation for document images. The proposed scheme replaces native classification by an intelligent combination of indexing

and classification. An edge noise removal technique for binary images is presented by Hoang et al. [33]. The Optimal precision parameter employed in their work shows that there exists a linear relationship between the binary level of the noise. Rusinol et al. [34] proposed multimodal page classification for the documents that are used in a banking application. The visual and texture features are merged to classify the documents used for administration. Hierarchical distribution of pixel intensities is used in the visual description and latent semantic analysis in the textual description.

From the literature, it is learned that

- The existing logo and signature-based document retrieval methods use the logo or signature component as a query. There is a need for the development of automatic logo/signature detection from the query to implement logo-based and signature-based document image retrieval with improved performance.
- In the modern age, many of the documents consist of the face or finger-print impression of a person. Such documents require novel document image retrieval techniques which shall be based on the face or fingerprint present in the documents.
- The globalization demands the acceptance of multi-lingual documents with different languages and scripts. This demands a need for language-based document image classification and retrieval algorithms.

Hence the objective of the proposed work includes analysis of documents and development of novel techniques to retrieve documents stored in the database

1.7 Problem Definition

Development of novel and efficient techniques for searching the documents of interest by using a query (for instances of the logo, signature, face, latent fingerprint, language) document image.

1.8 Challenges in Design and Implementation

Following are the major challenges in implementing the successful document image analysis and retrieval algorithms.

-
- **Degradation of documents:** The document images are degraded due to the reasons such as excessive duty noise, ink blobs, poor quality of the paper and ink. This causes the inaccurate or reduced performance of document image analysis algorithms.
 - **Language dependency:** The character shapes, way of writing and the orientation makes document retrieval as a language-dependent issue.
 - **Standard datasets:** Only a handful of document image datasets are publicly available. These datasets include documents comprising logo, signature and different layouts. The available datasets are insufficient to develop new methods of document retrieval such as face-based, fingerprint-based and language-based.

The proposed research addresses the above challenges as follows.

1. Image enhancement techniques and noise reduction filters are used to address document degradation problem.
2. The proposed document retrieval techniques are language-independent because the retrieval is based on logo, signature, face and fingerprint.
3. During the proposed research work, three datasets were developed to evaluate and compare the results of new document retrieval methods.

1.9 Evaluation Strategies

Three important parameters precision, recall and F-measure are used for evaluation of the proposed document retrieval algorithms and techniques. These parameters are discussed below.

- **Precision:** It is defined as the “ratio of relevant documents retrieved to the total number of documents retrieved” and computed using equation (1.1).

$$Precision (P) = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}} \quad (1.1)$$

- **Recall:** The recall is the “ratio of the number of relevant documents retrieved to the total number of relevant documents available in the database”. It is computed using equation (1.2)

$$Recall (R) = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents available in the database}} \quad (1.2)$$

- **F-measure:** It gives the effective performance of the document retrieval system and computed using equation (1.3).

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1.3)$$

1.10 Summary of Contributions

Fig. 1.5 provides a pictorial representation of the proposed research work. The main contributions of this research are summarized below.

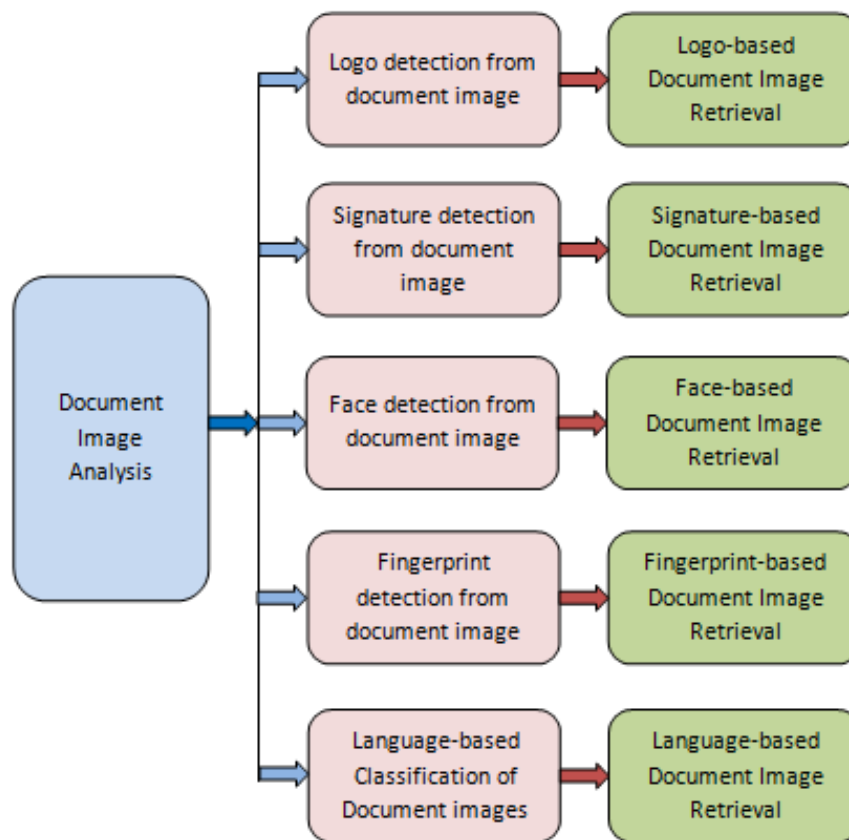


Fig.1.5 Pictorial Representation of Proposed Research Work

- Logo detection and logo-based document image retrieval:** The earlier approach for logo detection and retrieval are based on DWT features. To improve the performance of logo detection, the proposed method employed the properties of connected components. Two sets of features are proposed for logo-based document retrieval namely derived Singular Value Decomposition (SVD) features and singular values of the singular matrix. The results obtained with the proposed method outperform compared with the earlier approach [35].

-
- ii. **Signature detection and signature-based document image retrieval:** The similarity metrics used for matching and retrieval of the signature play a vital role and influence on the performance. Therefore the seven distance metrics namely, Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard are investigated for single and multi-level DWT based features in the proposed automatic signature-based document image retrieval. The combination of multi-level DWT based features with city block distance provided better performance.
 - iii. **Face/Photo detection and face-based document image retrieval:** The document images such as identity cards, Permanent Account Number (PAN) cards, passports, certificates include face/photo of a person. The approaches used for document image retrieval in the literature are based on the title, signature, logo and layout. Therefore to access or search the documents of interest a face/photo based document retrieval is proposed. Two sets of features based on SVD and GLCM are used in the proposed method. The GLCM based features provided better retrieval performance in comparison with [35], [36] and [37].
 - iv. **Fingerprint detection and finger-print based document image retrieval:** Recently the important documents in many organizations are embedded with fingerprint impression of the person for authentication due to high security. Property registration, letters related to legal issues, bank transaction records are some of the examples. To retrieve such documents of interest, an automatic fingerprint-based document image retrieval technique is proposed. The DWT based features of the connected components with SVM classifier is employed for detection of fingerprint impression from the document. Two sets of features namely DWT based LBP and SWT based LBP are proposed for fingerprint-based document retrieval. The proposed feature extraction schemes provided promising results compared to [38] and [39].
 - v. **Language-based classification and retrieval of document images:** The document images having content with different languages gives rise to the need for language-based document classification and retrieval techniques. The methods used for language/script identification are based on segmentation of the document images at word-level, line-level or block level. The word-level and line-level based classification techniques are computationally expensive. To reduce the time for classification and retrieval of the documents, the multi-resolution HOG

features with segmentation free technique is proposed. The proposed system is investigated on document images of Kannada, Marathi, Telugu, Hindi and English languages. The multi-resolution HOG features provided better results in comparison with [38], [39] and [40].

1.11 Organization of the Thesis

This thesis is organized into seven chapters.

- Chapter 1 provides an overview of document image analysis, motivation for the research, literature survey, problem definition, challenges and the major contributions.
- Chapter 2 describes the proposed automatic logo detection and logo-based document image retrieval technique using Singular Value Decomposition (SVD) based features.
- Chapter 3 gives an insight into the proposed signature-based document retrieval employing multi-level DWT features. It also provides a comparison of signature-based document retrieval results using different distance metrics.
- Chapter 4 discusses the proposed face/photo based document image retrieval methods employing SVD and GLCM based features.
- Chapter 5 presents the proposed fingerprint-based document image retrieval method using multi-resolution LBP features.
- Chapter 6 explains the proposed methods for language-based classification of document images and retrieval using multi-resolution HOG features.
- Chapter 7 concludes the overall research work and provides future directions for further research.

1.12 Summary and Conclusion

Document image analysis and retrieval is an essential application for accessing the documents from a huge database. Several methods such as title, logo, signature and layout based document retrieval techniques are proposed in the literature. There is scope for improving the performance of existing document retrieval techniques. Also, there is a

need for the development of new document retrieval techniques based on the face/photo, fingerprint and language. The performance of the approach used for logo detection and retrieval [35] can be improved by using new technique and set of features. This motivated us to develop an efficient automatic logo-based document image retrieval in the next chapter.

Chapter 2

Automatic Logo-based Document Image Retrieval

Abstract of the Chapter: The growth in technology has led to a drastic increase in document images in various sectors namely government organizations, banking, insurance companies, digital libraries, etc. Automatic logo-based document retrieval is an intelligent technique which allows retrieval of documents based on the logo detected in the query document. This chapter proposes a method for logo detection from the document image and retrieval of the relevant documents from the database using an algebraic mathematical tool Singular Value Decomposition (SVD). The proposed system is tested on the publicly available database Tobacco 800. An average logo detection of 90.06% is achieved with the proposed method. Two sets of features based on SVD are proposed for logo matching and logo-based retrieval of the documents. Various set of experiments are conducted for testing the proposed features for logo-based document retrieval. An average precision of 84% is obtained by using the proposed retrieval algorithm.

2.1 Introduction

The logo is a graphical symbol that includes some cryptic text, used to authenticate many of the documents by several organizations. The logos are classified as graphical logo, text logo and a mixed logo. As the name suggests, a graphical logo is designed entirely by using some typical graphics, a text logo includes the name of the company or organization with some specific design and a mixed logo is the combination of both graphics and the text. Fig. 2.1 shows a sample of the three varieties of logos that are available in Tobacco – 800 database [41].



Fig. 2.1 Graphical, Text and Mixed logo examples

The earlier methods proposed for logo detection provide poor logo detection and retrieval results. This motivated us to improve the performance by proposing an algorithm for automatic logo detection based on the energy possessed by the connected components. The main contribution of this chapter is proposing two sets of SVD based features for logo-based document image retrieval. The proposed logo detection and logo-based document retrieval systems have provided promising results compared with the existing method [35].

2.2 Problem Statement

The objective of the proposed system is:

- (1) Detecting and extracting a logo from the given query document and
- (2) To retrieve documents from the database those have a similar logo as that of the query.

Statement: Let $Doc = \{Doc_1, Doc_2, Doc_3, \dots, Doc_N\}$ is a set of 'N' documents and 'Doc_Q' is a query document. The objective is to extract the logo object 'L' from the query document and retrieve a set of 'K' documents given by $Doc_R = \{Doc_1, Doc_2, Doc_3, \dots, Doc_K\}$, that have the similar logo 'L' as that of query document 'D_Q'.

2.3 Related Work

Seiden et al. [42] developed a logo detection method by segmenting the document image into a small set of connected components. Later they used a set of sixteen region-based features to differentiate the segments as logo and non-logo segments. The presented method was evaluated on a set of 130 business documents. Their method suffered from detecting textual logos. Logo detection technique using the spatial density of the pixels is proposed by Pham [43]. This method divides the document image into small windows of fixed size. For each window, the spatial density of the pixels is computed and the window with higher spatial density is assumed to have logo part of the document. The method is tested using the UMD logo database having 105 logos with different conditions. The main advantage of this method was computationally cheap and practically simpler to implement.

Novel logo detection employing a multi-scale boosting strategy was presented by Zhu and Doermann [44]. At the initial level, the connected components with fisher classifier are used to detect a set of probable logo candidates. Then a cascaded set of classifiers are used for logo detection at the final level. They tested the method on realistic complex

document images of Tobacco 800 dataset. Logo detection and recognition system employing spatial and structural features were developed by Hassanzadeh and Pourghassem [45]. The goal of their work was to detect and combine separated parts of logos present in the document. Histogram of object occurrence-based feature is proposed for detecting the logo parts and a morphological dilation is employed for merging. Logo recognition is implemented by using K-NN classifier. The presented technique was investigated on a database of Maryland University.

Nejad and Faez [46] proposed logo extraction method with two steps. In the first step, the location of the logo is identified using horizontal and vertical analysis of pyramidal tree structure. Later for logo extraction, they applied the boundary extension of feature rectangle. The K-NN classifier is used for logo recognition. A DWT based logo detection technique was presented by Shirdhonkar and Kokare [35]. This method divides the document into smaller regions, whose size is almost the same as that of logos. For each smaller region, they compute the energy of wavelet coefficients and classify the region having the highest energy as logo part of the document image.

Wang and Chen [47] proposed a new method of logo detection using boundary extension of feature rectangle. “A feature rectangle is a minimum virtual rectangle which fully embraces at least one foreground pixel (black) with four edges consisting of all background pixels (white) and has minimum inner area”. With the assumption that the logos will have a white background, a seed pixel of 3×3 neighbor is defined. All the neighboring pixels from the seed are included as part of the logo, till white pixels are encountered. This approach will create a set of probable logo candidates, which are then fine-tuned for accurate logo detection using a decision tree. Zhu and Doermann [48] proposed logo based document retrieval. The results of logo detection proposed in [44] were improved using a cascade of classifiers. A two-dimensional shape context features are proposed for document retrieval. These features are matched using neighborhood graph matching for ranking the documents. Jain and Doermann [49] developed a logo retrieval approach using Speed Up Robust Features (SURF). An indexing scheme combining local features and geometrical constraints was presented for a huge sized database of documents.

Le et al. [50] presented a new technique of document retrieval based on logo spotting and recognition. Initially, they match the key points of query logo and the documents stored in the database in the Scale Invariant Feature Transform (SIFT) feature space. The logos are

segmented using spatial density-based approach and homography. The number of matched key points is used as a metric for ranking and retrieval of documents.

The main contribution of this chapter is presenting an efficient logo detection method using algebraic tool Singular Value Decomposition (SVD) [51]. The idea behind the proposed logo detection is that the connected component of the document contributing highest energy represents the logo part of the document. In the proposed method, the possible logo candidates are detected by selecting the connected components based on their area property. Later the energy of these logo candidates is computed using SVD and the candidate with maximum energy is treated as a logo. Two sets of features based on SVD are proposed to implement logo-based document retrieval namely (i) Derived SVD features and the (ii) Singular values obtained after SVD based decomposition.

2.4 Proposed Logo Detection Method

Algorithm 2.1 lists the sequence of steps used in the proposed logo detection technique. Preprocessing, area-based thresholding, singular value decomposition and logo detection are the important steps used in the algorithm. These steps are described in the following sections.

Algorithm 2.1: Logo detection and Extraction from Document Image using SVD

1. Begin
2. **Input:** Document image.
3. **Output:** Extracted logo from the document.
4. Read document image from database.
5. Find connected components and perform area based thresholding to get pool of possible logo candidates. Let $A[1,2, .. N]$ represents pool of possible logo candidates.
6. Apply SVD to each possible logo candidate A_i and decompose into matrices U_i , S_i and V_i . Where U_i , V_i are unitary orthogonal matrices and S_i is diagonal matrix.
7. Calculate energy of each possible candidate A_i using equation (2.1)
$$E_i = \frac{1}{M \times N} \left\{ \sum_{k=1}^M \sum_{l=1}^N [U_i(k, l) + S_i(k, l) + V_i(k, l)] \right\} \quad (2.1)$$

Where E_i is energy of each possible logo candidate A_i , M and N are size of matrices U_i , S_i and V_i .
8. Detected logo = Candidate from the pool with $\text{Max}(E_i)$.
9. Extract the detected logo from the document image.
10. End

2.4.1 Preprocessing

Initially, the given query document image is converted into binary format, where each pixel is associated with intensity value 1 and 0. The pixels with values '0' and '1' correspond to black and white-colored pixels respectively. Usually, the document images suffer from clutter noise which appears in the form of horizontal and vertical stripes at the edges. This type of noise may lead to inaccurate detection of logo candidates. Therefore to remove such noise, the image is scanned for a continuous string of 0s in horizontal and vertical directions. The strings with a length of more than 200 are replaced with background pixels. The length 200 is chosen empirically to avoid removal of lines which could be part of logos. After removal of horizontal and vertical strips at the edges, the image is passed through a simple median filter to eliminate the impulse noise.

2.4.2 Area-based Thresholding

The aim of this step is to find the possible set of logo candidates from the query document. In this step, the connected components whose area is more than 25% of the largest one are considered as the probable set of logo candidates. This step helps to reduce the processing of all the connected components and speed up the process of logo detection.

2.4.3 Singular Value Decomposition

The SVD is used to compute the energy of probable logo candidates obtained from the previous step. SVD is an algebraic technique used in several image processing algorithms. Applying SVD decomposes the image into independent components, where each component contributes its own energy.

Let $A[1,2, \dots, N]$ is an array of probable logo candidates. Applying SVD to 'A_i' results into decomposition comprising of $U_i[M,N]$, $S_i[N,N]$ and $V_i[N,N]$. The $U_i[M,N]$ is a column orthogonal matrix, $S_i[N,N]$ is a diagonal matrix and the $V_i[M,N]$ is an orthogonal matrix satisfying the following mathematical conditions.

- a) $A = U \times S \times V^T$, 'U' is an orthogonal matrix with columns being Eigenvectors of $A \times A^T$ and 'S' is a diagonal matrix with elements S_1, S_2, \dots, S_n ; where each 's_i' represent singular values of 'A'. The values of 's_i' are given by equation (2.2)

$$S_i = \sqrt{\text{Eigenvalue of } A \times A^T} \quad (2.2)$$

b) In the same way, ‘V’ is also an orthogonal matrix with column values representing Eigenvectors of $A^T \times A$.

2.4.4 Logo Detection and Extraction

The energy ‘ E_i ’ of the probable logo candidates ‘ A_i ’ is computed using equation (2.1). The candidate ‘ A_i ’ having maximum energy from the pool is considered as actual logo candidate of the document. By using the coordinates of the detected logo, it can be extracted from the document image.

2.5 Proposed Logo-based Document Image Retrieval

The architecture of the proposed Logo Based Document Image Retrieval (LBDDIR) is depicted in Fig. 2.2. Preprocessing, logo detection/extraction, feature computation, logo matching and document retrieval are the sequence of steps used in the proposed system. The preprocessing, logo detection and extraction steps are discussed in section 2.4. The remaining steps are discussed in the following sections.

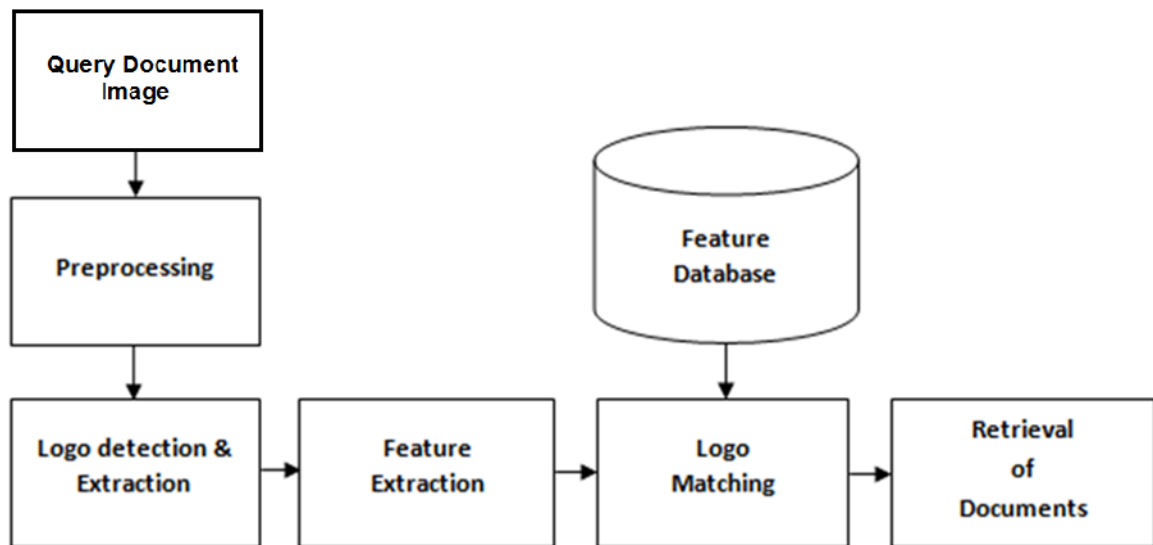


Fig. 2.2 Architecture of the Proposed Logo-based Document Retrieval System

2.5.1 Feature Computation of Detected Logo

Two sets of features namely (i) SVD based derived features and (ii) Singular values of the singular matrix are proposed for logo-based document retrieval. The method used for computing these features is described below.

- **SVD based Derived Features**

Since SVD can be applied on a square matrix, the extracted logo from the document is resized to have ‘N’ rows and ‘N’ columns. Let $L[N:N]$ is a square sized logo and decomposed by applying SVD given by equation (2.3)

$$L[N:N] = U \times S \times V^T \quad (2.3)$$

The ‘U’, ‘S’ and ‘V’ in the equation represent decomposed matrices after application of SVD. In the proposed work, the energy and standard deviation of matrices ‘U’, ‘S’ and ‘V’ are used to construct the first set of features. The equations (2.4), (2.5) and (2.6) are used to compute the energy and (2.7), (2.8) and (2.9) are used to compute the standard deviation respectively.

$$E_U = \frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N |U(i,j)| \quad (2.4)$$

$$E_S = \frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N |S(i,j)| \quad (2.5)$$

$$E_V = \frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N |V(i,j)| \quad (2.6)$$

Where, ‘ E_U ’, ‘ E_S ’ and ‘ E_V ’ represent the energy of matrices ‘U’, ‘S’ and ‘V’.

$$SD_U = \sqrt{\frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N (U(i,j) - \mu_U)^2} \quad (2.7)$$

$$SD_S = \sqrt{\frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N (S(i,j) - \mu_S)^2} \quad (2.8)$$

$$SD_V = \sqrt{\frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N (V(i,j) - \mu_V)^2} \quad (2.9)$$

Where, ‘ SD_U ’, ‘ SD_S ’ and ‘ SD_V ’ are the standard deviation of the matrices ‘U’, ‘S’ and ‘V’. Similarly ‘ μ_U ’, ‘ μ_S ’ and ‘ μ_V ’ represent their mean values. Finally, the feature vector ‘FV’ is constructed using equation (2.10)

$$FV = \{E_U, E_S, E_V, SD_U, SD_S, SD_V\} \quad (2.10)$$

- **Singular Values of a Singular Matrix**

The ‘S’ is a diagonal matrix resulted by applying SVD on the logo $L[N:N]$. It consists of singular values, which are obtained using equation (2.11).

$$S_i = \sqrt{\text{Eigenvalues of } L^T \times L} \quad (2.11)$$

Where ‘L’ represents a matrix with logo intensity values and ‘L^T’ is its transpose. The singular values bear much important information of the image [50] and hence they can form a good set of features for logo matching and retrieval of the documents. In the experimentation, the extracted logo is resized to 50×50 pixels to obtain a feature vector of 50 singular values which is represented by equation (2.12).

$$FV = \{S_1, S_2, \dots, S_{50}\} \quad (2.12)$$

Where, “S₁, S₂, ..., S₅₀” in the equation (2.12) are the singular value features.

2.5.2 Logo Matching

The computed feature vectors of logos corresponding to each document are stored in the feature database. Let FDB[1:N] holds features of ‘N’ documents and ‘FVQ’ is logo features of the query document. The documents present in the database are indexed based on the similarity between ‘FDB’ and ‘FVQ’. The Canberra distance given by equation (2.13) is used to compute the similarity.

$$CanDist(j) = \sum_{i=1}^t \frac{|FVQ_i - FDB_i|}{|FVQ_i| + |FDB_i|} \quad \text{for } j = 1, 2, \dots, N \quad (2.13)$$

The ‘CanDist’ is an array holding similarity distance values between the query and the other documents of the database. The lowest distance corresponds to the closest match and vice-versa.

2.5.3 Retrieval of Documents

To retrieve the documents user is asked to submit a number of documents he or she wish to retrieve. Let ‘K’ is the number of documents to be retrieved from the database. Now top ‘K’ documents based on the similarity are accessed and displayed on the console. The algorithm 2.2 lists the sequence of steps employed in the proposed document retrieval method.

Algorithm 2.2: Logo-based Document Image Retrieval

1. Begin
Input: Query document image.
Output: Retrieved documents from database
2. Let $FDB[1:N]$ = Database of feature vectors for N documents.
3. Read query document image.
4. Preprocess the query document image.
5. Extract logo from the document.
6. Compute features of extracted logo and create feature vector for query document FVQ.
7. Obtain Canberra distance between $FDB[1:N]$ and FVQ.
$$CanDist[i] = CanberraDist(FDB_i, FVQ)$$

Where $CanDist[1:N]$ holds Canberra distance between feature of query document and data base of documents.
8. Sort and rank all the documents based on distance values.
9. Retrieve top-K documents and display the results.
10. End

2.6 Experimental Results

The proposed logo detection and logo-based document retrieval system are tested using the following database in the experimentation and the corresponding results are provided in subsequent sections.

2.6.1 Image Database

The publicly available Tobacco-800 database is used for testing the system. It is a subset of IIT CDIP and has 42 million pages of document images. A total of 266 document images comprising a variety of logos have been selected to test the proposed algorithms.

2.6.2 Performance of Proposed Logo Detection Method

Fig. 2.3 shows the sample result of logo detection. It includes input document, probable logo candidates and the detected logo. The parameter logo detection rate is used to assess the performance of the proposed algorithm. It is the ratio of correctly detected logos from document images to the total number of input documents.

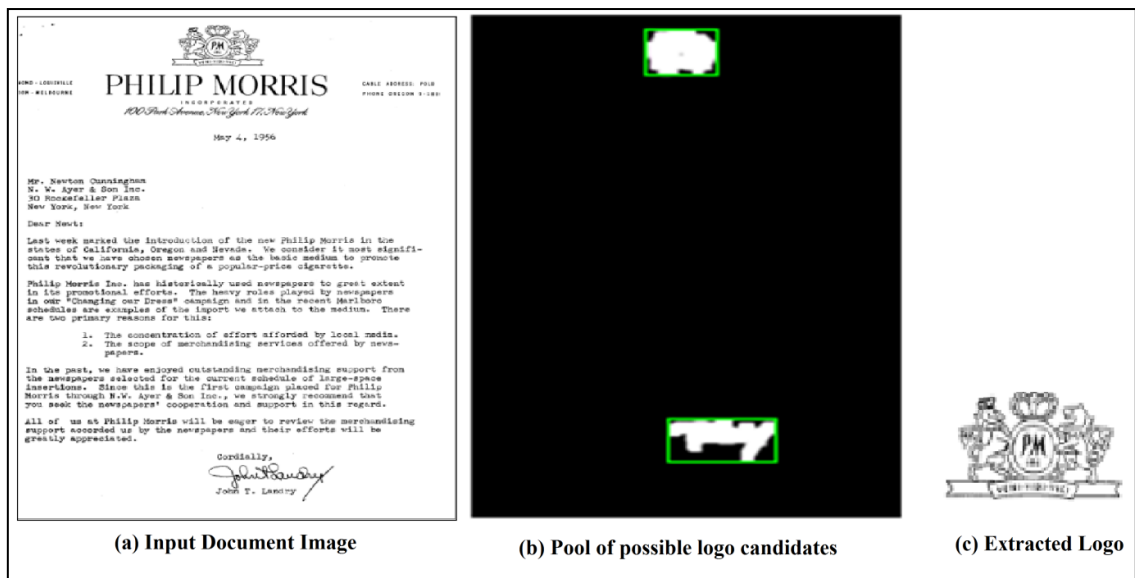









Fig. 2.3 Sample Logo Detection Result

Table 2.1 Logo Detection Results

Sl. No.	Logo type	Number of Documents	Detection Rate (%)	
			Earlier Method [35]	Proposed Method
1		54	90	94
2		65	70.7	96.9
3	<i>Lorillard</i>	54	77.7	98
4		10	30	80
5	B&W	42	38	64
6		23	85.7	91
7		31	13	96.7
8		6	100	100
9		10	40	90
Average detection rate			60.56	90.06

Equation (2.14) is used to compute the logo detection rate.

$$\text{Logo detection rate} = \frac{\text{Number of correctly detected logos}}{\text{Number of logos present in the ground truth}} \quad (2.14)$$

The performance of the algorithm may be influenced by the type of logos. Hence the document images containing the different type of logo are carefully chosen to evaluate the performance. The results obtained are compared with the technique presented in [35]. The results obtained for different category of logos is tabulated in Table 2.1. It reveals that the proposed system outperforms compared with the earlier technique [35].

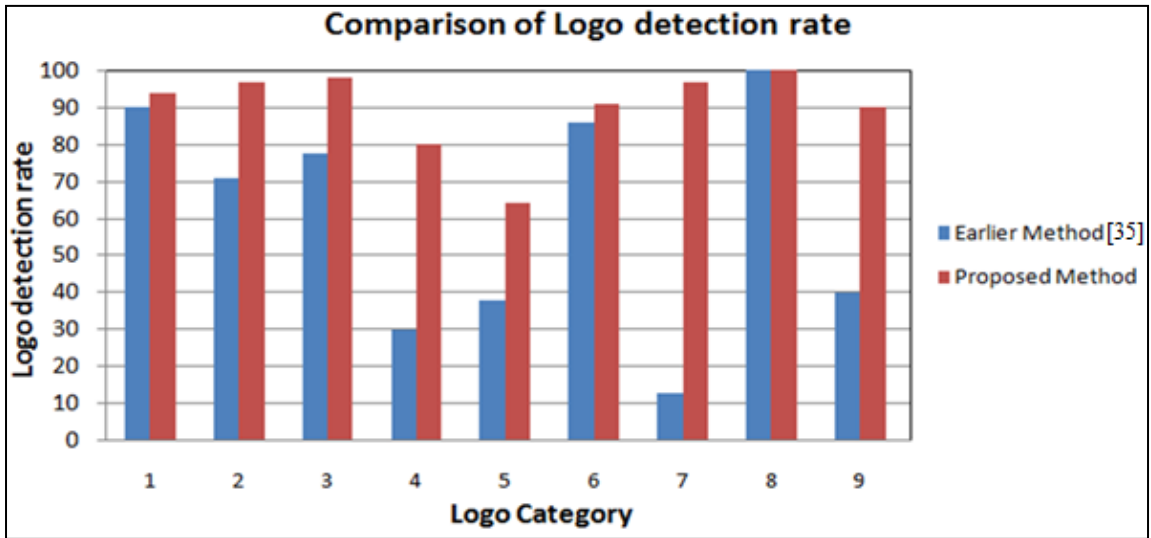


Fig. 2.4 Comparison of Logo Detection Results

Fig. 2.4 shows a graphical comparison of the results for the document images with different category of logos. The logo detection rate obtained with the proposed algorithm for each category of documents is much better compared to the earlier method. An average detection rate is also taken into account for comparing the logo detection performance. The average detection rate is obtained using equation (2.15).

$$\text{Average logo detection rate} = \frac{1}{NLC} \sum_{i=1}^{NLC} \text{logo detection rate} \quad (2.15)$$

Where 'NLC' is the number of different category of logos considered for evaluation. The proposed method has given an average logo detection rate of 90.6%.

2.6.3 Performance of Logo-based Document Retrieval

Fig. 2.5 shows a sample result using the proposed system. It includes a query document and the top 8 retrieved documents. In the result shown, out of 8 retrieved documents all the 8 are relevant documents yielding a precision of 100%.

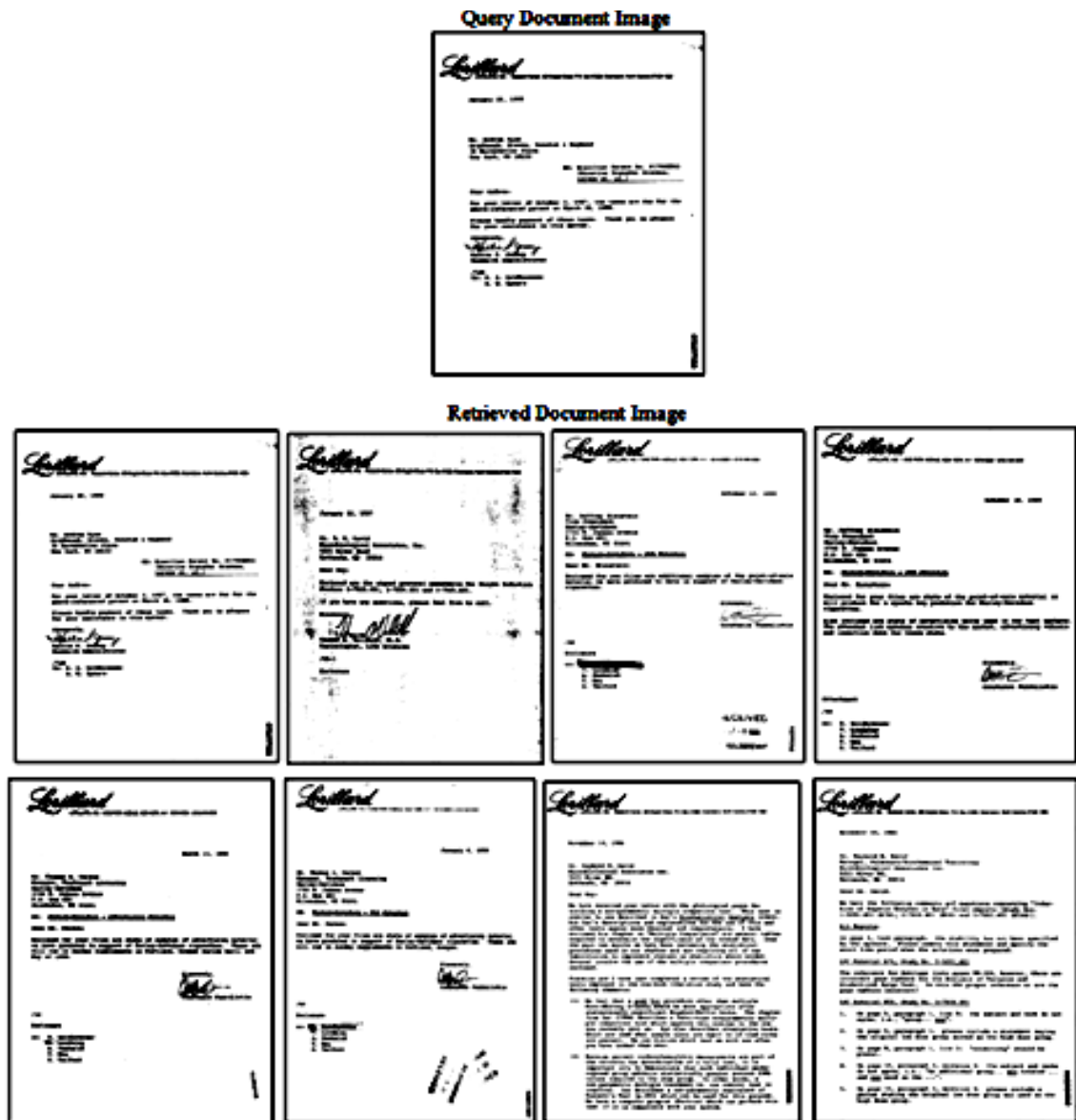


Fig. 2.5 Sample Result of Logo-based Document Retrieval

Experiments are conducted by choosing a random query document to retrieve top10 documents. The precision is used as an assessment parameter to evaluate the results of logo-based document retrieval. The results computed using the proposed method are compared with DWT based features [35] and shown in Table 2.2. Fig. 2.6 shows a graphical comparison of results. The results imply that the proposed method provides better performance. The document retrieval using derived SVD features gave a performance of 76% and the singular values 84%.

Table 2.2 Experimental Results

Query Document	Precision		
	Earlier Method[35]	Proposed Method (with derived SVD features)	Proposed Method (with singular value features)
1	30	40	60
2	30	60	60
3	70	70	90
4	30	30	30
5	50	100	100
6	60	90	100
7	40	100	100
8	70	70	100
9	40	100	100
10	60	100	100
Average Precision	48%	76%	84%

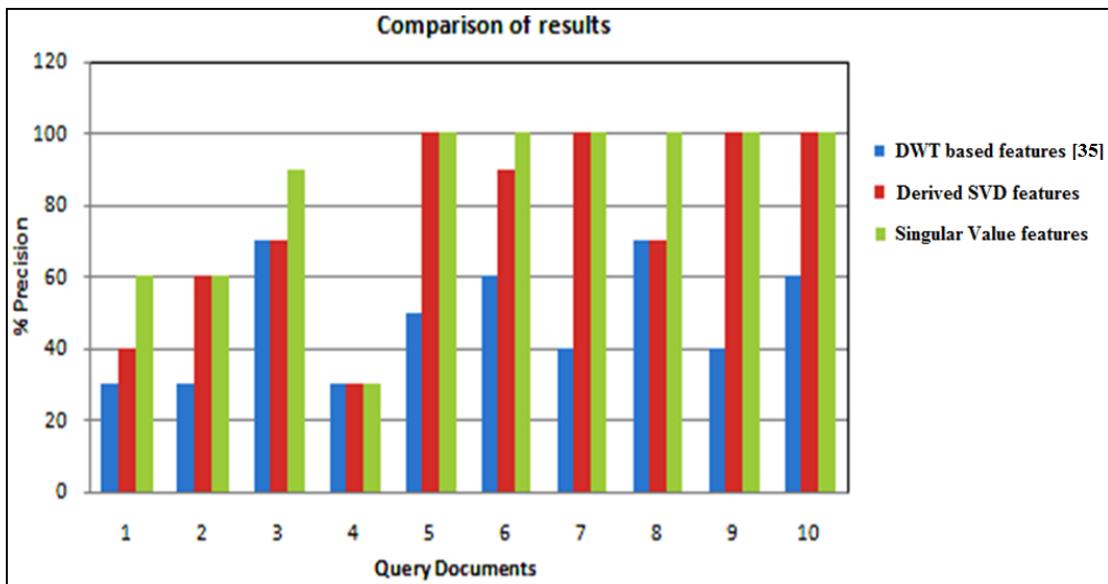


Fig. 2.6 Graphical Comparison of Logo-based Document Retrieval Results

The size of the derived feature vector is 6 and whereas the number of singular values used in the experiment is 50. As the derived features are less in number, the cost of computation is less and it is best suited for large databases.

An experiment is also conducted using N/2 and N/4 singular values as features to observe the variation in retrieval performance. As 'N' is 50 in the proposed algorithm N/2 happens to be 25 and N/4 to be 12. The results obtained are tabulated in Table 2.3. It can be observed that, even by reducing the number of features to 50% and 25%, the singular

values prove to be a good set of features and provide comparable results with a less computational cost.

Table 2.3 Results with N/2 and N/4 Singular Values

Query Document	Precision in (%)		
	Using 'N' singular value features	Using 'N/2' singular value features	Using 'N/4' singular value features
1	60	60	50
2	60	50	50
3	90	80	70
4	30	30	30
5	100	90	80
6	100	100	100
7	100	100	100
8	100	90	80
9	100	100	100
10	100	100	90
Average Precision	84%	80%	75%

2.7 Summary and Conclusion

This chapter proposed two sets of features based on the mathematical tool SVD for implementation of logo-based document image retrieval. The proposed singular value features outperform in comparison with DWT based features by providing a precision of 84%. The logo-based document retrieval gives good results but it suffers from a limitation. Some of the documents do not contain the logo but do contain a signature. So there is a need for the development of signature-based document retrieval. This motivated us to develop automatic signature-based document image retrieval in the next chapter.

Chapter 3

Signature-based Document Image Retrieval

Abstract of the Chapter: The signature-based document retrieval has drawn the attention of many researchers due to its wide applications. It aims to provide a solution towards the retrieval of documents from the huge database depending on the query signature. This chapter proposes an automatic signature detection and signature-based document retrieval using multi-level DWT features. The chapter also provides a comparison of different distance metrics which are used for matching and retrieving of the documents. Namely, the seven distance metrics such as Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard are used for assessing the performance of the retrieval results. The experiments show that the city block distance metric provide better retrieval results using multi-level DWT based features. A precision of 80% is obtained using the city block distance metric in combination with multi-level DWT features.

3.1 Introduction

A signature is someone's name or nick-name written in a distinctive fashion as a mark of identity and intent. Document images such as forms, cheques, notices, circulars, etc. comprise of some textual information with the signature of the authorized person. Such documents rarely contain any logo and hence logo-based document retrieval cannot be used in such a scenario. The signature present in the document image provides an opportunity for indexing of the documents. Therefore to retrieve these types of documents the signature-based document image retrieval is proposed. The general method of signature-based document retrieval includes feature extraction from the query signature and matching these features with signatures that are present in the documents of a database for retrieval. The proposed method adopted a two-stage approach; the first stage deals with the detection and extraction of signature from the document and the second stage deals with the retrieval of documents based on the signature.

The main contribution of this chapter is proposing new techniques for signature detection and signature-based document image retrieval. The proposed signature detection algorithm employs length of the connected components to identify the probable set of

signatures at the coarse level and a distance-based classifier at a finer level. Signature-based document retrieval is proposed using multi-level DWT based features. As the distance metrics play a vital role in matching and retrieval of the document images, experiments are conducted using seven distance metrics. The city block distance provided the highest precision of 80% in comparison with other metrics.

3.2 Problem Statement

The problem formulated is the retrieval of document images using the signature extracted from the submitted query document image. Given a set of document images and a query document, the goal is to retrieve the documents that have the same signature as that of the query.

3.3 Related Work

A lot of algorithms and techniques for signature detection and retrieval are presented by the researchers. This section describes the state of art techniques developed towards segmentation of document signature and the signature-based document retrieval.

Djeziri et al. [52] introduce the use of filiformity for detection of signature from the document. “Filiformity is defined for two topological measures and it differentiates the contour lines of the signatures from the handwritten lines which are being isolated”. A technique for signature extraction from the bank cheques including other documents is presented by Madasu et al. [53]. They used a sliding window that has an area approximately equal to the signature area. The two parameters entropy and the density of the window are used to detect the signature and a fuzzy related theory is used for signature verification. Chalechale et al. [54] worked on Arabic and Persian documents and proposed a method for signature-based classification with the retrieval of documents. They used geometric characteristics of the signatures in their method. The link between the feature vector of signature and the documents is established to retrieve the documents. Average retrieval rate was used as an evaluation parameter for assessing retrieval results in their work.

Signature-based document retrieval using shape-based features is presented by Srihari et al. [55]. The normalized correlation distance metric is used for matching the features. Zhu et al. [56] used multi-scale saliency features for signature extraction. The saliency function designed by them is such that, the functional value increases with the curve length and decreases with the curvature. The technique developed was applicable to

different languages. A method for extracting the signature from the labeled region is proposed in [57]. The salient contour, which is obtained by detecting the skeleton from the document, is used for signature extraction. Srinivasan et al. [58] proposed the use of conditional random fields (CRF) for signature-based document retrieval. The CRF is used for both signature extraction and also for indexing the documents.

A three-stage approach for signature extraction from documents is presented by Mandal et al. [59]. The algorithm employs word-level features for locating the signature and also to separate overlapping text. The CRF minimization energy concept with skeleton analysis is used for classification of actual signature strokes from the text. Roy et al. [60] presented signature-based retrieval for the documents having a cluttered background. They characterized the signature object by spatial features computed from recognition result of background blobs. A codebook of the background blobs is employed for indexing and the Generalized Hough Transform (GHT) is used to detect the query signature. A two-stage method for signature detection was proposed by Mandal et al. [61]. The word wise components with gradient-based features are used for segmentation and classification of the signature block of the document. The SIFT (Scale-Invariant Feature Transform) descriptors and Spatial Pyramid Matching (SPM)-based approaches are used for signature recognition in the second stage. In both the stages, SVM is used as a classifier.

A technique to detect signature from the document image is presented by Cuceloglu and Ogul [62]. They used a two-phase connected component labeling approach for signature segmentation. The method is also investigated with a combination of multiple features with SVM classifier. Signature matching method using pre-filtering was proposed to reduce the search space by Schulz and Sablatnig [63]. They employed shape context distance to perform pre-filtering and a Thin-Plate Spline (TPS) transformation for feature matching. Nurdiyanto and Hermanto [64] proposed a signature recognition technique using the neural network. They used Shannon entropy to extract the features and a Probabilistic Neural Network for signature recognition. Seyyid Ahmed Medjahed [65] provided a comparison of different feature extraction techniques to classify the images in different applications. Belhallouche and Kpalma [66] presented shape adaptive DWT for region-based retrieval. The advantage of this method is a number of coefficients after transformation will be same as that of a number of pixels in the region.

This chapter presents a two-stage approach for signature-based document retrieval using multi-level DWT based features. The first stage is concerned with the extraction of

signature from the query document and the second stage with document retrieval. The main contribution of this work is to provide an experimental comparison of similarity measures for signature-based document retrieval using DWT based features.

3.4 Proposed Methodology

The single-level DWT provides a multi-resolution analysis of the image but fails to acquire minute features of the image. More accurate features can be acquired to improve the performance of matching and classification using multi-level DWT. But the number of DWT levels to be used depends on the size of the image. In this research, as the size of the signature happens to be small compared to general images, a two-level DWT is found suitable to extract the features. As discussed in the previous section, the proposed method is implemented in two stages: (i) Signature detection and extraction from the query document and (ii) Signature-based document retrieval.

3.5 Signature Detection and Extraction

Algorithm 3.1 enlists the steps used in the process of signature detection and extraction.

Algorithm 3.1: Signature Detection and Extraction	
1.	Begin Input: Query document, Output: Extracted signature
2.	Read document image containing signature as query.
3.	Preprocess the query document.
4.	Find connected components and perform length based thresholding to get possible signature candidates.
5.	Apply multi-level DWT to all possible signature candidates and obtain feature set FSPC[1:N]. Where FSPC represents an array containing feature set of possible candidates.
6.	Find the distance between FSPC[1:N] and FSSS _i [1:N] using Canberra distance and store in CanDist[i]. Where FSSS _i [1:N] will contain feature set of sample signatures.
7.	Detected signature candidate = candidate with lowest(CanDist).
8.	Extract the detected signature candidate.
9.	End

The idea behind signature extraction is, first finding a set of probable signature candidates based on the length of the connected components and later detecting the signature

component by applying a classifier. The steps used in Algorithm 3.1 are briefly explained in the following sections.

3.5.1 Preprocessing

Processing of the document images in their original form is time consuming due to the presence of a large amount of data. Hence, this step is used for image conversion to minimize the data and also to remove noise from the document. The given query image is converted to grayscale and then to binary form using the Otsu method. The pixels with value '0' and '1' represent black and white colors respectively. Initially, a median filter of size 3×3 is used to remove the impulse noise present in the document. Later a morphological dilation is carried out to join the small gaps between the pixels that are created during image conversion. This step also helped to merge the broken lines of the signature.

3.5.2 Finding Probable Signature Candidates

Generally, the length of the signature component will be large compared to other components of the document. This visual feature is used to detect the initial set of signature candidates. Hence, in this step, the connected components of the preprocessed document are identified and the length-based thresholding of the components is performed. In the proposed method the length chosen for detection of the signature candidates is depending on the length/size of the characters present in the document. Empirically all the connected components whose length is double than that of the average length of the characters are considered as probable signature candidates.

3.5.3 Feature Extraction

The multi-level DWT [67] based features are used to detect the signature component from a pool of probable signature candidates obtained from the previous step. Multi-level DWT features help to improve the performance of signature detection. In the proposed method the number of levels is limited to two. This section provides a brief introduction to single-level DWT and then explains the method used for obtaining proposed multi-level DWT features.

Single Level DWT: Morlet and Grossman initiated the term wavelet in the design of Morlet wavelets. Meyer presented wavelet with the orthogonal property during the year

1984. “The orthogonal property states that the information obtained by one wavelet will be entirely independent of the information captured by another wavelet”.

The mother wavelet ‘ $\Psi_{a,b}(t)$ ’ is given by equation (3.1) is the basis for deriving all the kernel functions in DWT.

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi \frac{t-b}{a} \quad (3.1)$$

The term ‘a’ and ‘t-b’ in the equation are referred to as scaling and the translation parameters. The fraction $\frac{1}{\sqrt{a}}$ is the normalization factor which assures the uniform distribution of energy among the wavelets. The Fig.3.1 shows the structure of mother wavelet.

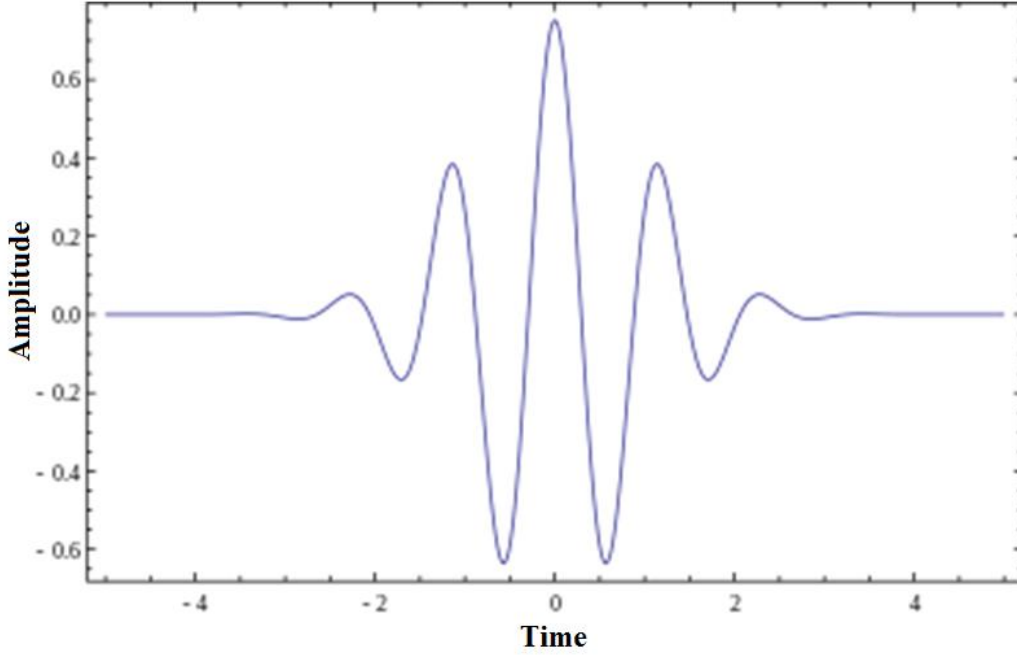


Fig. 3.1 Plot of a Mother Wavelet

The discrete wavelet transform of a two-dimensional signal $f(x,y)$ is given by equations (3.2) and (3.3)

$$W_{\phi}(j_0, m, n) = \frac{1}{\sqrt{M \times N}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \phi_{j_0, m, n}(x, y) \quad (3.2)$$

$$W_{\Psi}^i(j_0, m, n) = \frac{1}{\sqrt{M \times N}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \Psi^i_{j_0, m, n}(x, y) \quad (3.3)$$

Where,

- j_0 is an arbitrary scale.
- $W_0(j_0, m, n)$ is a function of approximation of $f(x,y)$ with the scale j_0 .
- $W_{\downarrow}^i(j, m, n)$ are the horizontal, vertical and diagonal details with the scale $j \geq j_0$.
- M and N are the row and column dimensions of the input image

A series of low-pass and high-pass filters as shown in Fig. 3.2 are employed to obtain DWT. The letters ‘H’ and ‘G’ used in the diagram represent low-pass and high-pass filters respectively. The low-pass filtering and high-pass filtering are the results of convolving pixels of the image with moving-average and the moving-difference masks. The notations $\downarrow 2$ and $\downarrow 1$ are used for representing down-sampling of columns and rows respectively. The process is initially applied along the rows and then to the columns of an image. This process leads to four sub-bands as listed below.

- CA – Approximate sub-band that contains the down-sampled original image.
- CH – Horizontal details of an input image.
- CV – Vertical details of an input image.
- CD – Diagonal details of an input image.

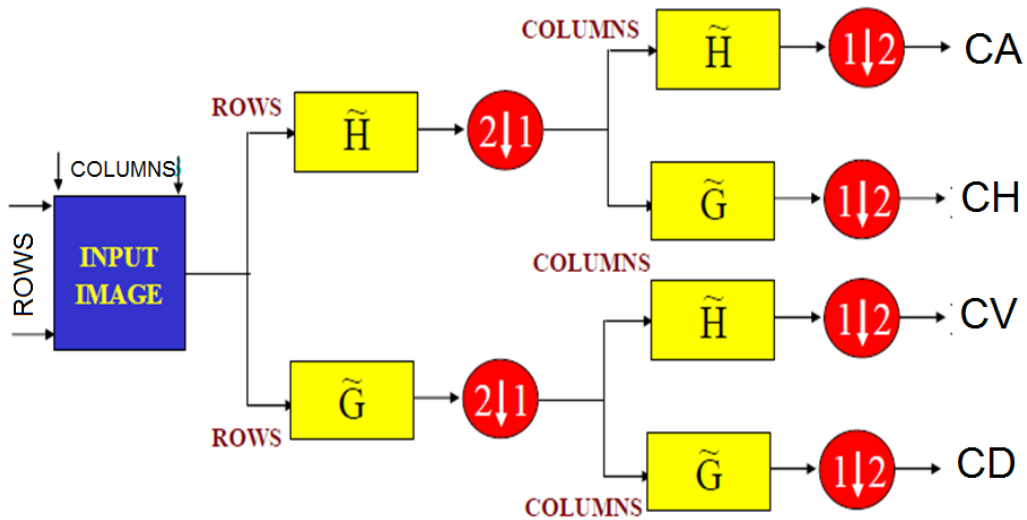


Fig. 3.2 DWT Decomposition of an Image

Proposed Multi-level DWT Features: The two-level DWT is obtained by applying DWT on approximation coefficients of the first level. Fig. 3.3 shows the method of applying two-level DWT. The CA1, CH1, CD1 and CV1 are approximate, horizontal, diagonal and vertical coefficients resulted by applying DWT on the signature component.

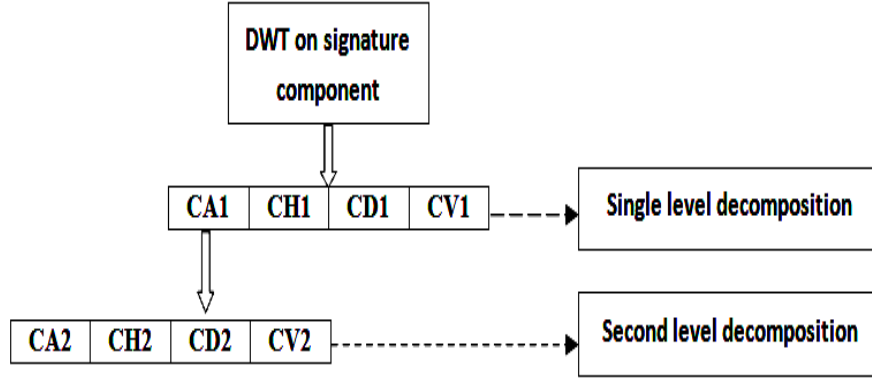


Fig. 3.3 Two-level DWT

To obtain the second level, DWT is again applied on CA1 as shown in Fig. 3.3. This results in second-level coefficients CA2, CH2, CD2 and CV2. The energy and standard deviation of the coefficient matrices are used to form the feature vector. The equations (3.4) to (3.11) are used to compute energy and standard deviation of the coefficients.

$$E_{CA1} = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N |CA1(i, j)| \quad (3.4)$$

$$SD_{CA1} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N (CA1(i, j) - \mu_{CA1})^2} \quad (3.5)$$

Where, E_{CA1} , SD_{CA1} and μ_{CA1} in equations (3.4) and (3.5) represent energy, standard deviation and mean of approximate coefficient matrix CA1.

$$E_{CH1} = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N |CH1(i, j)| \quad (3.6)$$

$$D_{CH1} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N (CH1(i, j) - \mu_{CH1})^2} \quad (3.7)$$

Where, E_{CH1} , SD_{CH1} and μ_{CH1} in equations (3.6) and (3.7) represent energy, standard deviation and mean of horizontal coefficient matrix CH1.

$$E_{CD1} = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N |CD1(i, j)| \quad (3.8)$$

$$SD_{CD1} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N (CD1(i,j) - \mu_{CD1})^2} \quad (3.9)$$

Where, E_{CD1} , SD_{CD1} and μ_{CD1} in equations (3.8) and (3.9) represent energy, standard deviation and mean of diagonal coefficient matrix CD1.

$$E_{CV1} = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N |CV1(i,j)| \quad (3.10)$$

$$SD_{CV1} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^N (CV1(i,j) - \mu_{CV1})^2} \quad (3.11)$$

Where, E_{CV1} , SD_{CV1} and μ_{CV1} in equations (3.10) and (3.11) represent energy, standard deviation and mean of vertical coefficient matrix CD1. Similarly, E_{CA2} , E_{CH2} , E_{CD2} , E_{CV2} , SD_{CA2} , SD_{CH2} , SD_{CD2} and SD_{CV2} are computed for second-level DWT coefficients. Let FV1 is a set of features from first level DWT, as given by equation (3.12) and FV2 is second level DWT features given by equation (3.13).

$$FV_1 = \{E_{CA1}, E_{CH1}, E_{CD1}, E_{CV1}, SD_{CA1}, SD_{CH1}, SD_{CD1}, SD_{CV1}\} \quad (3.12)$$

$$FV_2 = \{E_{CA2}, E_{CH2}, E_{CD2}, E_{CV2}, SD_{CA2}, SD_{CH2}, SD_{CD2} \text{ and } SD_{CV2}\} \quad (3.13)$$

Now the first and second level DWT features are combined to construct the final feature vector FV as shown in equation (3.14).

$$FV = \{FV_1\} \cup \{FV_2\} \quad (3.14)$$

3.5.4 Signature Detection and Extraction

In this step, the features of probable signature candidates are matched with features of stored sample signatures using Canberra distance. Three sample signatures per author are used in the proposed system while testing the algorithm. Let $FSPC_i$ and $FSSS_i$ feature vectors of probable signature components and sample signatures with 'NF' number of features. The Canberra distance between these features is computed using equation (3.15).

$$CanDist(j) = \sum_{i=1}^{NF} \frac{|FSPC_i - FSSS_i|}{|FSPC_i| + |FSSS_i|} \text{ for } j = 1 \text{ to } N \quad (3.15)$$

In the above equation, $\text{CanDist}[1:N]$ is the distance array holding the similarity distance between the features of ‘P’ probable candidates and sample signatures. The probable signature component with the lowest distance is considered as the required signature. The pixels corresponding to this particular component is stored in a 2D array and used as an extracted signature in the retrieval process.

3.6 Signature-based Document Retrieval

In this step, the features of detected signature from query document are compared with the signatures of the documents stored in the database. Algorithm 3.2 shows the steps used in the proposed document retrieval process.

Algorithm 3.2: Signature-based Document Retrieval
<ol style="list-style-type: none"> 1. Begin Input: Extracted signature from query document. Output: Retrieved documents. 2. Let $\text{FDB}_i[1:N]$ holds features of signatures corresponding to the documents stored in the database and $\text{FVQ}[1:N]$ holds multi-level DWT features of extracted signature from query document. 3. Find the similarity distance between $\text{FVQ}[1:N]$ and $\text{FDB}_i[1:N]$ using different distance metrics such as Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard. $\text{SimDist}[i] = \text{Distance}(\text{FVQ}[1:N], \text{FDB}_i[1:N])$ 4. Sort the documents based on similarity values and display top ‘K’ documents on the console. 5. End

The extracted signature is used as an input to the Algorithm 3.2. The array $\text{FVQ}[1:N]$ is used to store Multi-level DWT features of extracted signature. Let $\text{FDB}_i[1:N]$ holds pre-computed features of signatures corresponding to all the ‘N’ documents of the database. To retrieve the documents, two important steps feature matching and ranking of the documents are used. These steps are described below.

3.6.1 Feature Matching

This step is used to find the extent of similarity between the signature of the query document and signatures of documents stored in the database. The results of document

retrieval are influenced by the type of similarity metric used for matching. Hence in the proposed system investigates the performance of the retrieval using seven distance metrics: Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard [68]. Table 3.1 shows the distance metrics used with their corresponding mathematical equations. The ‘X’ and ‘Y’ in the equations are two feature vectors to be matched with dimension ‘N’.

Sl. No	Distance metrics	Equation for computing distance
1	Canberra	$\sum_{i=1}^n \frac{ X_i - Y_i }{ X_i + Y_i }$
2	Euclidean	$\sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$
3	City block	$\sum_{i=1}^n X_i - Y_i $
4	Chebychev	$\text{Max} \{ X_i - Y_i \}$
5	Cosine	$1 - \frac{\sum X_i Y_i}{\sum \sqrt{X_i^2 Y_i^2}}$
6	Hamming	$\sum_{i=1}^n X_i - Y_i $ The distance is 1 if $X_i \neq Y_i$ and 0 if $X_i = Y_i$
7	Jaccard	$1 - \frac{ X \cap Y }{ X \cup Y }$

3.6.2 Ranking of the Documents

Let SimDist[1:N] is a set of distance values between the signature of query document and database documents. These distance values are arranged in ascending order so as to get the lowest distance value at the top. The lowest distance corresponds to the closest match and vice-versa. The documents are also arranged and ranked based on the computed distance values. Based on the users request top ‘K’ number of matching documents are accessed from the database and displayed on the console.

3.7 Experimental Results

The proposed algorithms are investigated on the following database and their performance is described in the subsequent sections.

3.7.1 Image Database

A total of 360 scanned documents which are signed by 18 different authors are used for evaluating the performance. These documents are chosen such that, they contain printed English text with a mixture of graphics such as logos, figures and also tables.

3.7.2 Performance of Signature Detection and Extraction

Some sample document images with the result of signature extraction are shown in Fig. 3.4.

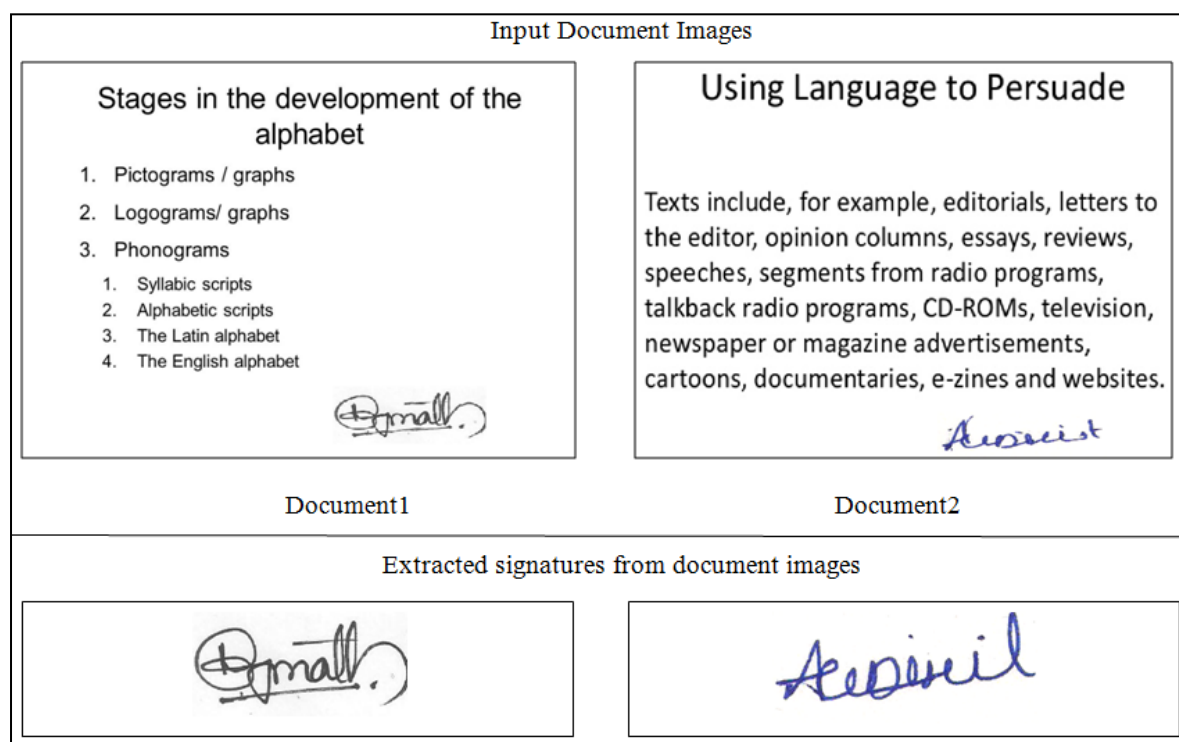


Fig 3.4 Signature Extraction Results

The signature detection rate is used as a parameter to evaluate the proposed signature detection algorithm. It can be defined as the ratio of the number of successful signatures detected from the documents to the total number of documents used for testing. Equation (3.16) is used for computing the signature detection rate.

$$\text{Signature detection rate} = \frac{\text{No. of successful signatures detected}}{\text{Total number of documents with signatures}} \quad (3.16)$$

During the evaluation, the detection is considered as successful only if the detected area of the document contains at least 70% of the signature. The proposed signature detection algorithm is evaluated on a data set of 360 document images and a detection rate of 86.6% is achieved.

3.7.3 Performance of Signature-based Retrieval

Fig. 3.5 shows the sample result of signature-based document retrieval with a query document and top 4 retrieved documents.

Query Document

Using Language to Persuade

Texts include, for example, editorials, letters to the editor, opinion columns, essays, reviews, speeches, segments from radio programs, talkback radio programs, CD-ROMs, television, newspaper or magazine advertisements, cartoons, documentaries, e-zines and websites.

Accipit

Top 4 Retrieved Documents

Using Language to Persuade

Texts include, for example, editorials, letters to the editor, opinion columns, essays, reviews, speeches, segments from radio programs, talkback radio programs, CD-ROMs, television, newspaper or magazine advertisements, cartoons, documentaries, e-zines and websites.

Accipit

WRITING OF THE NEWS

*News is very important thing in today's media scenario and there are a numbers of factors to modify the importance of news in actual practice.

*The policy of news medium may increase or diminish the importance of the story. The class of viewers and listeners that dominates the audience of a channel determines largely what is news for that medium.

*The amount of time available on television determines whether is told briefly or in detail and thus time alters the value of a news story.

*Repeating the same news also sometimes decreases the importance of a news story.

Accipit

Stages in the development of the alphabet

1. Pictograms / graphs
2. Logograms/ graphs
3. Phonograms
 1. Syllabic scripts
 2. Alphabetic scripts
 3. The Latin alphabet
 4. The English alphabet

Accipit

Outcome 1 Reading and Responding

In this area of study the range of texts expands to include a variety of text types and genres, including print, non-print and multimodal texts. (e.g. novel, anthologies of poetry, short stories, scripts for radio, television or stage, narrative films, documentary films, CD-ROMs, and hyperfiction).

Accipit

Fig. 3.5 Sample Result of Signature-based Retrieval

The two parameters, viz. recall (R) and precision (P) are used to evaluate the performance of the document retrieval results. The testing of the algorithm is carried out by retrieving top 20 documents by choosing the query document randomly by the user. Table 3.2 shows the results obtained using single-level DWT and Table 3.3 shows the results obtained using multi-level DWT features with different distance metrics. The average recall and average precision values are also computed for quick interpretation of the results and shown in Tables 3.2 and 3.3.

Table 3.2 Recall and Precision results using Single-level DWT

Query Document	Canberra		Euclidean		City Block		Chebychev		Cosine		Hamming		Jaccard	
	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P
1	13.33	40	16.66	50	13.33	40	13.33	40	3.33	10	33.33	100	33.33	100
2	13.33	40	16.66	50	13.33	40	16.66	50	00	00	33.33	100	33.33	100
3	26.66	80	30	90	26.66	80	30	90	30	90	00	00	00	00
4	33.33	100	33.33	100	33.33	100	33.33	100	33.33	100	00	00	00	00
5	26.66	80	26.66	80	26.66	80	30	90	30	90	26.66	80	26.66	80
6	33.33	100	33.33	100	33.33	100	33.33	100	33.33	100	00	00	00	00
7	23.33	70	16.66	50	23.33	70	20	60	13.33	40	33.33	100	33.33	100
8	20	60	20	60	20	60	20	60	00	00	00	00	00	00
9	26.66	80	26.66	80	26.66	80	26.66	80	33.33	100	00	00	00	00
10	20	60	20	60	20	60	20	60	00	00	00	00	00	00
Average	23.66	71	24	72	23.66	71	24.33	73	17.66	53	12.66	38	12.66	38

Table 3.3 Recall and Precision results using Multi-level DWT

Query Document	Canberra		Euclidean		City Block		Chebychev		Cosine		Hamming		Jaccard	
	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P	%R	%P
1	23.33	70	16.66	50	20	60	13.33	40	20	60	33.33	100	33.33	100
2	23.33	70	20	60	23.33	70	10	30	6.66	20	33.33	100	33.33	100
3	30	90	16.66	50	30	90	20	60	20	60	00	00	00	00
4	26.66	80	30	90	26.66	80	26.66	80	23.33	70	00	00	00	00
5	30	90	30	90	30	90	13.33	40	10	30	26.66	80	26.66	80
6	26.66	80	16.66	50	26.66	80	10	30	26.66	80	00	00	00	00
7	23.33	70	20	60	20	60	16.66	50	23.33	70	33.33	100	33.33	100
8	26.66	80	23.33	70	26.66	80	23.66	70	13.33	40	00	00	00	00
9	20	60	33.33	100	33.33	100	33.33	100	23.66	70	00	00	00	00
10	30	90	30	90	30	90	23.33	70	20	60	00	00	00	00
Average	26	78	23.66	71	26.66	80	19	57	18.66	56	12.66	38	12.66	38

Fig. 3.6 shows a graphical comparison of results obtained using both single-level and multi-level DWT features.

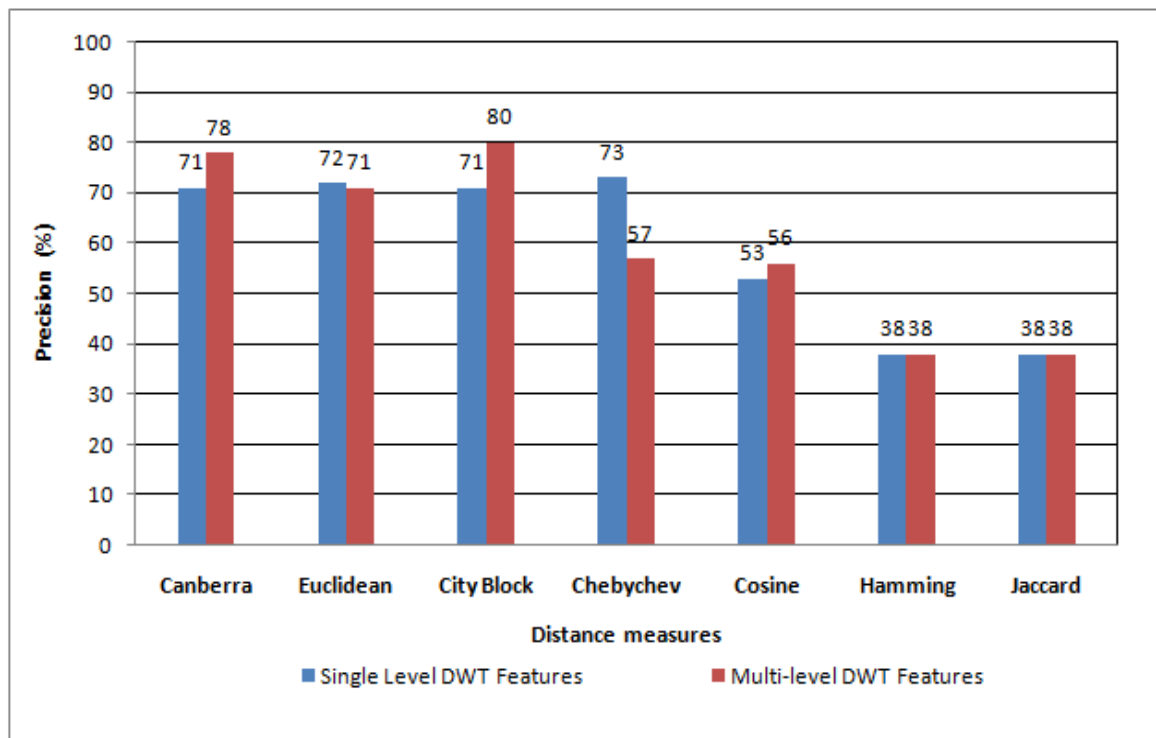


Fig. 3.6 Graphical Comparison of Results

The comparison of experimental results reveals that,

- Multi-level DWT feature extraction scheme provided good results as compared to single-level DWT features with most of the similarity metrics considered for experimentation.
- The city block similarity metric provided better retrieval result (80% precision) in comparison with other distance metrics when multi-level DWT features are used. Canberra distance provided 78% precision which is slightly less compared to a city block. Hamming and Jaccard provided 38% precision which is the lowest among all the distance metrics.
- The Chebychev similarity metric provided the highest precision of 73% with single level DWT features and 57% with multi-level DWT features which is an interesting point observed in the experimentation and needs further analysis. The variation of performance is due to less number of features and limited directional information.
- A precision of 72% and 71% for single-level and multi-level DWT is observed with Euclidean distance. From the experimentation, it seems that Euclidean distance provides better performance with less number of features.

3.8 Summary and Conclusion

The automatic signature-based document retrieval is an essential tool required for searching and accessing the documents. This chapter proposed a multi-level DWT based feature extraction scheme for signature-based document retrieval. The main contribution of this chapter is an investigation of the different distance metrics for matching the signature features. The city block distance metric provided a precision of 80% with two-level DWT based features. The logo and signature-based document retrieval gave promising results but suffer from the limitation. Both the retrieval methods cannot be used for the documents such as identity cards, passports, certificates etc., which may be embodied with photo/face. This motivated us to develop face/photo based document image retrieval in the next chapter.

Chapter 4

Face-based Document Image Retrieval

Abstract of the Chapter: A large number of documents nowadays are found with face/photo of a person. The passports, identity cards, license cards, certificates are few of the example for such documents. The face-base document image retrieval allows for retrieving the documents based on face/photo that the query document consists of. This chapter proposes an automatic face-based document image retrieval using two feature extraction schemes. The first scheme employs the mathematical tool Singular Value Decomposition (SVD) to construct the features and the second scheme extends the concept of Gray Level Co-occurrence Matrix (GLCM) for RGB color image with a simple feature reduction scheme. The proposed feature extraction schemes are tested on a database of 810 documents. The proposed GLCM based feature extraction scheme outperform by providing Mean Average Precision (MAP) of 82.66%.

4.1 Introduction

Recently a lot of the document images are found with face/photo of a person. Few of the examples include passport, identity cards issued by organizations, voter identity cards, certificates, driving license, etc. These documents are also uploaded by the users//authors to the web for the purpose of document verification, account transaction, voting in an election, to obtain a subsidy from the government, etc. The traditional document image retrieval techniques like logo, signature and layout based fail to provide accessing of such document images. This motivated us for the development of face/photo-based document image retrieval as part of this research work. The face/photo-based document retrieval also permits for accessing person-specific documents from a huge database.

The major contribution of this chapter is that, it proposes two feature extraction schemes (1) SVD based features and (2) GLCM based features. The proposed feature extraction schemes are evaluated on a total of 810 document images derived by using the face images of face94 [69] database. The results are compared to feature extraction schemes

used in [35], [36] and [37]. It is observed that the proposed GLCM based features outperform by providing MAP of 82.66%.

4.2 Problem Statement

Given a set of document images that include face/photo of a person, the objective is to retrieve only those documents which have the same face/photo as that of the query document.

4.3 Related Work

This section provides state of art techniques proposed by the researchers for face image recognition, matching and retrieval.

Kekre et al. [70] presented the method for face recognition using Walshlet pyramid. The features are obtained at the different levels of the decomposed image by applying Walshlets pyramid. A kernel machine-based discriminant analysis technique for the nonlinearity of face patterns is presented in [71]. This has solved the problem of a small sample size that was existing in the task of face recognition. Ahmadi et al. [72] proposed a Fuzzy-based Hybrid Learning Algorithm (FHLA) for Radial Basis Function Neural Network (RBFNN) for face recognition. They combined gradient and linear least squared techniques in FHLA for adjusting RBF parameters and also the weights of the neural network.

A face detection technique by employing skin-based colored features has been presented by Weng et al. [73]. The 2-dimensional Discrete Cosine Transform (DCT) is used for extracting the skin-based features and the neural network classifier is used for detection. Tan and Triggs [74] proposed a face recognition method using heterogeneous features. Gabor wavelets and Local Binary Patterns (LBP) are combined for capturing the small appearance with a broader range of details. The principal component analysis is employed for the reduction of resulting features. The Kernel Discriminative Common Vector (KDCV) is then applied for obtaining discriminant nonlinear features in the proposed methodology.

Leng et al. [75] proposed the use of DCT features in the task of face and palm recognition. They described the process of selecting suitable DCT coefficients for an improved discrimination effect and also presented “Dynamic Weighted Discrimination Power Analysis (DWDPA)” for better performance. Park and Jain [76] proposed a

technique for matching and retrieval of the face images using soft biometrics. The proposed method by them suggests the use of demographic information (e.g., gender and ethnicity) and facial marks (e.g., scars, moles, and freckles) for boosting the overall performance of matching and retrieval of face images. Chen et al. [77] presented two orthogonal techniques with the help of attribute-enhanced sparse coding and attribute embedded inverted indexing to retrieve the face images. They used a combination of low-level with high-level features of the human face to improve retrieval performance. Face recognition and retrieval technique employing the fiducial point features were described by Pannirselvam and Prasath [78]. The fiducial points including eyebrow, eye, iris, nose and mouth were extracted to construct the feature vectors and the Euclidean similarity metric was applied for face image matching and retrieval. Dubey et al. [79] used local S, U, V and D sub-bands obtained from SVD of an image. Particularly the descriptors formed by S-sub band are employed for the purposed of face retrieval.

In the literature, it is found that most of the work presented for face detection and recognition is focused on features of individual parts of the face such as eyes, nose, lips etc. Computing individual features of the face for document retrieval becomes an expensive and time-consuming task. Therefore to reduce the number of features the proposed system uses an entire face image as a single object. In this work we propose two feature extraction schemes for face/photo detection and document retrieval: (1) SVD based features and (2) GLCM based features.

In logo-based document retrieval, it was learned that the singular values of an image can form an important feature set. However, using all the singular values for retrieval is expensive; hence we proposed an alternate method for reducing singular values features. In the literature, it is found that various statistics of GLCM are used for image matching and classification. This work proposes an alternate scheme of using GLCM features which are computationally less expensive.

4.4 Proposed Face/Photo-based Document Retrieval System

Fig. 4.1 depicts the architecture of the proposed face-based document retrieval system. Face/photo detection, feature extraction and document retrieval are the building blocks of the architecture. These blocks are discussed in the following sections.

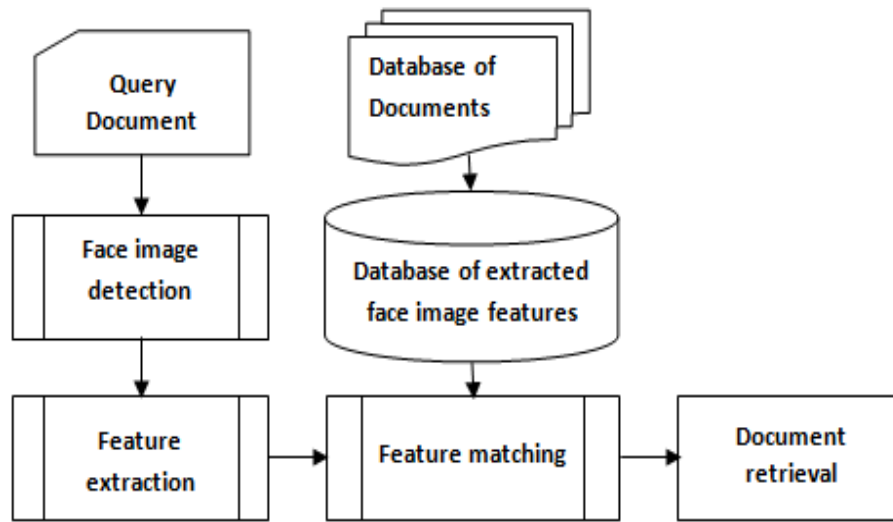


Fig. 4.1 Proposed Face/Photo-based Document Retrieval System

4.4.1 Face Image Detection

In the document images like identity cards, passport, voter id, bank passbooks, driving license; the part of the document containing face or photo of a person will have the highest energy in comparison with the other parts of the document. This basic idea is used for face/photo detection from the documents. Algorithm 4.1 enlists the steps employed for face/photo detection block of the proposed system.

Algorithm 4.1: Face/Photo Detection from Query Document	
1.	Begin
2.	Input: Query document Output: Face image
3.	Read the query document image.
4.	Convert RGB image to Binary image.
5.	Find the connected components of the input document and compute their energy.
6.	Consider the connected component possessing maximum energy as the face and extract the coordinate points and size of this component.
7.	Using co-ordinate points and the size information extract the face/photo from the color query document.
8.	End

As shown in the proposed system, the document containing face/photo of a person is used as an input. The color images are converted into grayscale using the equation (4.1)

$$I = 0.2989 \times R + 0.5870 \times G + 0.114 \times B \quad (4.1)$$

Where 'R', 'G' and 'B' are the red, green and blue components of the color image and 'I'

is the converted grayscale image. Then an Otsu method [80] is applied to convert the grayscale image 'I' into binary form. To detect the face image in the document, the connected components of the document are found by employing 8-neighbors. Connected components of an image are connected set of dark pixels. Each connected component possesses energy depending on the density of pixels that makes the connected set. Let $CC(i,j)$ is a 2D array representing the connected component with a dimension of 'M' rows and 'N' columns. Equation (4.2) is used in the proposed method to compute the energy 'E_{CC}' of such components.

$$E_{CC} = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N |CC(i,j)| \quad (4.2)$$

In the next step, the connected component possessing the maximum energy is considered as the face/photo to be extracted. Using the location of this particular component, the face/photo region from the document is extracted and used for further processing. A sample result of face detection is presented in Fig. 4.2. The identity card of a person is used as an input document for face/photo extraction in the sample result.

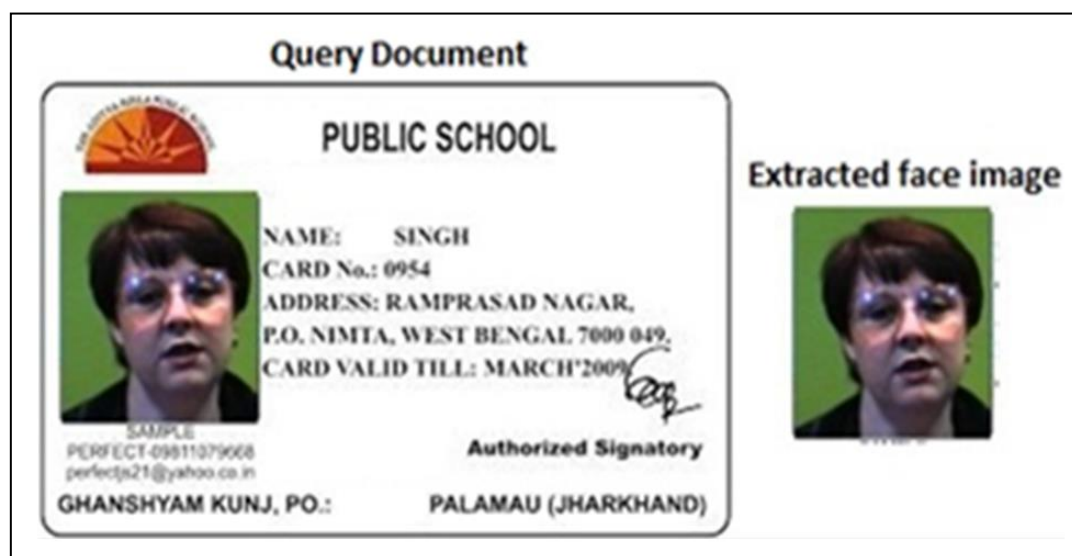


Fig 4.2 Sample Result of Face/Photo Extraction

4.4.2 SVD based Feature Extraction

SVD is a mathematical tool widely used in image processing applications such as image analysis and representation due to its robustness. Application of SVD on an image 'I' decomposes into 3 matrices 'U', 'S' and 'V' as shown in equation (4.3).

$$SVD(i) = U \times S \times V^T \text{ for } i = 1 \text{ to Number_of_Components} \quad (4.3)$$

Where ‘U’ is a column orthogonal matrix, ‘S’ is a singular matrix with only non-zero elements at diagonal and ‘V’ being the orthogonal matrix whose column values are Eigenvectors of image ‘I’ and its transpose. This has motivated us to use the singular values of the ‘S’ matrix in the proposed logo-based document retrieval work. This chapter presents a simple technique to reduce the number of singular value features. Algorithm 4.2 enlists the steps used in proposed feature extraction.

Algorithm 4.2: SVD Based Feature Computation	
1. Begin	
Input: Face image (F)	
Output: Feature Vector (FV)	
2. Resize the face image into 64×64 pixels.	
3. Divide the face image ‘F’ into four blocks as F ₁ , F ₂ , F ₃ and F ₄ with size 32×32.	
4. Decompose F ₁ , F ₂ , F ₃ , F ₄ using SVD and obtain set of singular values {S ₁ }, {S ₂ }, {S ₃ } and {S ₄ }.	
5. Merge diagonal elements of {S ₁ }, {S ₂ }, {S ₃ } and {S ₄ } and store in vector D using equation (4.4)	$D[1:128] = \{S_1\} U \{S_2\} U \{S_3\} U \{S_4\} \quad (4.4)$
6. Construct the Feature Vector ‘FV’ by adding every four consecutive elements of D using equation (4.5).	$FV [1:32] = \{\sum_{i=1}^4 D(i, i), \sum_{i=5}^8 D(i, i), \dots, \sum_{i=125}^{128} D(i, i)\} \quad (4.5)$
	Where, each element of ‘FV’ is trace of the singular matrices.
7. Return ‘FV’	
8. End	

Let ‘F’ is an input face image obtained from the face detection algorithm. Initially, the image is resized to 64×64 pixels and then divided into four blocks F₁, F₂, F₃ and F₄ of size 32×32 for feature extraction. These four blocks are decomposed by applying SVD. Equation (4.6) shows the decomposed matrices ‘U_i’, ‘S_i’ and ‘V_i’ for ith block ‘F_i’.

$$F_i[32:32] = U_i \times S_i \times V_i^T, \text{ for } i = 1 \text{ to } 4 \quad (4.6)$$

Let ‘S₁’, ‘S₂’, ‘S₃’ and ‘S₄’ are the singular matrices with size 32×32 holding the square root of Eigen-values. The 32 diagonal elements from each singular matrix are collected and merged to form a set of 128 features and stored in the vector ‘D’. In the last step, the sum of every four consecutive elements of ‘D’ are computed and merged to construct a

final feature vector 'FV' of size 32. Each element of 'FV' represents a trace of singular matrices 'S₁', 'S₂', 'S₃' and 'S₄'.

4.4.3 Gray Level Co-occurrence Matrix (GLCM) Based Feature Extraction

GLCM is a square matrix of dimension 'N' rows and 'N' columns, used to gather a variety of statistical parameters. The features provided by GLCM are widely used in image analysis and classification [81]. The GLCM will have the same dimension as that of the number of the gray levels of an image. Hence the number of gray levels used in image plays a vital role in computing GLCM. For example, an 8-bit image with 256 gray levels leads to a GLCM of size 256×256. Each element at location (i,j) in GLCM indicates the frequency of occurrence of two pixels whose gray level values are 'i' and 'j' respectively, with a distance of 'k' in the direction specified by the displacement vector. Fig. 4.3 shows an example of GLCM for a 6×6 binary image 'I' with a unit displacement vector in the diagonal direction.

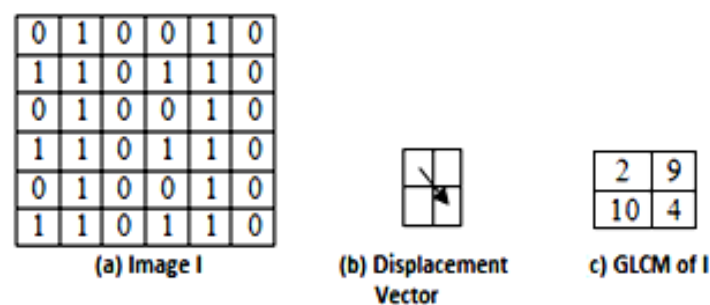


Fig. 4.3 GLCM of an Image 'I'

The first element in the GLCM of image 'I' provides the occurrence of the pixel with gray level value (0,0) in the diagonal direction. In the same way, the other elements of GLCM give occurrence of pixels with values (0,1), (1,0) and (1,1). The GLCM obtained for an image provides the texture information for image analysis. In the literature, the four parameters namely the energy, contrast, correlation and homogeneity known as Haralick features which are widely used for image analysis.

The success story of GLCM motivated us to use it for the purpose of face-based document image retrieval. This chapter proposes a technique to extend the usage of GLCM for an RGB color image and suggest the use of diagonal elements of GLCM to construct a feature vector instead of using complex and computationally expensive statistical parameters. The reason for using the diagonal elements of GLCM is that they

are the non-zero values and provide more unique features of the image [81]. Algorithm 4.3 lists out the steps used in the proposed GLCM based feature extraction scheme.

Algorithm 4.3: GLCM Based Feature Computation	
1.	Begin Input: RGB face image Output: Feature vector.
2.	Separate Red, Green and Blue channels of the face image.
3.	Divide each R, G and B into 8 intensity levels.
4.	Calculate co-occurrence of R, G and B matrices.
5.	Compute: <ul style="list-style-type: none"> a. $FV1 = \{\text{Diag}(\text{CMR})\}$ /*CMR is co-occurrence matrix for R component*/ b. $FV2 = \{\text{Diag}(\text{CMG})\}$ /*CMG is co-occurrence matrix for G component*/ c. $FV3 = \{\text{Diag}(\text{CMB})\}$ /*CMB is co-occurrence matrix for B component*/
6.	Merge FV1, FV2 and FV3 $FV = \{FV1\} \cup \{FV2\} \cup \{FV3\}$
7.	Return FV.
8.	End

The algorithm uses the extracted face image of the document as an input. The RGB color images are represented as a 3-dimensional array comprising of red (R), green (G) and blue (B) colors. Initially, these 'R', 'G', 'B' color components of the face image are separated and stored in individual 2D arrays. The elements in these arrays represent the intensity of values of red, green and blue colors. To reduce the size of the GLCM matrix, the numbers of intensity levels are reduced to 8 by performing intensity mapping. The intensity mapping is obtained by dividing the intensity values of the pixels by 8 and taking an integral part of the quotient. This operation is performed using equation (4.7).

$$\text{Mapped intensity value} = \text{Integer} \left(\frac{\text{Pixel intensity value}}{8} \right) \quad (4.7)$$

Let 'CMR', 'CMG' and 'CMB' are the computed co-occurrence matrices of red, green and blue colored pixels. The dimension of each co-occurrence matrix is 8×8, as the number intensity levels of the pixels are reduced to 8.

Instead of using computationally expensive statistical parameters, the proposed algorithm uses principal diagonal elements of 'CMR', 'CMG' and 'CMB' to construct the feature vector. Since the diagonal elements of co-occurrence matrices will have non-zero values,

they form unique features for image matching. Equations (4.8), (4.9) and (4.10) are used to obtain diagonal elements as follows.

$$FV1 = \forall CMR(i, j), \text{ where } i = j \quad (4.8)$$

$$FV2 = \forall CMG(i, j), \text{ where } i = j \quad (4.9)$$

$$FV3 = \forall CMR(i, j), \text{ where } i = j \quad (4.10)$$

Where, ‘FV1’, ‘FV2’ and ‘FV3’ are the vectors each consisting of 8 elements. Finally, these 3 vectors are merged together to form the feature vector ‘FV’ of size 24 using equation (4.11).

$$FV = \{ FV1 \} U \{ FV2 \} U \{ FV3 \} \quad (4.11)$$

4.4.4 Document Retrieval

The Mahalanobis similarity metric is used for matching the features of query face image and the pre-stored face image features of all the documents. Let ‘FVQ’ is a feature vector of query face image and ‘FDB’ is an array of features of faces/photos present in the database of documents. The Mahalanobis distance between ‘FVQ’ and ‘FDB’ using positive definite covariance ‘C’ is given by equation (4.12)

$$Mdist(FVQ, FDB_i) = \sqrt{(FVQ - FDB_i)^T C^{-1} (FVQ - FDB_i)} \quad (4.12)$$

Where ‘Mdist’ holds the distance values between ‘FDB’ and ‘FVQ’. The documents of the database are arranged based on their distance values computed. Lowest distance value corresponds to the closest match and vice-versa. Now depending on the user request top ‘K’ number of documents are accessed and displayed on the user screen. The proposed feature extraction scheme with a Mahalanobis distance metric has provided good retrieval results. Algorithm 4.4 provides the steps used in the document retrieval process.

Algorithm 4.4: Retrieval and Ranking of Documents

1. Begin
2. **Input:** $FDB[1:N]$ /* Feature vector of 'N' documents stored in database*/
 $FVQ[1:N]$ /* Feature vector of query document*/
Output: Top-K retrieved documents
3. Compute Mahalanobis distance ' $Mdist(FVQ, FDB_i)$ ' between FVQ and each FDB_i using equation (4.12) for finding a match between query face image and each of indexed documents.
4. Sort the distance values stored in ' $Mdist$ ' and corresponding documents for ranking.
5. Display top-K documents, whose features are close to the query document based on sorted ' $Mdist$ ' values.
6. End

4.5 Experimental Results

The database used for evaluation of the proposed face-based document retrieval is explained in the following section and its performance is discussed in the subsequent section.

4.5.1 Image Database

The database is created by combining a variety of documents and the face images borrowed from the database face94 [69]. The database is formed to include documents of 27 members, specifically 16 male and 11 female persons. 30 documents per person are created, which leads to a total of 810 document images ($30 \text{ documents} \times 27 \text{ members}$) for testing. These documents include sample identity cards, passports, certificates, PAN cards etc. with varying size and quality of the print

4.5.2 Performance of Face-based Document Retrieval

The performance of the proposed method is assessed using three evaluation parameters precision, recall and F-measure. Fig. 4.4 depicts the sample result obtained from proposed face-based document retrieval. The sample result is shown for the top 4 retrieved documents. The 3 documents are related to the query out of 4 retrieved documents in the result. This gives a precision of 75%, recall of 30% suppose there are 10 relevant documents present in the database and an F-measure of 42.86%.

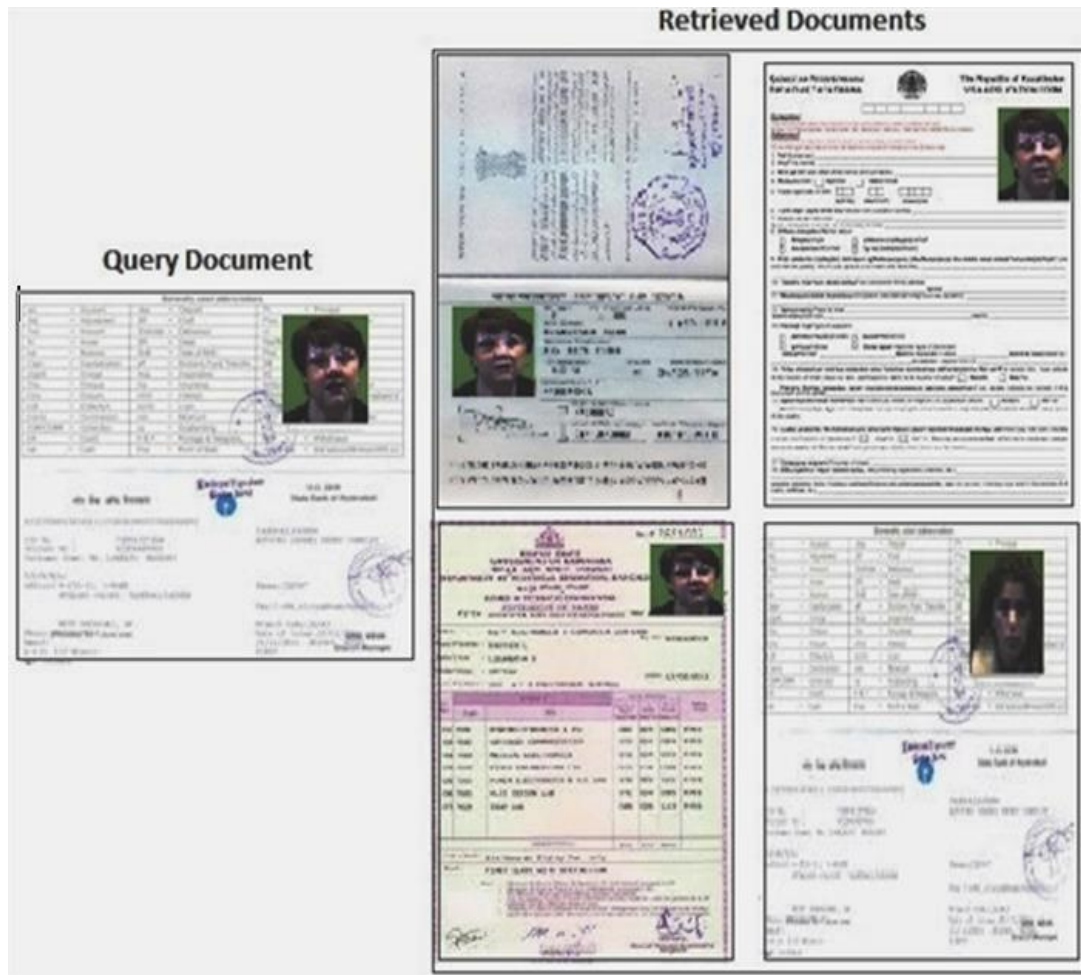


Fig. 4.4 Sample Document Retrieval Result

The proposed method is tested by selecting a query document randomly by the user. Precision, recall and F-measure are computed by retrieving Top 1, Top 5, Top 10, Top 15 and Top 20 documents for all 27 classes. Thus there will be a set of 27 precision, recall and F-measure results for each Top 1, Top 5, Top 10, Top 15 and Top 20 retrieval results. For easier interpretation, the average precision (AP) and average recall (AR) are computed and tabulated as shown in Table 4.1. The table also provides AP and AR values computed using the state of art techniques. The proposed GLCM based features outperform by providing a MAP of 82.66%.

Table 4.1 Average Precision and Average Recall

Top matches	Haralick Features[36]		DWT Features [35]		Proposed SVD based features [37]		Proposed GLCM based features	
	AP	AR	AP	AR	AP	AR	AP	AR
Top 1	100	3.3	100	3.3	100	3.3	100	3.3
Top 5	60.74	10.12	65.93	12.96	77.78	12.96	89.62	15.26
Top 10	46.29	15.43	49.46	18.03	65.57	20.23	84.81	28
Top 15	35.81	17.91	40.78	19.87	52.25	21.83	74.81	41.71
Top 20	29.81	19.86	35.89	22.01	40.15	24.27	64.07	42.71

Fig. 4.5 shows a graphical comparison of AP values obtained using the proposed method and the earlier techniques.

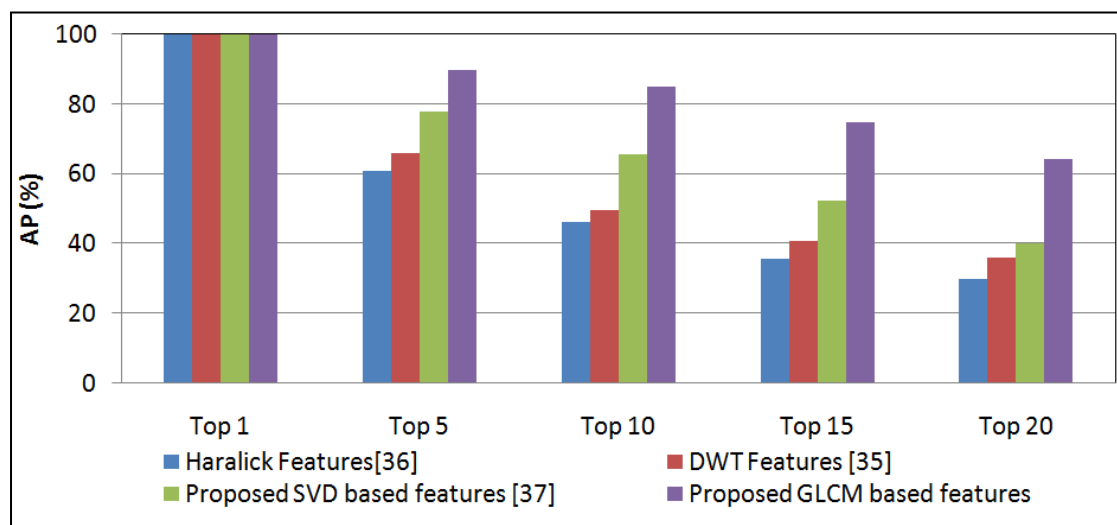


Fig. 4.5 Graphical Comparison of Average Precision

It can be observed that the proposed GLCM based features outperform in comparison with the others. Table 4.2 provides the F-measure values computed for each Top 1, Top 5, Top 10, Top 15 and Top 20 retrieval results.

Table 4.2 Comparison of F-measure

Top matches	Haralick Features [36]	DWT based Features [35]	Proposed SVD based features [37]	Proposed GLCM based features
Top 1	6.34	6.34	6.34	6.34
Top 5	17.35	21.91	22.22	26.06
Top 10	23.13	26.42	30.92	42.10
Top 15	23.87	26.72	30.79	53.55
Top 20	23.84	27.29	30.25	51.24

Fig. 4.6 shows a graphical comparison of f-measure using proposed feature extraction schemes and other techniques. It is evident from the comparison that the proposed GLCM based features provided good F-measure compared with other methods.

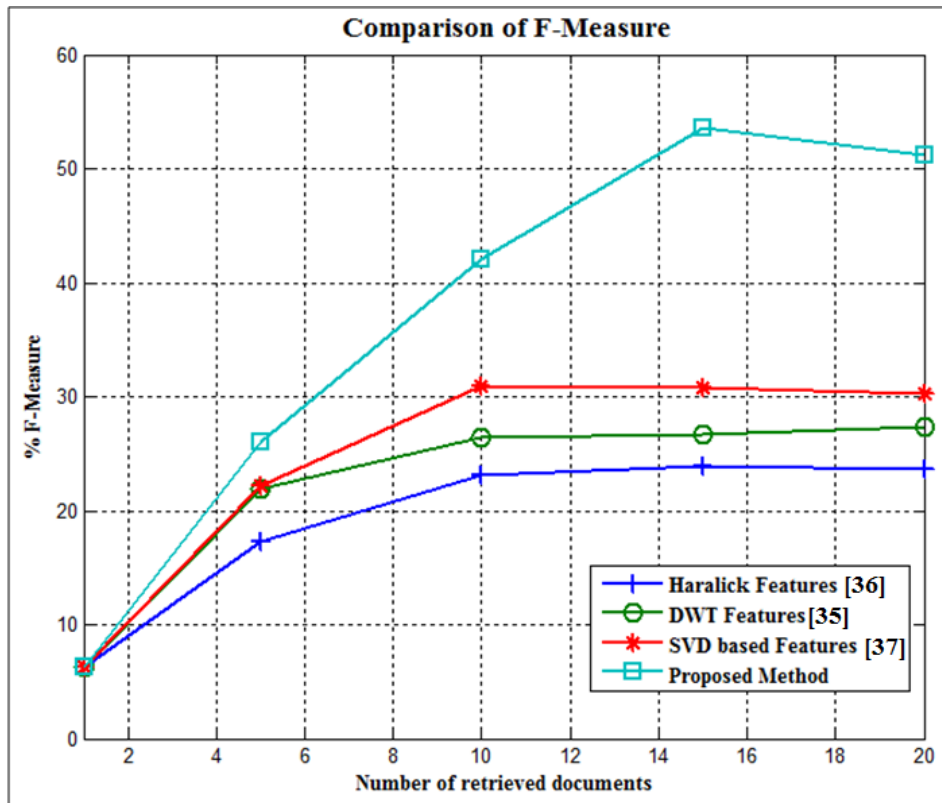


Fig. 4.6 Comparison of F-measure

4.6 Summary and Conclusion

The face-based document image retrieval is an essential application for document images such as identity cards, passports, certificates, license cards, etc. This chapter proposed the use of SVD and GLCM based features for face/photo based document image retrieval. The proposed GLCM based features provided better results compared to feature extraction schemes used in [35], [36] and [37]. Nowadays many of the documents include fingerprint impression to provide more security and authentication. The face/photo based document retrieval is a useful application but it cannot be used for retrieving the documents that contain fingerprint impression. This motivated to develop fingerprint-based document image retrieval in the next chapter.

Chapter 5

Fingerprint-based Document Image Retrieval

Abstract of the Chapter: Recently most of the documents are authenticated using fingerprint impression. The examples for such documents are property registration, banking transactions, insurance documents, etc. The goal of fingerprint-based document retrieval is to provide an easier way of accessing, browsing or searching of such document images. This chapter proposes a simple but efficient fingerprint detection algorithm for documents images using Discrete Wavelet Transform (DWT) based features and Support Vector Machine (SVM) classifier. Two sets of features namely (1) DWT based Local Binary Pattern (LBP) and (2) SWT based LBP are proposed for the fingerprint-based document retrieval process. The proposed fingerprint print detection has given a detection rate of 98.87% for 1100 document images and the retrieval method has given Mean Average Precision (MAP) of 73.08% for a set of 1200 document images.

5.1 Introduction

To provide high security and authentication the recent documents have been included with fingerprint impression instead of a signature. Property registration, banking transactions, insurance documents, etc. are the examples that comprise of fingerprint impression of authorized persons. “Fingerprint represents a unique pattern of ridges and valleys of the surface of fingers [82]”. The traditional method uses paper-ink whereas, a group of sensors has been employed for producing a finger-print impression in an electronic form. In general, a three-level hierarchy is employed for representation of fingerprint friction ridge information. These include a pattern of fingerprints, minute points and ridge contours. Generally, level 1 features are employed for matching latent fingerprint whereas level 2 and level 3 features are used in fingerprint identification systems [83].

To access document images containing fingerprint impression an efficient fingerprint detection technique and the development of a novel feature extraction scheme are the major challenges in the implementation of fingerprint-based document retrieval. The main contribution of this chapter is that, it proposes a simple but yet an efficient

fingerprint detection module and two sets of features namely DWT-based LBP and Stationary Wavelet Transform (SWT) based LBP features. To obtain the proposed sets of features initially, the image is decomposed into four sub-bands: approximate, horizontal, vertical and diagonal coefficients [84] using DWT/SWT. Then the LBP of the sub-bands are computed and the histograms of LBP values are used to construct the features. During the retrieval process, the standardized Euclidean distance is used to improve the performance of the system [85]. The proposed feature extraction schemes provided promising results compared to [38] and [39].

5.2 Problem Statement

The objective of this chapter is to present new techniques for retrieval of document images which comprise the fingerprint of a person similar to that of a query document image.

5.3 Related Work

Many of the algorithms and techniques are proposed by researchers for matching the fingerprints for recognition and verification. These algorithms and techniques are discussed in this section.

Jiang et al. [86] presented fingerprint retrieval technique by employing the features that are obtained from the orientation field and the dominant ridge distance. They proposed a new distance measure which averages the unit vector with phase doubled the orientation for matching the features of fingerprints. Chen et al. [87] proposed a fuzzy feature match (FFM) technique for matching the local triangular features collected from deformed fingerprints. To estimate the similarity, the feature vectors are normalized before applying the distance metric. He et al. [88] presented a three-stage algorithm for global comprehensive similarity. In the first stage, they built a minutia-simplex comprising of a pair of minutiae with their textures that include a transformation-variant and invariant set of features. In the second stage, ridge-based relative features associated with minutiae are used for grouping the minutia depending on their affinity with the ridge. In the third stage, they presented the relationship between transformation and the comprehensive similarity of two fingerprint images in the form of a histogram.

Liu et al. [89] proposed a database clustering model based fingerprint search technique for narrowing the search space. They used multi-scale orientation field-based features as the primary set of features and the average ridge distance as secondary features. A

modified K-means clustering algorithm was applied to partition the orientation feature space into clustering. The fingerprint friction ridge details consist of 3 levels of features. They are level 1 comprising of pattern, level 2 that include minutia points and level 3 consisting of pores and ridge contours. Jain et al. [90] proposed the use of Level 3 features to match high-resolution fingerprints. Zegarra et al. [91] proposed wavelet feature-based fingerprint retrieval method with 3 important tasks: feature extraction, similarity measurement and indexing of the features. They used different types of wavelets namely DWT, tree-structured DWT and the Gabor wavelets for the decomposition of the given image. The features are formed by computing the energy and standard deviation of decomposed fingerprint images.

Jain and Feng [92] proposed a method for matching latent fingerprints with rolled fingerprints. The proposed system employs the use of a quality map in addition to minutiae. Nanni and Lumini [93] presented the hybrid fingerprint matching technique using LBP based features. Initially, the 2 fingerprints are matched and aligned using their corresponding minutiae and then decomposed into non-overlapping windows. These non-overlapping windows are convolved with the Gabor filters to construct LBP histograms. Jung and Lee [94] proposed a method for classification of fingerprints employing the probabilistic approach using features of ridges. Bharkad et al. [95] suggested the use of discrete wavelet packet transform by neglecting horizontal coefficients to obtain redundant features for matching of the fingerprint images.

A method for detection of the convex core point for different types of fingerprints was presented by Le et al. [96]. A modified complex filter known as the semi-radial filter is proposed in their method for detection of rotational symmetries of core points. The vertical variation feature is used for removing spurious core points. Cappelli and Ferrara [97] developed a method for fingerprint retrieval using the combination of a level-1 and level-2 set of features. A hybrid fusion-based technique is used for evaluation of various score and ranking of the fingerprints. Shalaby and Ahmed [98] proposed the use of a multi-level structural approach to recognize the fingerprints, by decomposing them into regions using multi-level features. Paulino et al. [99] proposed a technique to match the latent fingerprints. They used descriptor based Hough transforms to align the fingerprints. The orientation field is used to measure the similarity of fingerprints in the proposed method. Arun et al. [100] proposed a texture-based finger knuckle print recognition method with the help of features formed using LBP variants. They used Local Directional

Pattern (LDP), Local Derivative Ternary Pattern (LDTP) and Local Texture Description Framework based Modified Local Directional Pattern (LTDF-MLDN) based feature extraction in their proposed method. Nearest neighbor and Extreme Learning Machine (ELM) classifiers are used for the classification task. Rodrigues et al. [101] presented a technique to recognize the finger knuckle prints. The Sobel operator was used for detecting the edges. Different similarity metrics are used for recognition of binarized images.

Tzalavra et al. [102] provided the comparative analysis of 3 multi-resolution transform-based features namely DWT, SWT and Fast Discrete Curvelet Transform (FDCT) for assessing breast tumors in Dynamic Contrast-Enhanced Magnetic Resonance Images (DCE-MRI). FDCT based features provided better performance in comparison with the other two. Qayyum et al. [103] employed SWT for feature extraction for the recognition of facial expressions. In particular, the combinations of vertical and horizontal coefficients are used for obtaining muscle movement information.

From the literature, it is learned that the major challenges for implementing fingerprint-based document retrieval include an efficient fingerprint detection technique and a suitable set of features for fingerprint matching and retrieval. The main contributions of this chapter include:

- (1) Proposing an efficient and simple technique for fingerprint detection using DWT based features and SVM classifier.
- (2) Proposing the use of DWT and SWT based LBP feature extraction schemes to match and retrieve the fingerprint-based documents.

5.4 Proposed Method for Fingerprint-based Document Retrieval

The architecture of the proposed fingerprint-based document retrieval is depicted in Fig. 5.1. The major building blocks of the system include fingerprint detection, feature extraction, matching and retrieval of documents from the database. These blocks are discussed in the subsequent sections.

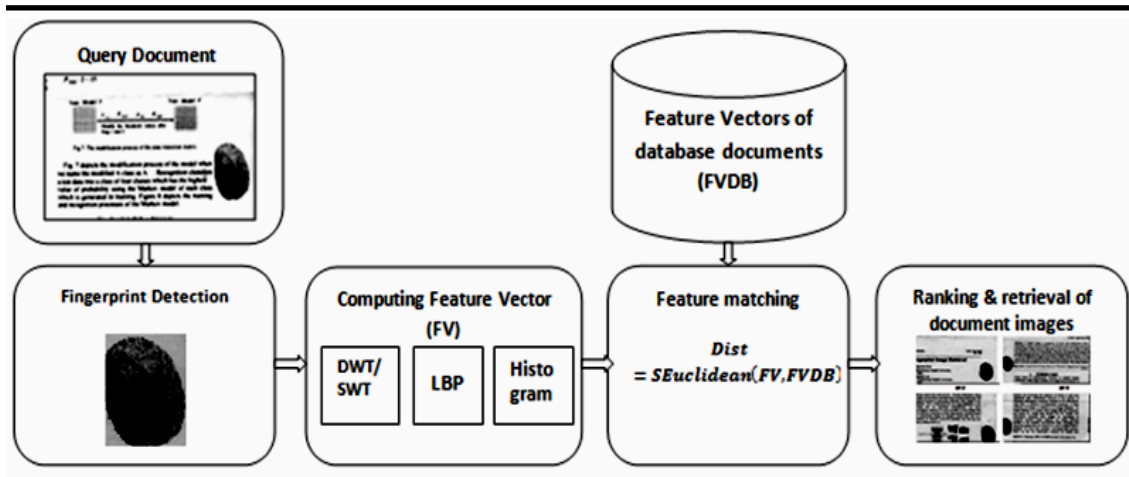


Fig. 5.1 Proposed Architecture of Fingerprint-based Document Retrieval

5.4.1 Fingerprint Detection

Fig. 5.2 shows the proposed system for fingerprint detection. It includes a two-phase approach: a training phase and a testing phase. The proposed system is motivated by the technique presented for discrimination of handwritten and printed text [104]. In the training phase, 140 patches comprising of a variety of text, logos, fingerprint and different symbols are used. The patches are collected from the first 100 document images of the database. Initially, the patches from the submitted query document are obtained using connected component analysis, their DWT based features are extracted and then these patches are classified as fingerprint and non-fingerprint patches using SVM classifier.

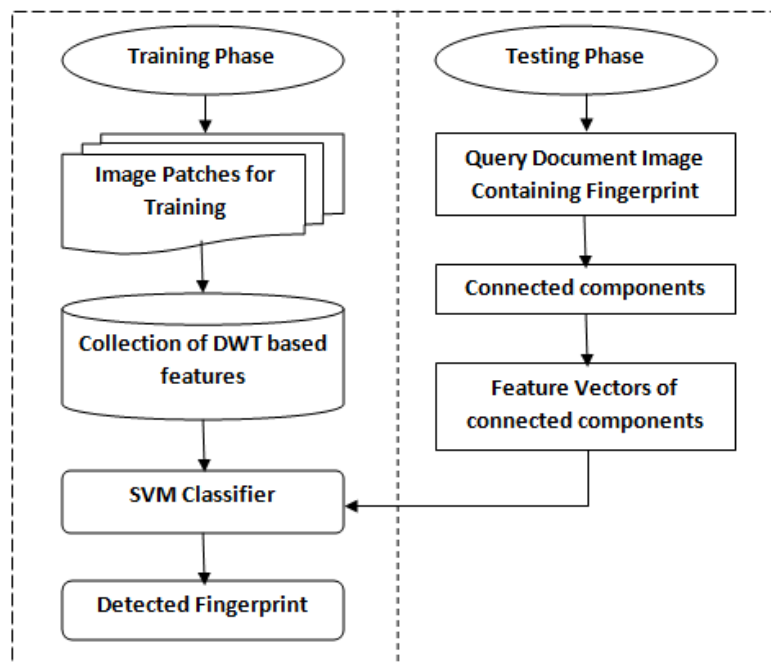


Fig. 5.2 Fingerprint Detection System

-
- **Feature Extraction:** To make the system simpler and faster, the DWT based features are used in the training phase. The 140 image patches that are considered for training the system are divided into four sub-bands by applying DWT. Let $C1(x,y)$, $C2(x,y)$, $C3(x,y)$ and $C4(x,y)$ are the resulted approximate, horizontal, vertical and diagonal sub-bands respectively. The energy and standard deviation of these sub-bands computed using equation (5.1) and (5.2) are used to construct the feature vector.

$$E_i = \frac{1}{M \times N} \sum_{m=1}^N \sum_{n=1}^N |C_i(x,y)| \quad \text{for } i = 1 \text{ to } 4 \quad (5.1)$$

$$SD_i = \sqrt{\frac{1}{M \times N} \sum_{x=1}^N \sum_{y=1}^N (C_i(x,y) - \mu_{ci})^2} \quad \text{for } i = 1 \text{ to } 4 \quad (5.2)$$

Where,

- ‘ E_i ’ and ‘ SD_i ’ represent energy and standard deviation of i^{th} sub-band.
- ‘ M ’ and ‘ N ’ are the row and column dimension of sub-bands.
- μ_{ci} refers to the mean of i^{th} sub-band.

The final feature vector is formed using the equation (5.3)

$$FV = \{ E_1, E_2, E_3, E_4, SD_1, SD_2, SD_3, SD_4 \} \quad (5.3)$$

- **SVM Classifier:** A binary SVM classifier is used to detect whether an image patch has a fingerprint or not. The SVM uses a hyper-plane to separates the data points into two groups for classification. The optimal hyperplane is given by equation (5.4).

$$OHP = W^T \phi(x) + b \quad (5.4)$$

Where ‘ W ’ indicates the normal vector, ‘ b ’ indicates an offset vector of the hyperplane with the kernel function ‘ $\phi(x)$ ’. Fig. 5.3 shows an example classification of circles and squares using hyperplane of SVM classifier. Different Kernel functions used in SVM viz., polynomial, linear, Sigmoid and Gaussian. A linear kernel function is used in the proposed system. For fingerprint detection, the extracted DWT based features of 140 image patches are used to train the classifier. Out of 140 image patches, 42 comprise of fingerprints and the other 98 image patches comprise of a variety of data including text, tables, logos and some horizontal and vertical strips.

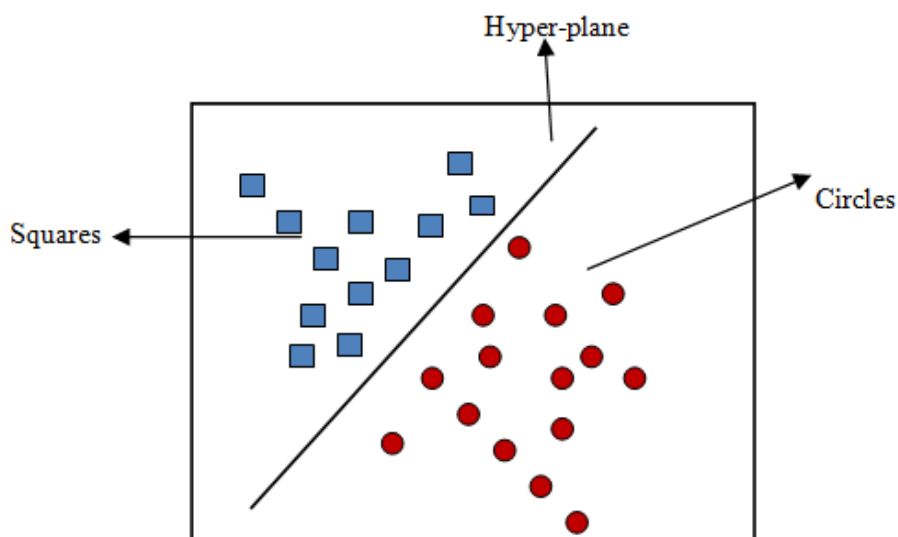


Fig. 5.3 Classification Example using SVM

The query document that consists of a fingerprint impression is given as an input to the system. The three sequences of steps namely binarization, morphological dilation and generating the image patches are used in preprocessing of the query document. An Otsu [80] gray-level thresholding method is used in the process of binarizing the image. Then the image patches are generated by employing the connected components [105] of the binary image. The small pieces of the image such as tiny text, fingerprint ridges, lines, curves and other nearby graphical elements are merged into a single patch by performing the morphological dilation on the image. The disk-shaped structuring element having a radius of 20 numbers of pixels is chosen in the dilation process. The shapes, as well as the size of the structuring element, are chosen empirically by conducting the experiments. Once the image patches are formed, their DWT based features are used as test features for SVM classifier to detect whether an image patch comprises of fingerprint or not. Finally by knowing the coordinates of the patch containing fingerprint is extracted from the document for further processing.

5.4.2 Computing Feature Vector for Document Retrieval

Two sets of feature extraction schemes (1) DWT based LBP features and (2) SWT based LBP features are proposed for fingerprint-based document retrieval. To construct these features the DWT/SWT is applied on the detected fingerprint image, LBP of each sub-bands are computed and finally, the histograms of these LBP features are merged to form the feature vector as shown in Fig. 5.4. These steps are described in the following sections.

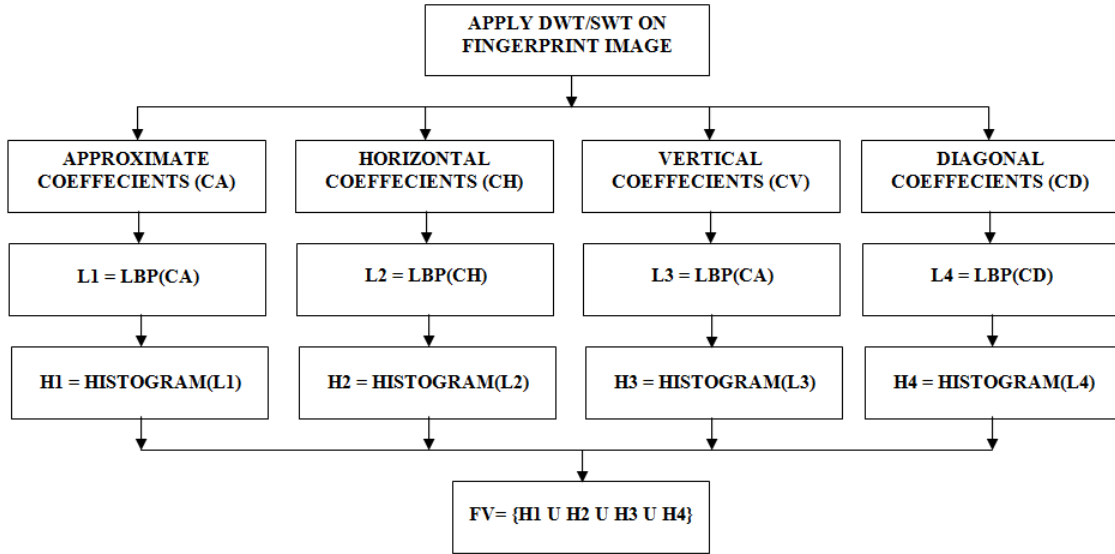


Fig. 5.4 Proposed Feature Extraction Scheme(s)

- Applying DWT/SWT:** The DWT and SWT are the most popular techniques employed in the multi-resolution analysis of an image. The DWT/SWT are obtained by using a series of low pass and high pass filters on row-wise and column-wise down-sampled image [91]. Application of DWT leads to the decomposition of an image of size ‘M’ rows and ‘N’ columns into four sub-bands namely: approximate, horizontal, vertical and the diagonal. The dimensions of all the sub-bands will be $M/2 \times N/2$. The SWT is similar to DWT except that the down-sampling process is eliminated. Due to this, the decomposed sub-bands of SWT will have the same size as that of the original.

Let $fpimg(x,y)$ is a two-dimensional function that represents extracted fingerprint from the query document. Equations (5.5) and (5.6) are used for obtaining the sub-bands by applying DWT/SWT.

$$[CA, CH, CV, CD] = DWT(fpimg(x, y)) \quad (5.5)$$

$$[CA, CH, CV, CD] = SWT(fpimg(x, y)) \quad (5.6)$$

The ‘CA’, ‘CH’, ‘CV’ and ‘CD’ in the equations represent approximate, horizontal, vertical and diagonal coefficients of a fingerprint image.

- Computation of LBP:** The LBP is computed on each 3×3 cells of ‘CA’, ‘CH’, ‘CV’ and ‘CD’ to obtain the texture information. “LBP codes are generated by multiplying

the binary threshold values assigned to each of the pixels of $N \times N$ cell and summing up the result [106]”. Algorithm 5.1 shows the steps used for obtaining LBP code for a cell.

Algorithm 5.1: Computation of LBP code for a 3×3 Cell

1. Begin
 - Input:** 3×3 Cell
 - Output:** 8-bit LBP Value
2. Let $C = f(x, y)$ is center pixel of 3×3 Cell and $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8$ are 8-neighbors of C .
3. For $i=1$ to 8 repeat
 - If $C_i > C$ then
 - $C_i = 1$
 - Else
 - $C_i = 0$
 - Endif
- Endfor
4. $LBP(x,y) = C_1 \times 2^7 + C_2 \times 2^6 + C_3 \times 2^5 + C_4 \times 2^4 + C_5 \times 2^3 + C_6 \times 2^2 + C_7 \times 2^1 + C_8 \times 2^0$
5. End

Let ‘ L_1 ’, ‘ L_2 ’, ‘ L_3 ’ and ‘ L_4 ’ are the LBP of the sub-bands ‘CA’, ‘CH’, ‘CV’ and ‘CD’ obtained using the Algorithm 5.1.

- **Formation of Feature Vector:** The histograms are used in order to reduce the number of LBP codes computed for sub-bands of an image. The histogram is a discrete function that gives the occurrence of particular data values that fall into disjoint categories called bins. Let ‘ H_1 ’, ‘ H_2 ’, ‘ H_3 ’ and ‘ H_4 ’ are the histogram values computed for the sub-bands ‘ L_1 ’, ‘ L_2 ’, ‘ L_3 ’ and ‘ L_4 ’ each with 64 numbers of bins. Finally these histograms are merged to generate a final feature vector ‘FV’ of size 256 as shown in equation (5.7).

$$FV = \{ \{H1\} \cup \{H2\} \cup \{H3\} \cup \{H4\} \} \quad (5.7)$$

5.4.3 Fingerprint Matching and Document Retrieval

The purpose of this step is to match the features of query fingerprint impression with the features of fingerprints belonging to a database of documents and retrieve more relevant documents. The Euclidean distance is widely used in the literature for matching the features. The problem with normal Euclidean distance is that the features with larger values dominate and the features with lower values contribute very less in the resulted similarity values. This particular issue is addressed by standardized Euclidean distance where all the features contribute almost equal in the estimation of similarity value. Hence the standardized Euclidean distance is employed in the proposed method to compute the similarity. Equation (5.8) is used for computing the similarity metric.

$$StdEucDist(i) = \sqrt{(FVQ - FDB) \times V^{-1} \times (FVQ - FDB)^T}, \text{ for } i = 1 \text{ to } N \quad (5.8)$$

Where,

- ‘StdEucDist’ is a vector to hold computed similarity values for N number of documents.
- ‘FVQ’ is the feature vector of the query document image.
- ‘FDB’ is pre-extracted fingerprint features of documents stored in the database.
- ‘V’ is the n -by- n diagonal matrix whose j^{th} diagonal element is $S(j)^2$ and ‘S’ being the vector of standard deviations.

After computing the distance values, the documents present in the database are indexed based on their distance values with respect to the query document. Now based on the user request top ‘K’ number of documents are accessed from the database and displayed on the user console. Algorithm 5.2 provides the complete list of steps used for fingerprint-based document image retrieval.

Algorithm 5.2: Fingerprint-based Document Image Retrieval.

1. Begin

Input: FDB = Features of extracted fingerprints from document images stored in the database.

fpimg(x,y) = Extracted fingerprint image.

Output: FVQ = Fingerprint feature vector of query document image

D[1:N] = Top N Retrieved Document Images

2. Preprocess the image

a. Color to grayscale conversion

b. Resize the image to 256×256

3. /* Feature extraction from fingerprint image */

a. Apply DWT/SWT on fpimg(x,y) and decompose into approximate (CA), Horizontal (CH), Vertical (CV) and Diagonal (CD) coefficients.

b. Compute LBP of CA, CH, CV and CD

L1=LBP(CA)

L2=LBP(CH)

L3=LBP(CV)

L4=LBP(CD)

c. Obtain histogram features of L1, L2, L3 and L4 with 64 bins.

H1=Histogram(L1)

H2=Histogram(L2)

H3=Histogram(L3)

H4=Histogram(L4)

d. Concatenate H1, H2, H3 and H4 to get a final feature vector 'FV'.

FV = {H1 U H2 U H3 U H4}

4. Perform matching between query fingerprint features FVQ and indexed database fingerprint features FDB using standardized Euclidean distance.

5. Rank the documents based on distance value and retrieve top 'K' documents from the database.

6. End

5.5 Experimental results

The database used for testing the proposed fingerprint detection and document retrieval is explained in the below section. The performance evaluation of the system is provided in the subsequent sections.

5.5.1 Image Database

A total of 50 members are considered for database creation. The 24 left thumb impressions per member on the variety of documents using a blue colored ink pad are collected. Thus a total of $50 \times 24 = 1200$ documents are generated. These documents are then scanned with a resolution of 300×300 using HP M1005 scanner.

5.5.2 Performance of Fingerprint Detection System

The database was having 1200 documents out of which 100 are used for training the system and the other 1100 documents are used for testing. Fig. 5.5 depicts two sample results of fingerprint detection module. The detected fingerprint is marked with a red colored rectangle box.

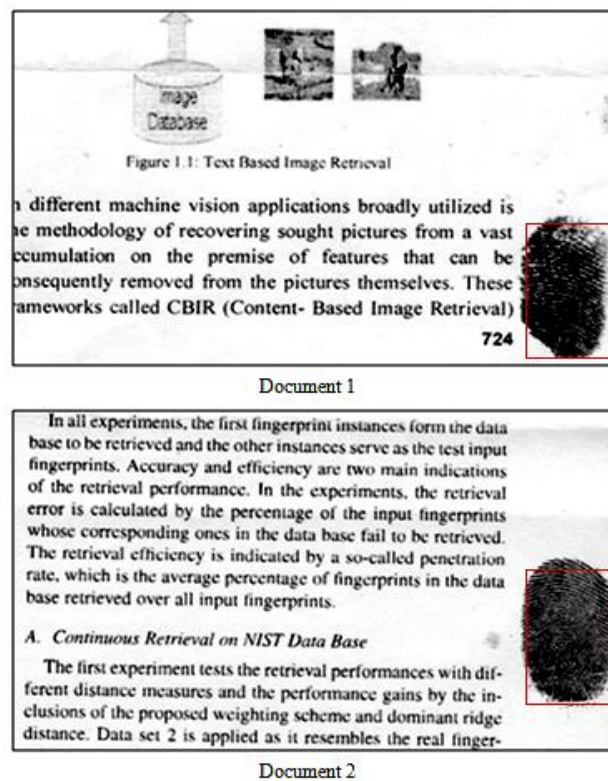


Fig. 5.5 Sample Results of Fingerprint Detection

An image patch detected with at least 80% of the fingerprint impression is treated as successful detection during the evaluation of the proposed scheme. The detection rate is used to estimate the performance of the method. It is defined as the ratio of a number of documents in which fingerprints are successfully detected to the total number of documents considered for the experiment. The equation (5.9) is used for computation of detection rate.

$$Detection\ Rate = \frac{Number\ of\ documents\ with\ successful\ detection}{Total\ number\ of\ documents} \quad (5.9)$$

The proposed fingerprint detection method provided the detection rate of 98.87%, which makes more suitable to be used as the first step of fingerprint-based document retrieval.

5.5.3 Performance of Fingerprint-based Document Retrieval

Core-i3/4GB RAM/windows8 machine with MATLAB is used for the implementation of the proposed algorithm. It took 1.015 seconds for detection of fingerprint and 1.1725 seconds for document retrieval of top 20 documents.

The precision and recall are used for evaluating the performance of the proposed fingerprint-based document retrieval algorithm. An exhaustive set of experiments are carried out for evaluation of the proposed system. In each experiment, the user is asked to choose a query document randomly from each class. The precision and recall values are computed for retrieval of top 1, top 5, top 8, top 10, top 15 and top 20 documents and tabulated. A total of 300 queries are executed for assessing the performance of the developed algorithm. The average precision (AP) and average recall (AR) for 'N' number of queries are also computed for tabulated results using equation (5.10) and (5.11).

$$Average\ Precision\ (AP) = \frac{1}{N} \sum_{i=1}^N Precision_i \quad (5.10)$$

$$Average\ Recall\ (AR) = \frac{1}{N} \sum_{i=1}^N Recall_i \quad (5.11)$$

Fig. 5.6 shows a sample result with a query document and the top four retrieved documents using the proposed system. The mean of average precision (MAP) and mean of average recall (MAR) are also used for evaluation and comparison of the proposed feature extraction scheme with other methods. Table 5.1 lists out AP, AR, MAP and MAR values computed using proposed two sets of features also with conventional LBP and Histogram of Oriented Gradient (HOG) features.

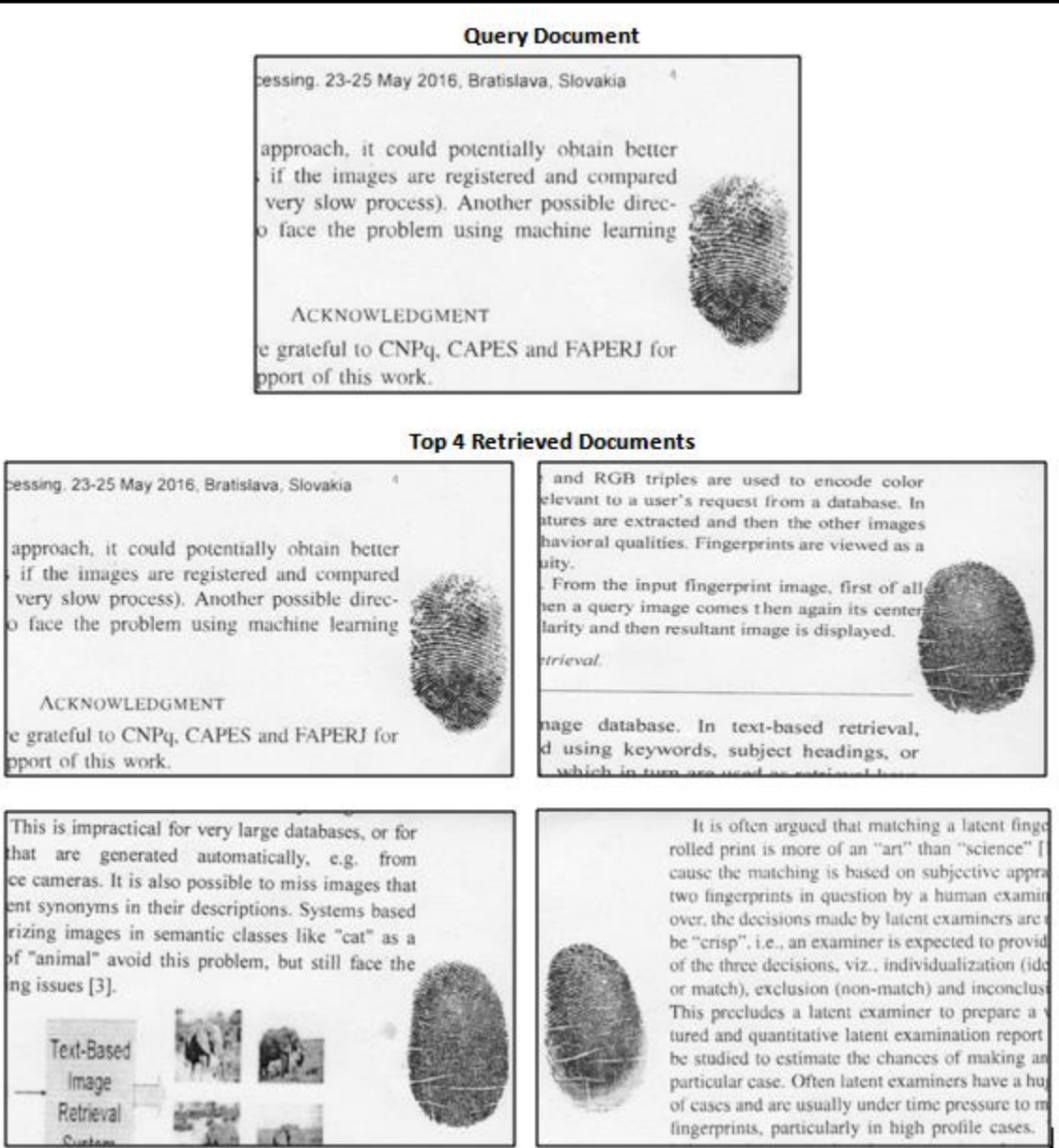


Fig. 5.6 Sample Result of Fingerprint-based Document Retrieval

Table 5.1 Average Precision and Recall

Number of Top Matches	HOG Features [38]		LBP Features [39]		Proposed Method-I (SWT based LBP Features)		Proposed Method-II (DWT based LBP Features)	
	Average Precision	Average Recall	Average Precision	Average Recall	Average Precision	Average Recall	Average Precision	Average Recall
TOP 1	100	4.16	100	4.16	100	4.16	100	4.16
TOP 5	76.34	15.97	75.66	15.85	78.26	16.3	86.96	18.16
TOP 8	68.11	22.56	66.85	22.28	69.02	23.00	72.83	24.28
TOP 10	65.60	27.37	62.6	26.09	65.56	27.35	68.7	28.62
TOP 15	54.38	33.67	53.33	33.33	54.49	34.05	58.26	36.41
TOP 20	47.19	39.14	46.52	38.77	47.39	39.39	51.74	43.11
MAP/MAR	68.60	23.80	67.49	23.41	69.12	24.04	73.08	25.79

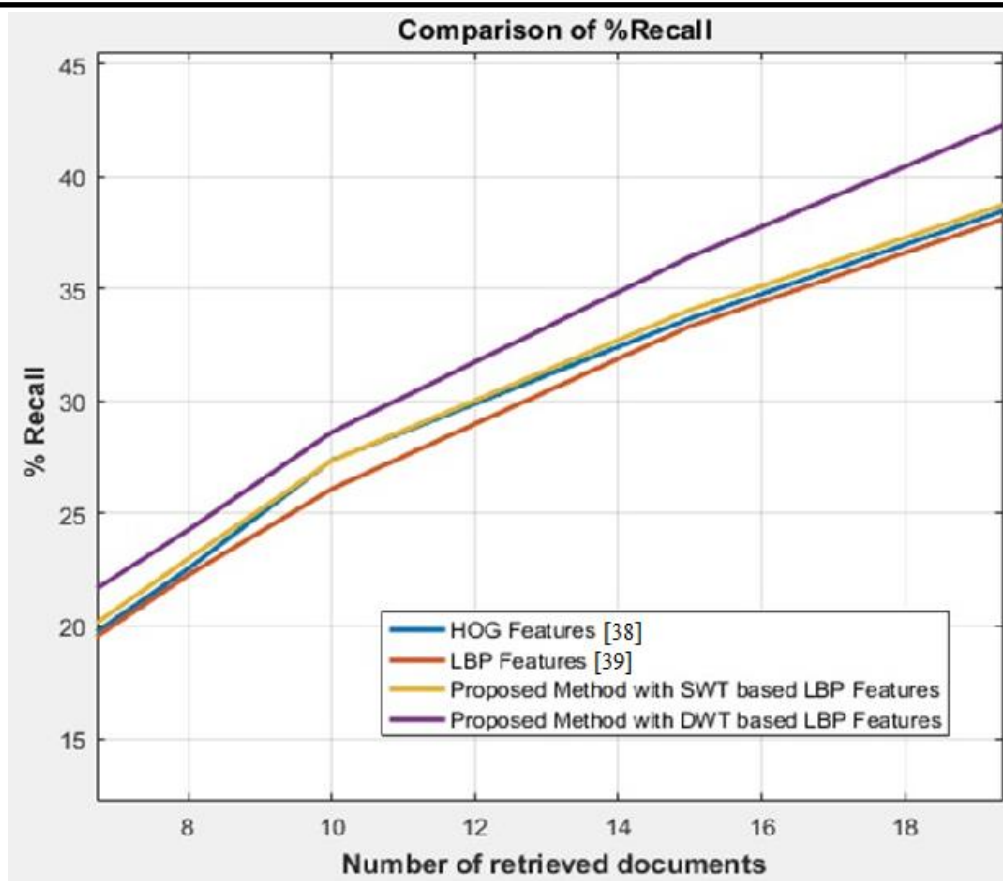


Fig. 5.7 Comparison of Average Recall

Fig. 5.7 provides a graphical comparison of average recall values obtained using proposed features with another set of features. The results are compared for retrieval of top 5, top 8, top 10, top 15 and top 20 documents from the database. The following observations can be made by comparison of the results.

- Both the proposed feature extraction schemes DWT and SWT based features provide better performance compared to existing feature extraction methods.
- The retrieval performance with proposed features increases as the numbers of retrieved documents is increased. This helps in the retrieval of more relevant documents from a huge database. The improved performance is because of LBP features computed for directional components of the fingerprint image by applying DWT/SWT.

5.6 Summary and Conclusion

The fingerprint-based document image retrieval is an essential application for document images using fingerprint print impression for authentication. This chapter proposed two feature extraction schemes: DWT based LBP and SWT based LBP. The proposed feature extraction schemes provided better results compared to existing techniques [38] and [39].

The logo, face/photo, fingerprint-based document retrieval are useful applications to overcome language dependency. But due to globalization most of the organizations have accepted the use of documents with multiple document images. When a huge database of multi-language documents is present, there is a need for language-based classification of document images and retrieval techniques. This motivated to develop language-based document image classification and retrieval application in the next chapter.

Chapter 6

Language-based Document Image Classification and Retrieval

Abstract of the Chapter: Use of multi-lingual documents gives rise to the need for language-based classification and retrieval of document images. The language-based classification helps to sort out the huge number of documents automatically with less amount of time. Whereas language-based retrieval provides a way of searching the document images based on the language used in the query document. To make the system computationally cheap and faster the proposed system uses block-level processing of the documents. This chapter proposes the use of multi-resolution Histogram of Oriented Gradient (HOG) features for both the language-based classification and retrieval of documents. A total of 1006 document images of Kannada, Marathi, Telugu, Hindi and English are used for evaluating the classification performance. An average classification rate of 87.02% is achieved using the proposed method. Since, most of the states in India have agreed upon a tri-lingual system, the proposed language-based retrieval system is tested for the documents with the combination of languages such as Kannada-Hindi-English, Marathi-Hindi-English and Telugu-Hindi-English.

6.1 Introduction

Most of the countries agreed upon the use of the multi-lingual system. This has lead to the existence of a huge number of documents with different languages in many organizations and multi-national companies. To classify or searching of such documents there is a need for automatic classification and retrieval techniques. Hence this chapter proposes language-based classification and retrieval algorithms.

The language-based classification and retrieval system is also helpful in applications like the implementation of OCR in digital libraries, text to speech conversion, development of document content understanding and many more. The language/script identification algorithms can be broadly classified into two categories depending on whether they employ global analysis or local analysis of the document images. The global analysis deals with the processing of the document at page-level or block-level. However local analysis is concerned with the processing of the document at word or line-level. The taxonomy of language/script based classification system is depicted in Fig. 6.1.

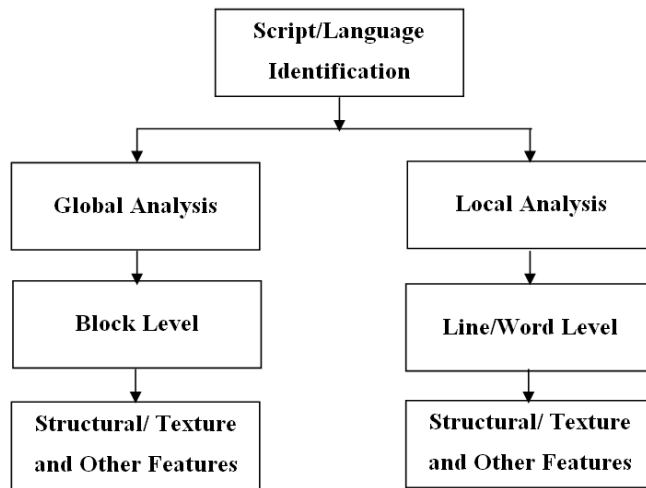
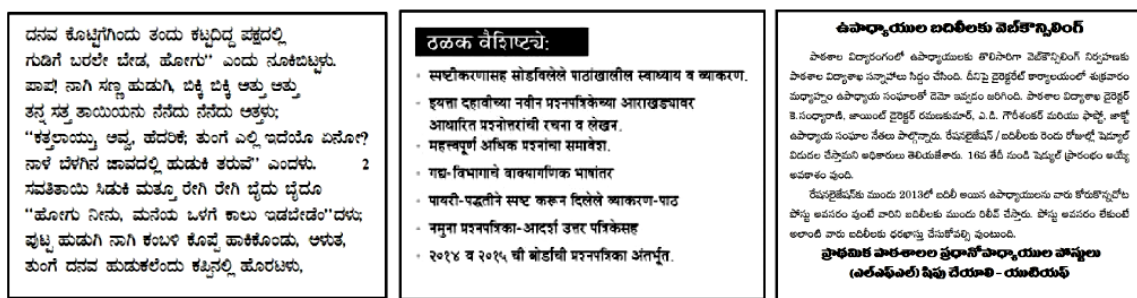


Fig. 6.1 Taxonomy of Language/Script Identification

The structural, texture or a hybrid set of features are applied on the block-level analysis of documents. However, the local analysis requires segmentation of the document into words or lines and then application of the features for analysis. As the global level analysis is segmentation free, it is computationally cheaper and faster compared to analysis. On the other side, the local analysis has an advantage of an accurate analysis of the documents which result in improved classification and retrieval performance.

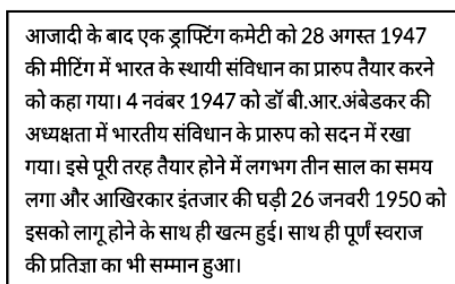
The main contribution of this chapter is proposing an efficient language-based classification and retrieval system using multi-resolution HOG features for printed document images of five different languages namely Kannada, Marathi, Telugu, Hindi and English. Fig. 6.2 shows an example of document images belonging to these five languages.



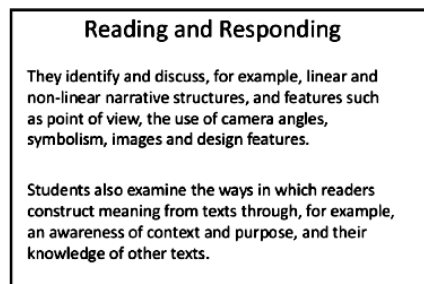
(a) Kannada Document

(b) Marathi Document

(c) Telugu Document



(d) Hindi Document



(e) English Document

Fig. 6.2 Sample Documents of Kannada, Marathi, Telugu, Hindi and English

The Kannada, Marathi, Telugu and Hindi are official languages of India. The Kannada and Telugu are originated from Brahmi alphabets and they are separated into two different sets between 12th and 15th century. There are 16 vowels and 34 consonants with more than 250 shapes. Telugu has 16 vowels, 3 vowel modifiers with 41 consonants. Hindi and Marathi language are originated from the Devanagari script. The Hindi includes 12 vowels and 34 consonants however Marathi has 16 vowels with 36 consonants. The horizontal line that connects the letters of a word is an important feature of Hindi and Marathi language scripts. English is a Latin-based language with 5 vowels and 21 consonants. Fig. 6.3 depicts vowels and consonants of all the five languages used for testing.

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ
a	ā	i	ī	u	ū	r̄	r̄̄
ಏ	ಐ	ಋ	ೠ	ಋ	ೠ	ಅಂ	ಅಃ
e	ē	ai	oi	o	ō	am	ah
ಕ	ಖ	ಗ	ಘ	ಢ			
ka	kha	ga	gha	ṅa			
ಚ	ಛ	ಜ	ಝ	ಞ			
ca	cha	ja	jha	ña			
ಟ	ಠ	ಡ	ಢ	ಣ			
ta	tha	da	dha	ṇa			
ತ	ಠ	ದ	ಢ	ನ			
ta	tha	da	dha	na			
ಪ	ಫ	ಬ	ಭ	ಮ			
pa	pha	ba	bha	ma			
ಯ	ರ	ಲ	ವ	ಲ			
ya	ra	la	va	la			
ಷ	ಸ	ಹ	ಝ	ರ			
ṣa	śa	śa	ha	ra			

(a) Kannada Vowels and Consonants

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ
a	ā	i	ī	u	ū	r̄	r̄̄
ಏ	ಐ	ಋ	ೠ	ಋ	ೠ	ಅಂ	ಅಃ
e	ē	ai	oi	o	ō	am	ah
ಕ	ಖ	ಗ	ಘ	ಢ			
ka	kha	ga	gha	ṅa			
ಚ	ಛ	ಜ	ಝ	ಞ			
ca	cha	ja	jha	ña			
ಟ	ಠ	ಡ	ಢ	ಣ			
ṭa	ṭha	ḍa	ḍha	ṇa			
ತ	ಠ	ದ	ಢ	ನ			
ta	tha	da	dha	na			
ಪ	ಫ	ಬ	ಭ	ಮ			
pa	pha	ba	bha	ma			
ಯ	ರ	ಲ	ವ	ಲ			
ya	ra	la	va	la			
ಷ	ಸ	ಹ	ಝ	ರ			
ṣa	śa	śa	ha	ra			

(b) Telugu Vowels and Consonants

अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	Initial Vowels
a	ā	i	ī	u	ū	r̥	e	ai	o	au	
[ə]	[a]	[i]	[iː]	[u]	[uː]	[r̥]	[e]	[əi]	[o]	[əu]	
क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	Velar and Palatal	
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña		
[kə]	[kʰə]	[gə]	[gʰə]	[ŋə]	[t͡sə]	[t͡ʃə]	[d͡ʒə]	[d͡ʒʰə]	[ɲə]		
ट	ठ	ड	ढ	ण	त	थ	द	ध	न	Retroflex and Dental	
ṭa	ṭha	ḍa	ḍha	ṇa	ta	tha	da	dha	na		
[ʈə]	[ʈʰə]	[ɖə]	[ɖʰə]	[ɳə]	[tə]	[tʰə]	[də]	[dʰə]	[nə]		
प	फ	ब	भ	म	य	र	ल	व	Labial and Semivowel		
pa	pha	ba	bha	ma	ya	ra	la	va			
[pə]	[pʰə]	[bə]	[bʰə]	[mə]	[jə]	[rə]	[lə]	[və/wə]			
श	ष	स	ह	ळ	क्ष	ज्ञ	श्च	Fricative, Retroflex Liquid and biconsonantal groups			
śa	ṣa	sa	ha	ḷa	kṣa	jña	śra				
[ʃə]	[ʃə]	[sə]	[ɦə]	[ɭə]	[kʂə]	[ɟnə]	[ʃrə]				

(c) Marathi Vowels and Consonants

Vowels and vowel diacritics														
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	अं	अः	अँ	
a	ā	i	ī	u	ū	r̥	e	ai	o	au	aṅ	aḥ	aṁ	
[ʌ]	[a]	[i]	[iː]	[u]	[uː]	[r̥]	[e]	[eɪ]	[o]	[əu]	[ɔ]	[ɔḥ]	[ɔṁ]	
प	पा	पि	पी	पु	पू	पृ	पे	पै	पो	पौ	पंपः	पाँ		
pa	pā	pī	pī	pu	pū	pr̥	pe	pai	po	pau	pāṁ	pāṁ		
[pə]	[pə]	[pə]	[pə]	[pə]	[pū]	[pr̥]	[pe]	[pai]	[po]	[pau]	[pə]	[pə]		
Consonants														
क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	ठ	ड	ढ	ण
ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña	ṭa	ṭha	ḍa	ḍha	ṇa
[kə]	[kʰə]	[gə]	[gʰə]	[ŋə]	[t͡sə]	[t͡ʃə]	[d͡ʒə]	[d͡ʒʰə]	[ɲə]	[ʈə]	[ʈʰə]	[ɖə]	[ɖʰə]	[ɳə]
त	थ	द	ध	न	प	फ	ब	भ	म	य	र	ल	व	
ta	tha	da	dha	na	pa	pha	ba	bha	ma	ya	ra	la	va	
[tə]	[tʰə]	[də]	[dʰə]	[nə]	[pə]	[pʰə]	[bə]	[bʰə]	[mə]	[jə]	[rə]	[lə]	[və]	
Additional consonants (used in loanwords from Persian, Arabic & English)														
श	ष	स	ह	क़	ख़	ग़	ज़	झ	फ़	ड़	ढ़			
śa	ṣa	sa	ha	qa	ḫa	ga	za	zha	fa	ra	ṛha			
[ʃə]	[ʃə]	[sə]	[ɦə]	[qə]	[xə]	[ɣə]	[zə]	[zə]	[fə]	[rə]	[r̥ʰə]			

(d) Hindi Vowels and Consonants

Vowels: A E I O U
Consonants: B C D F G H J K L M N P Q R S T U V W X Y Z
Vowels: a e i o u
Consonants: b c d e f g h i j k l m n o p q r s t u v w x y z

(e) English Vowels and Consonants

Fig. 6.3 Vowels and Consonants of Kannada, Telugu, Marathi, Hindi and English

The main contribution of this chapter is that it proposes a multi-resolution HOG features for language-based classification and retrieval of the document images. The proposed feature extraction schemes are found superior compared with the features used in [38], [39] and [40].

6.2 Problem Statement

The objective of this research work is to propose an efficient language-based classification and retrieval system. The language-based classification system aims to classify the document image automatically depending on the language used. However, the language-based retrieval aims to provide an easier and faster way of searching the documents based on the language used in the query document.

6.3 Related Work

Chaudhury et al. [107] developed a system for script identification of Indian languages. They employed Gabor filter-based features obtained from the connected components. The combination of different classifiers is suggested for the improvement of the results. Kulkarni et al. [108] presented the script identification technique based on visual features. A total of 8 visible features are used in combination with the Probabilistic Neural Network (PNN) in the proposed system. Padma and Vijaya [109] proposed the profile-based features and k-nearest neighbor classifier in their work of script identification for tri-lingual document images.

Pal and Chaudhuri [110] presented the method for recognition of English, Bangla, Arabic, Chinese and Devanagari script lines of a document image. They used the combination of shape, statistics and the water reservoir features. Rajput et al. [111] proposed the scheme for handwritten text identification with the help of DCT and the wavelet-based features. They employed global analysis with K-NN classifier. Mathematical and structure-based features along with a sequence of classifiers are applied for improving the results of script identification for Indian languages in [112].

Pardeshi et al. [113] employed multi-resolution spatial features for identification of the Indian scripts. The features are constructed using radon transform, DWT and DCT of the segmented words in their method. Tan et al. [114] proposed the use of word shape analysis for retrieving the text from the document images. Arani et al. [115] employed the Hidden Markov Model (HMM) for recognition of handwritten Farsi words. They used

Multi-Layer Perception (MLP) with an input of features obtained from image gradient, contour chain code and black-white transitions.

Djeddi et al. [116] proposed a method for writer recognition for handwritten document images of Greek and English languages. The run-length features with k-NN and SVM classifiers are used at the recognition phase. Roy et al. [117] presented HMM-based script identification for handwritten words of Indian languages. They employed the zone-wise segmentation of words to extract the features.

The method for document language detection was proposed by Spitz [118]. Initially, the script is categorized as Han-based or Latin-based employing upward concavities and later the shape-based features of the characters are used for deciding the language. Thanuja and Shreedevi [119] proposed the use of visual features obtained at word-level for recognition and retrieval of Kannada document images. Lakshmi [120] proposed a technique to identify the Telugu Palm Leaf characters. The 3D features of the characters are employed in their proposed method. Chandrakala [121] presented the recognition-free content-based document image retrieval for Kannada documents. They used a correlation technique for matching retrieval.

From the literature, it is learned that most of the language/script identification or retrieval systems focus on document analysis at word-level and line-level. The word-level and line-level analysis of the document images require segmentation of the document and provide better performance. The block-level analysis is segmentation free technique however provides little lower performance compared to word-level and line-level analysis techniques. But the carefully designed feature extraction scheme can improve the performance of document classification and retrieval system by adopting block-level processing. The main contribution of this work is presenting the use of multi-resolution HOG features for language-based classification and retrieval of the document images. Since these features are adopted at block level the system is computationally cheap and faster.

6.4 Proposed Language-based Classification System

Fig. 6.4 shows the proposed language-based classification system for document images with training and the testing phase. Preprocessing of the document, feature extraction and SVM classifier are the important blocks in the system architecture. These blocks are explained in the subsequent sections.

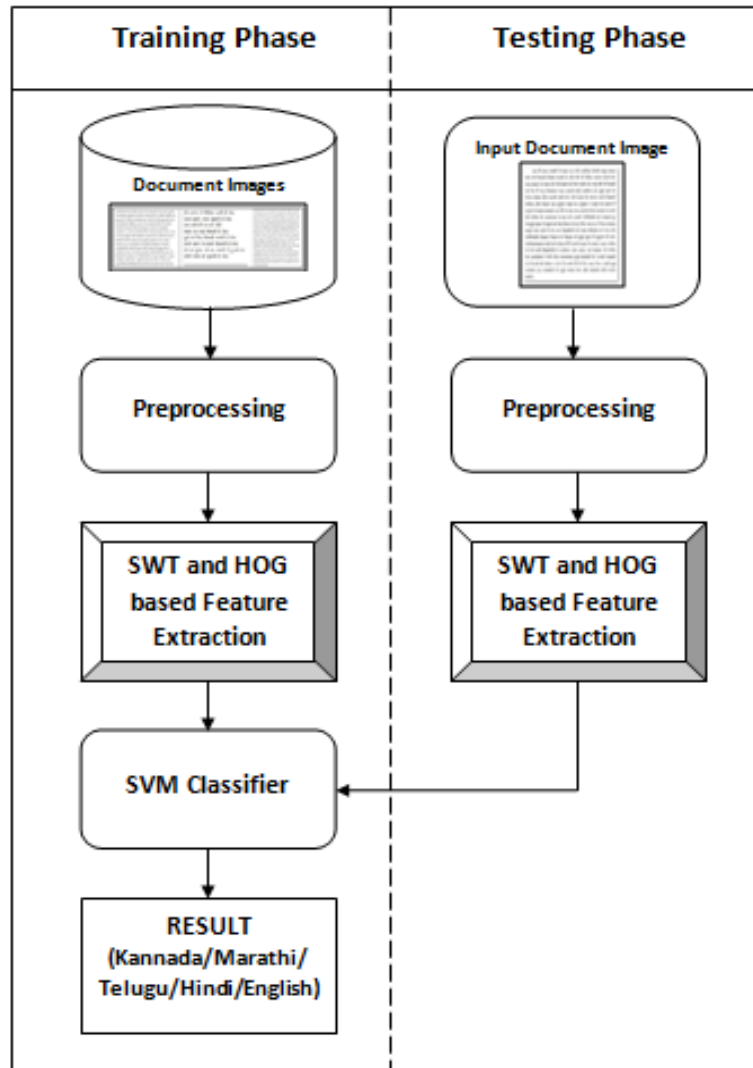


Fig. 6.4 Proposed Language-based Document Classification System

6.4.1 Preprocessing

The goal of this step is making the document image suitable for feature extraction. Initially, the documents are checked whether they are in color or grayscale and converted into grayscale if found in color. The equation (6.1) is used for conversion of RGB color image into the grayscale format. Let 'I' is the converted grayscale image.

$$I = 0.2989 \times R + 0.5870 \times G + 0.114 \times B \quad (6.1)$$

The 'R', 'G' and 'B' in the equation are intensity values of red, green and blue colors of pixels. The low contrast images cause poor classification performance due to the loss of some important features. Hence in the proposed method, to improve the quality of the image, it is passed through a low-pass filter followed by un-sharp masking with the help of a 3×3 mask shown in Fig. 6.5.

1	-1	-1
-1	8	-1
-1	-1	-1

Fig. 6.5 Un-sharp Mask

The algorithm 6.1 shows all the steps used for preprocessing of the document image in the proposed method.

Algorithm 6.1: Preprocessing of the Document	
1.	Begin Input: Document image Output: Preprocessed document image $D(x,y)$.
2.	Read the input document image.
3.	if input is color image Convert to gray-scale using equation (1). end if
4.	Perform un-sharp masking using mask shown in Fig. 6.3.
5.	Apply low-pass filter.
6.	End

6.4.2 SWT based Multi-resolution HOG Feature Extraction

The proposed system employs SWT based multi-resolution HOG features for language-based classification of documents. These features are obtained by computing HOG of the four decomposed coefficient matrices generated by applying SWT. Algorithm 6.2 provides the steps used for obtaining SWT based multi-resolution HOG features. To exploit the translation invariance property of the SWT, the proposed method employs SWT based multi-resolution HOG features. Let $D(x,y)$ is a resized version of the preprocessed document with a dimension of 256×256 pixels. The application of SWT on the image $D(x,y)$ results into four coefficient matrices namely D_{LL} , D_{LH} , D_{HL} and D_{HH} as given by equation (6.2).

$$SWT\{D(x,y)\} = \{D_{LL}, D_{LH}, D_{HL}, D_{HH}\} \quad (6.2)$$

Algorithm 6.2: Extraction of Multi-resolution HOG Features

1. Begin
 - Input:** Pre-processed document image $D(x,y)$
 - Output:** Feature vector (FV)
 2. Resize the document image $D(x,y)$ to 256×256 pixels.
 3. Apply stationary wavelet transform on the image $D(x,y)$.

$$[D_{LL}, D_{LH}, D_{HL}, D_{HH}] = \text{SWT} \{D(x,y)\}$$
 4. Extract HOG features from D_{LL}, D_{LH}, D_{HL} and D_{HH} .
 - a. $H_1 = \text{HOG}(D_{LL})$
 - b. $H_2 = \text{HOG}(D_{HL})$
 - c. $H_3 = \text{HOG}(D_{LH})$
 - d. $H_4 = \text{HOG}(D_{HH})$
 5. Concatenate H_1, H_2, H_3 and H_4 to construct feature vector.
 6. $\text{FV} = \{ H_1 \cup H_2 \cup H_3 \cup H_4 \}$
 7. End
-

The matrix D_{LL} represents approximation, D_{LH} horizontal, D_{HL} vertical and D_{HH} represent diagonal coefficient matrices. These four matrices are used for obtaining HOG features.

The use of HOG features was first proposed for detecting human faces by Dalal and Triggs [38]. The HOG with its various flavors is used in many applications such as hand detection, Pedestrian detection, face recognition, etc. In the proposed algorithm, the four coefficient matrices obtained using SWT are sub-divided into blocks with 2×2 cells. A total of 16 cells are formed out of four coefficient matrices. The dimension of the cells influences the performance of HOG features. Therefore the proposed system uses cells with a dimension of 128×128 pixels as well as 64×64 pixels for testing. The cells of size 128×128 yield 144 features whereas 64×64 yields 1296 features. After sub-dividing the blocks into cells, their gradient and the orientation are computed. Mathematically, the gradient of a 2D function $f(x,y)$ is given by equation (6.3).

$$\begin{bmatrix} \nabla x \\ \nabla y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} H_G(x,y) \\ V_G(x,y) \end{bmatrix} \quad (6.3)$$

Where ∇x is horizontal gradient and ∇y is vertical gradient values represented as $H_G(x,y)$ and $V_G(x,y)$ respectively. However, the gradient values in the image analysis applications are approximated to the difference between two successive pixels. The difference of

successive pixel intensities in a row provide horizontal gradient and the difference of pixel intensities in a column provide vertical gradient. Equations (6.4) and (6.5) are used for computation of horizontal and vertical gradient values in the proposed system.

$$H_G(x, y) = D_i(x + 1, y) - D_i(x - 1, y) \quad (6.4)$$

$$V_G(x, y) = D_i(x, y + 1) - D_i(x, y - 1) \quad (6.5)$$

Equations (6.6) and (6.7) are used to compute magnitude $M(x,y)$ and the direction of gradients $\Theta(x,y)$.

$$Mag(x, y) = \sqrt{H_G(x, y)^2 + V_G(x, y)^2} \quad (6.6)$$

$$\theta(x, y) = \tan^{-1} \frac{G_H(x, y)}{G_V(x, y)} \quad (6.7)$$

The orientation of gradient values computed for pixels of all the cells is used to construct the histograms. Let the H_1 , H_2 , H_3 and H_4 represent the histogram of oriented gradients of the four coefficient matrices D_{LL} , D_{LH} , D_{HL} and D_{HH} respectively. The final feature vector is obtained by merging the four histograms H_1 , H_2 , H_3 and H_4 by performing union operation as shown in equation (6.8).

$$FV = \{H_1 U H_2 U H_3 U H_4\} \quad (6.8)$$

Every histogram requires a range of distinct values known as bins. The proposed system uses histogram with 9 bins per cell to hold gradients values. As each coefficient matrix is sub-divided into four cells, $9 \times 4 = 36$ number of features is generated per matrix of coefficients. Thus a total of $36 \times 4 = 144$ number of features for each image is obtained for classification. Fig. 6.6 depicts a plot of SWT based multi-resolution HOG features obtained for a sample document image.

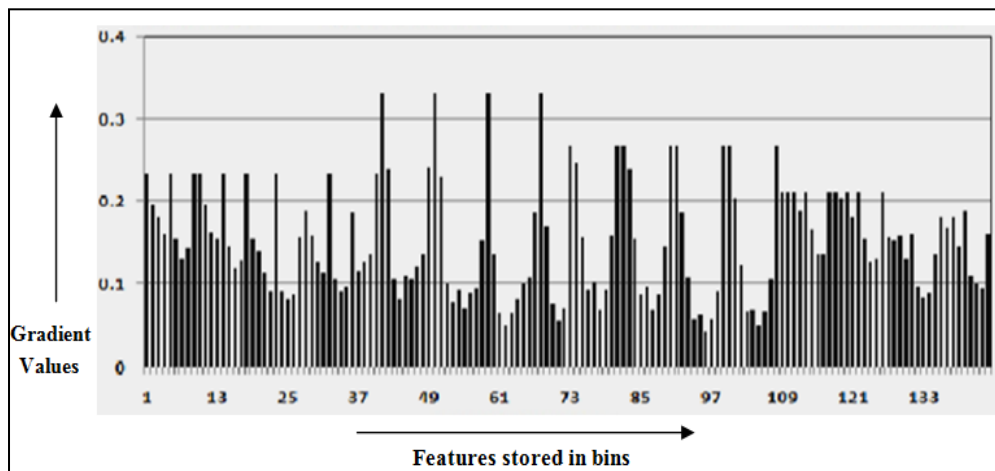


Fig. 6.6 SWT based Multi-resolution HOG Features of a Sample Document Image

6.4.3 SVM Classifier

As SVM provides high accuracy even for small training data set, the proposed system employed SVM for classification. It is trained using multi-resolution HOG features of 30% of the documents and the remaining 70% of the documents are used for testing.

6.5 Experimental Results of Language-based Classification

The proposed database used for language-based classification is discussed below and the performance of the system is provided in the subsequent section.

6.5.1 Image Database

The proposed language-based classification system is tested on a database of 1006 document images belonging to Kannada, Marathi, Telugu, Hindi and English languages. The documents are collected from the textbook, newspapers and the internet. The documents in the database comprise of printed text and a variety of graphics with varying resolution as well as size. Table 6.1 provides the details of the number of documents used from different languages.

Table 6.1 Details of the Database

Sl. No	Language	Documents
1	Kannada	197
2	Marathi	184
3	Telugu	198
4	Hindi	216
5	English	211
Total number of document images		1006

6.5.2 Performance of Language-based Classification System

Language detection rate is used as an assessment parameter for evaluating the performance of classification. It is computed as the ratio of a number of documents correctly classified to the total number of documents used for evaluation. Equation (6.9) is used for computing the language detection rate.

$$\text{Language detection rate} = \frac{\text{Number of document correctly classified}}{\text{Total number of documents}} \quad (6.9)$$

The average detection rate is also used for comparing the overall performance of the methods and is computed using equation (6.10).

$$\text{Average detection rate} = \frac{1}{N} \sum \text{Language detection rate} \quad (6.10)$$

The ‘N’ in the above equation represents the number of languages considered for evaluation of the system. The results obtained using the proposed algorithm is compared with feature extraction techniques used in [38], [39] and [40]. Table 6.2 provides the details of feature extractions schemes with the length of the feature vector used for comparison of the language-based classification system.

Table 6.2 Details of Feature Extraction Schemes

Sl. No.	Feature extraction method	Size of the feature vector
1	Rotation Invariant LBP (RILBP) [39]	640
2	HOG features [38]	324
3	Multi-resolution LBP features [40]	256
3	Proposed features with 64×64 cell size	1296
4	Proposed features with 128×128 cell size	144

Table 6.3 and 6.4 provide the results of classification using k-NN and SVM classifier respectively. The tabulated results show that the proposed feature extraction method provides promising results using both the classifiers. In particular, the proposed algorithm using cells with size 128×128 provided better results compared with 64×64 size.

Table 6.3 Comparison of Results using K-NN Classifier

Sl. No.	Language	Classification Accuracy (%)				
		RILBP [39]	HOG [38]	Multi-resolution LBP features [40]	Proposed Features with 64×64 cell	Proposed Features with 128×128 cell
1	Kannada	50	57.29	63.25	59.29	61.45
2	Marathi	56.31	77.49	81.43	68.93	81.55
3	Telugu	62.5	66.2	75.73	87.5	79.12
4	Hindi	98.4	94.11	97.97	99.49	100
5	English	84.82	83.33	77.78	82.2	88.48
Avg. Classification accuracy		70.406	75.684	79.232	79.482	82.12

Table 6.4 Comparison of Results using SVM Classifier

Sl. No.	Language	Classification Accuracy (%)				
		RILBP [39]	HOG [38]	Multi-resolution LBP features [40]	Proposed Features with 64×64 cell	Proposed Features with 128×128 cell
1	Kannada	55.2	65.23	67.7	76.04	75
2	Marathi	62.135	73.28	73.78	85.44	87.38
3	Telugu	72.22	86.11	91.66	80.56	80.58
4	Hindi	98.99	95.23	97.46	100	100
5	English	86.44	79.56	80.126	81.15	92.15
Avg. Classification accuracy		74.997	79.882	82.1452	84.638	87.022

Fig. 6.7 depicts a graphical comparison of the results obtained using various feature extraction methods with K-NN and SVM classifiers. Graphs are plotted considering average detection rate versus feature extraction techniques.

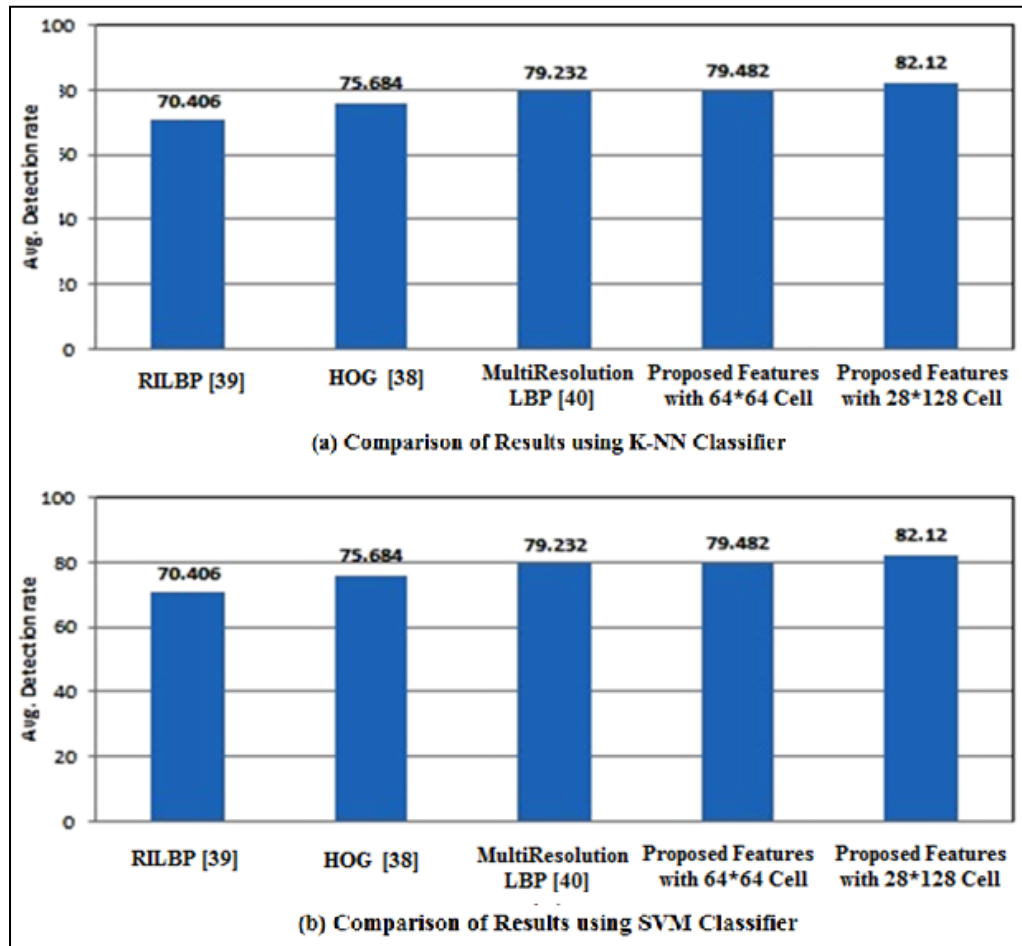


Fig. 6.7 Graphical Comparison of Results

The comparison of language-based classification results shows that,

- Proposed feature extraction scheme outperform with existing feature extraction schemes.

- The multi-resolution HOG features extracted using cells of the size of 128×128 (144 features) provided higher performance.
- SVM classifier outperforms in comparison with k-NN classifier irrespective of the features used for this particular application.

6.6 Proposed Language-based Document Image Retrieval System

The proposed language-based document retrieval employs the global approach for feature extraction. The scripts used by different languages are unique in nature and form visually distinctive features. The texture features obtained using multi-resolution HOG features are proposed for document image matching and retrieval. Fig. 6.8 depicts the building blocks such as preprocessing, feature extraction and similarity matching for language-based document retrieval system. The preprocessing operations used here are the same as that of the proposed language-based classification system explained in section 6.5.1. The remaining blocks feature extraction and similarity matching are explained in the following sub-sections.

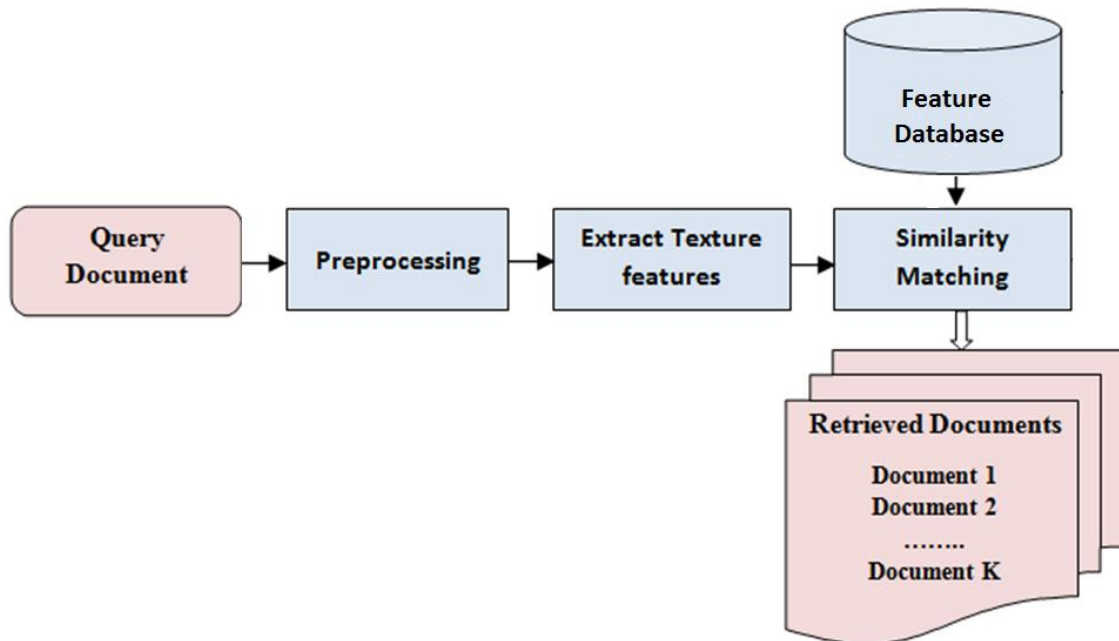


Fig. 6.8 Building Blocks of Proposed Language-based Document Retrieval System

6.6.1 Feature Extraction

The DWT based multi-resolution HOG features are used for language-based document image retrieval. Fig. 6.9 shows the conceptual view of feature extraction scheme used in the proposed method.

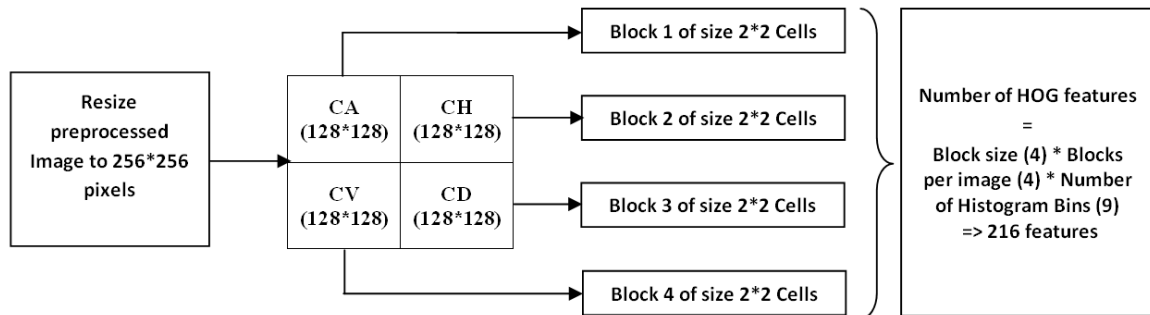


Fig. 6.9 Proposed DWT based Multi-resolution HOG Feature Extraction

The general method for obtaining HOG features includes division of the image into cells of suitable size, arranging these cells into blocks, computing the histogram of oriented gradients of the blocks and merging of the histograms to construct a feature vector. The size of the HOG feature vector is given by equation (6.11). In the equation, ‘BS’ represents block size, ‘NBPI’ is a number of overlapping blocks of the image and ‘NBINS’ represents the number of bins of the histogram.

$$\text{Size of HOG feature vector} = BS \times NBPI \times NBINS \quad (6.11)$$

Algorithm 6.3 provides the details of the steps used in the proposed feature extraction scheme.

Algorithm 6.3: Multi-resolution HOG Feature Extraction using DWT

1. Begin

Input: Preprocessed document image $D_p(x,y)$,

Output: Multi-resolution HOG features (HFV)

2. Resize $D_p(x,y)$ to 256×256 pixels.

3. Apply DWT on $D_p(x,y)$ to obtain approximate (CA), Horizontal (CH), Vertical (CV) and Diagonal (CD) coefficients. This results in 4 sub-bands of size 128×128 .

4. Divide each sub-band into cells of size 64×64 and organize a group of four cells as a block leading to a total of 4 blocks per image.

5. Compute HOG features of each sub-band using 9 bins in the histogram. Equations (6.12) and (6.13) are used to compute the gradient of pixels in horizontal ($G_H(x,y)$) and vertical directions ($G_V(x,y)$). The magnitude $M(x,y)$ and the direction $\Theta(x,y)$ of gradient values are computed using equations (6.14) and (6.15) respectively.

$$G_H(x,y) = D_p(x+1,y) - D_p(x-1,y) \quad (6.12)$$

$$G_V(x,y) = D_p(x,y+1) - D_p(x,y-1) \quad (6.13)$$

$$M(x,y) = \sqrt{G_H(x,y)^2 + G_V(x,y)^2} \quad (6.14)$$

$$\Theta(x,y) = \tan^{-1} \frac{G_H(x,y)}{G_V(x,y)} \quad (6.15)$$

Let H_1, H_2, H_3 and H_4 are the histogram of oriented gradients computed on the sub-bands CA, CH, CV and CD respectively.

6. Combine H_1, H_2, H_3 and H_4 to form the final feature vector HFV using Equation (6.16).

$$HFV = \{H_1, H_2, H_3, H_4\} \quad (6.16)$$

7. End

6.6.2 Similarity Matching

The Canberra distance is found to provide better matching results when the texture features are used [122]. Hence the proposed system employed Canberra distance to compute the similarity between query document features 'FVQ' and features of documents 'FDB' that are stored in the database. Equation (6.17) is used for the computation of Canberra distance of 'N' features.

$$Canberra(FVQ, FDB) = \sum_{i=1}^N \frac{|FVQ_i - FDB_i|}{|FVQ_i| + |FDB_i|} \quad (6.17)$$

After computing the distance, the database documents are sorted based on the distance value such that documents with the lowest distance at the top and vice-versa. Finally, ‘K’ number of top matching documents are retrieved and displayed on the user console.

6.7 Experimental Results of Language-based Document Retrieval

A Core i3/4GB RAM/windows8 machine with MATLAB is used for the implementation of the proposed system. It took 0.5625 seconds of time for retrieving top 50 documents when the query is submitted. Every state of India has adopted a trilingual system which includes a local/regional language, national language Hindi and a global language English. Hence the proposed system is tested for a combination of three trilingual datasets. The details of the database are discussed in the following section and the performance of the retrieval is discussed in the subsequent section.

6.7.1 Image Database

Three datasets having a combination of Kannada-Hindi-English, Marathi-Hindi-English and Telugu-Hindi-English documents are used for testing of tri-lingual language-based document image retrieval. Table 6.5 provides the details of the test database.

Table 6.5 Details of 3 Datasets

Sl. No	Dataset	Number of Document images					
		Kannada	Marathi	Telugu	Hindi	English	Total
1	Dataset1	94	--	--	196	191	481
2	Dataset2	--	110	--	196	191	497
3	Dataset3	--	--	72	196	191	459
Total number of document images with different languages							1437

6.7.2 Performance of Language-based Retrieval System

The average precision is used for evaluating the performance of the proposed system. Average precision is computed for retrieval of top 10, top 20, top 30, top 40 and top 50 documents from each dataset. 10 queries from each language leading to 30 queries per dataset are used for testing. Thus in total 90 queries for three datasets are executed to

evaluate the performance. Fig. 6.10 shows sample result of Kannada document image retrieval using dataset1.

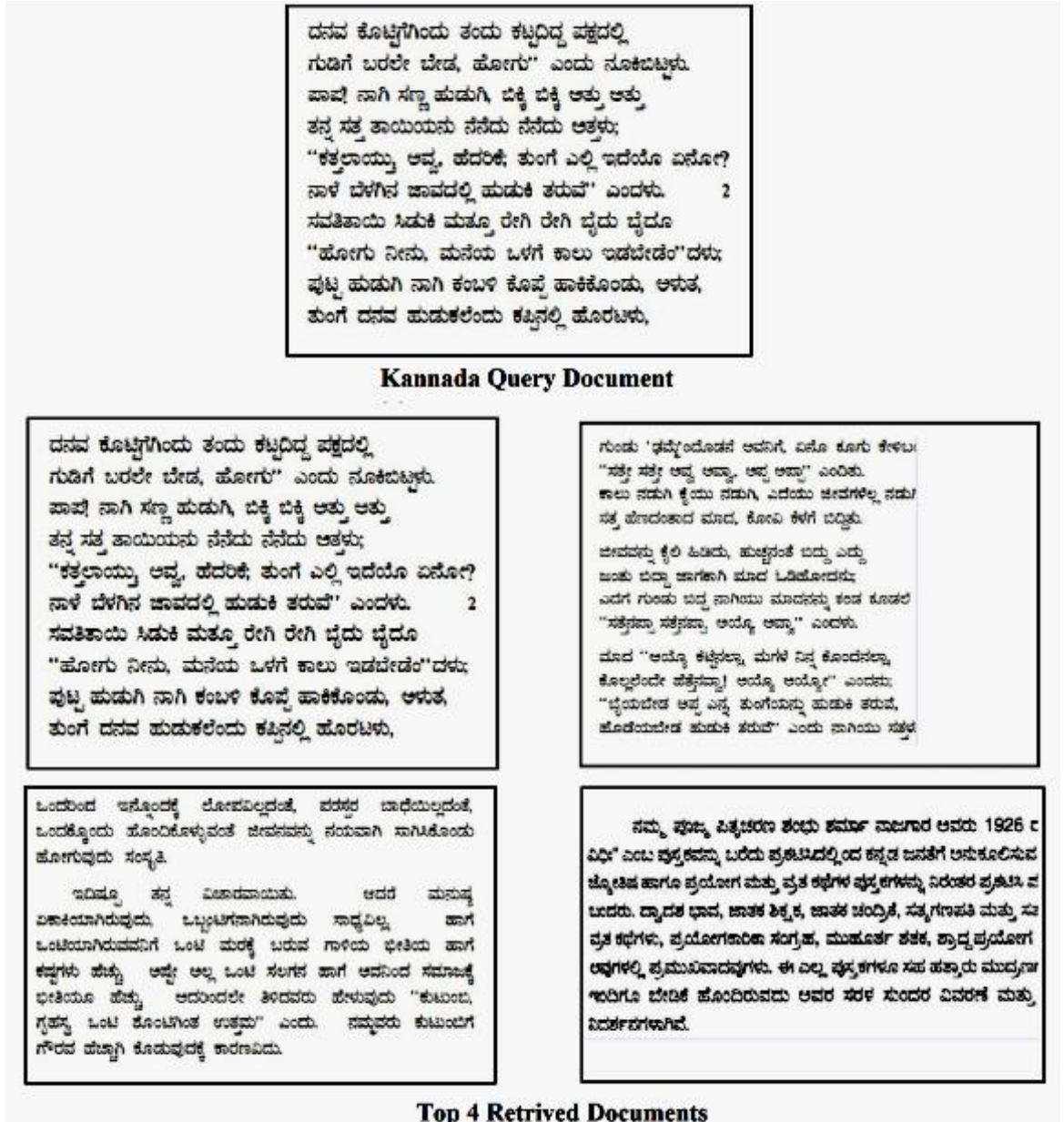


Fig. 6.10 Sample Result of Kannada Document Retrieval

Table 6.6 shows the results obtained for dataset1 using Rotation Invariant Local Binary Pattern (RILBP), HOG and proposed features. Fig. 6.11 shows the graphical comparison of the results.

Table 6.6 Average Precision for Dataset1

Number of top matches	AP using Rotation-invariant LBP (RILBP) Features [39]			AP using HOG Features [38]			AP with Proposed Method		
	Kannada	English	Hindi	Kannada	English	Hindi	Kannada	English	Hindi
10	86	96	98	88	91	99	99	100	100
20	68	97	97.5	81.5	78.5	87.5	97.5	100	100
30	60.67	93.66	94.33	75.33	80	89.33	97.33	99.75	100
40	57.25	93.5	94.75	73	76.75	87.25	96.25	99.75	99.75
50	53.4	86.8	88.6	72.4	69.2	82.6	96.2	99.6	99.8

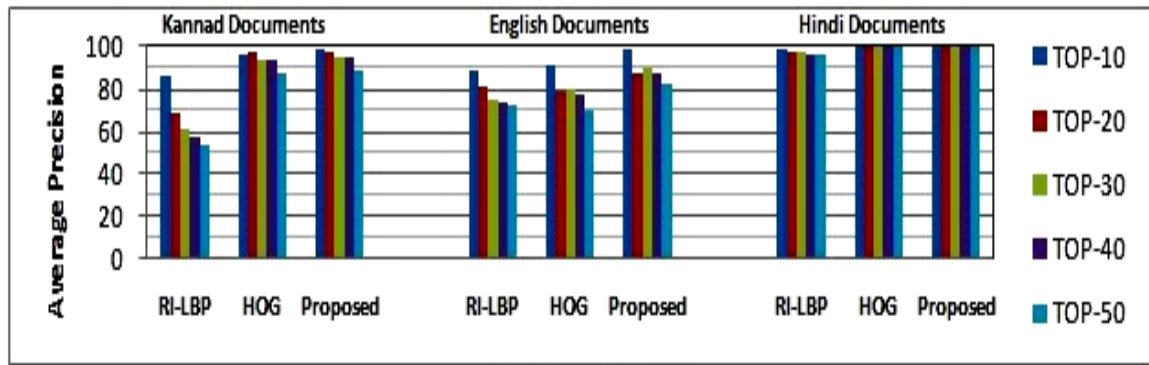


Fig. 6.11 Graphical Comparison of Results for Dataset1

Table 6.7 shows the results obtained for dataset2 and Fig. 6.12 shows the graphical comparison of the results.

Table 6.7 Average Precision for Dataset2

Number of top matches	AP using Rotation-invariant LBP (RILBP) Features [39]			AP using HOG Features [38]			AP with Proposed Method		
	Marathi	English	Hindi	Marathi	English	Hindi	Marathi	English	Hindi
10	78	82	86	88	94	100	97	100	100
20	71	76	82.5	81.5	88	93.5	97	100	100
30	66.33	73.33	79.33	77	82.67	91	96.67	100	100
40	64.25	71.5	75.75	75	77.5	87.25	96	99.75	99.75
50	61.8	69.4	72	73.4	75.4	83.14	95.4	99.4	99.8

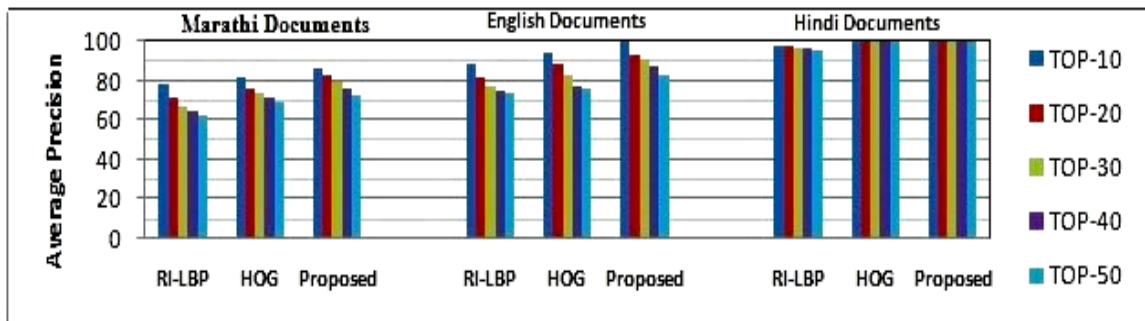


Fig. 6.12 Graphical Comparison of Results for Dataset2

Similarly Table 6.8 shows the results obtained for dataset3 and Fig. 6.13 depicts the graphical comparison of the results.

Table 6.8 Average Precision for Dataset3

Number of top matches	AP using Rotation-invariant LBP (RILBP) Features [39]			AP using HOG Features [38]			AP with Proposed Method		
	Telugu	English	Hindi	Telugu	English	Hindi	Telugu	English	Hindi
10	89	98.3	100	94	91	100	98	100	100
20	86.5	96	100	81	86	96.5	97	100	100
30	80.67	96	99.67	78.33	85.33	95.33	96.33	100	100
40	79.75	96	99.5	77	82	93	96.25	100	100
50	77	96	97.8	76.6	80.2	90.2	94.6	99.6	99.8

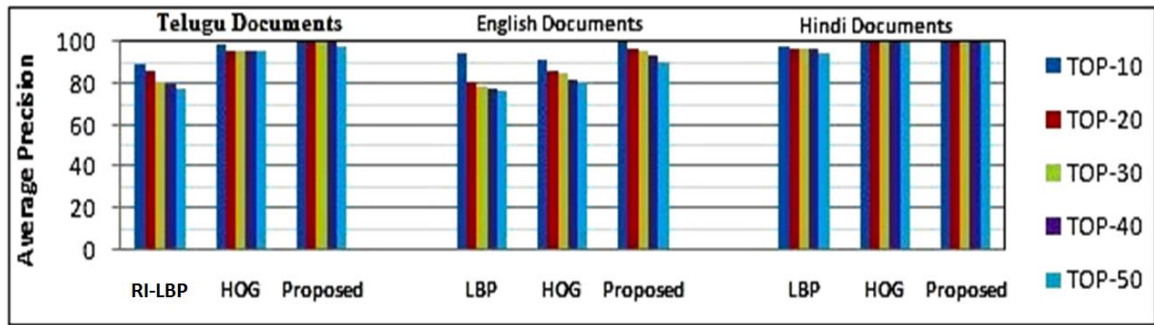


Fig. 6.13 Graphical Comparison of Results for Dataset3

The graphical comparison of results reveals that the proposed multi-resolution HOG features outperform compared to existing feature schemes for language-based document image retrieval system.

6.8 Summary and Conclusion

The language-based classification and retrieval system is an essential application for sorting and accessing the documents from a huge database consisting of multi-language documents. This chapter proposed the multi-resolution HOG features for classifying and retrieving the document images based on the language. The proposed method is tested for document images of Kannada, Marathi, Telugu, Hindi and English language. The proposed feature extraction scheme provided promising results compared to the existing state of art techniques.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

Document image analysis and retrieval system is an essential application for the implementation of a paperless office. The conclusions drawn out of this research work are as follows.

1. The document image analysis and retrieval is an interesting area of research and has many challenges. Chapter 1 provided an overview of the system, a detailed literature survey and the major contribution of the proposed research work. The survey carried out has been published in [123].
2. Logo-based document image retrieval is an essential tool for accessing and searching for the documents. Chapter 2 proposed an automatic logo-based document image retrieval using SVD based features. Automatic logo detection from the document is implemented based on the energy possessed by the connected components. The proposed logo detection algorithm provided a detection rate of 90.06%. Singular values of an image are found to be good set of features for logo-based document image retrieval. An average precision of 84% is achieved using proposed system. The results obtained with the proposed method outperform compared with the DWT based features. The results of this chapter have been published in [124], [125] and [126].
3. Logo-based document retrieval has the limitation of accessing documents which do contain logos and cannot be used for documents which contain signature only. To overcome this a signature-based document retrieval is proposed in Chapter 3. As the similarity metrics used for matching and retrieval of the signature play a vital role and influence on the performance, an experimental comparison of results using different similarity measures is presented. The seven distance metrics namely, Euclidean, Canberra, City-block, Chebychev, Cosine, Hamming and Jaccard are investigated for single and multi-level DWT based features. The combination of multi-level DWT

based features with city block distance provided better performance. A result of this chapter has been published in [127].

4. The document images such as identity cards, PAN cards, passports, certificates are embodied with photos. The logo and signature-based document retrieval techniques fail to access such documents. To overcome this Chapter 4 presents the use of SVD based and GLCM based features for face/photo based document image retrieval. The proposed feature extraction schemes provided a mean average precision of 82.66% which is very promising compared with Haralick features. The results of this chapter have been published in [37] and [128].
5. Recently the important documents in many organizations are embedded with fingerprint impression of the person for authentication to provide high security. Property registration, letters related to legal issues, bank transaction records are some of the examples. To retrieve such documents based on the fingerprint impression Chapter 5 proposed DWT/SWT based LBP features for fingerprint-detection and fingerprint-based document retrieval. Proposed fingerprint detection method has an accuracy of 98.87%. The fingerprint document retrieval with proposed features has provided a mean average precision of 73.08% for 1200 document images. Proposed feature extraction schemes provided promising results compared with existing schemes. Results of this chapter have been published in [40] and [129].
6. The acceptance for transaction in multiple languages creates a huge database of documents in different languages. This motivated us to propose the computationally cheap and faster techniques for language-based document classification and retrieval techniques using multi-resolution HOG features in Chapter 6. The proposed method is tested on document images of Kannada, Marathi, Telugu, Hindi and English languages. An average classification rate of 87.02% is achieved using the proposed system. The language-based retrieval system is tested for combination Kannada-English-Hindi, Marathi-English-Hindi and Telugu-English-Hindi languages. The proposed system provided very promising results compared with existing methods. Results of this chapter have been communicated [130-131].

7.2 Future Directions

Document image analysis and retrieval is an open area for research. The modern technology has developed new opportunities for the researchers in this domain. The following are the observations and directions for upcoming research in this particular domain.

1. The signature-based document retrieval presented in Chapter 2 employs a handwritten signature with printed documents. However, the detection of the signature from handwritten documents can be considered in future research. The language-based classification and retrieval of document images proposed in Chapter 6 can be further extended for handwritten documents.
2. Since document images captured nowadays are of high resolution, they require more processing time. Hence, there is a need for the development of computationally cheap and efficient feature extraction schemes to improve the speed of document image analysis and retrieval.
3. The retrieval time for searching the documents must be optimal. As the retrieval time is directly proportional to the dimension of the features used, the new and efficient feature reduction techniques can be developed.
4. The computerization of medical diagnosis has lead to a huge number of document images. This gives an opportunity for developing a document search engine for searching document images based on disease identified, patient details, symptoms identified, etc.
5. Digitization of documents in every field may require proper classification techniques for sorting and searching of the document images. The development of efficient sorting and searching of the document images that suit for a particular application is an open area of research.
6. The developments in image processing and artificial intelligence have given rise to many deep learning and machine learning algorithms. Use of deep learning and new machine learning algorithms can be investigated on document image analysis and retrieval.

Author's Publications

Book Chapters:

- [1] Umesh D. Dixit and M. S. Shirdhonkar, "Face-Biometric Based Document Image Retrieval using SVD Features," *Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing Series*, vol. 556, pp. 481-488, 2017.
- [2] Umesh D. Dixit and M. S. Shirdhonkar, "Language-based Document Images using Hybrid Texture Features," *Advances in Biometrics – Modern Methods and Implementation Strategies, Springer E-Book. (Accepted)*

Journals:

- [1] Umesh D. Dixit and M. S. Shirdhonkar, "A Survey on Document Image Analysis and Retrieval System," *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 2, pp. 259-270, 2015.
- [2] Umesh D. Dixit and M. S. Shirdhonkar, "Signature based Document Image Retrieval using Multi-level DWT Features," *International Journal of Image, Graphics & Signal Processing (IJIGSP)*, vol. 9, no. 8, pp.42-49, 2017.
- [3] Umesh D. Dixit and M. S. Shirdhonkar, "Face-based Document Image Retrieval System," *Procedia Computer Science, Elsevier*, vol. 132, pp. 659-668, 2018.
- [4] Umesh D. Dixit and M. S. Shirdhonkar, "Fingerprint-Based Document Image Retrieval," *International Journal of Image and Graphics (IJIG)*, vol. 19, no. 2, pp. 1-17, 2019.
- [5] Umesh D. Dixit and M. S. Shirdhonkar, "Preprocessing Framework for Document Image Analysis," *International Journal of Advanced Networking and Applications (IJANA)*, vol. 10, no.4, pp. 3911-3918, 2019.
- [6] Umesh D. Dixit and M. S. Shirdhonkar, "Language-based Document Image Retrieval System," *International Journal of Information Technology (IJIT), Springer. (Accepted)*

Conferences:

- [1] Umesh D. Dixit and M. S. Shirdhonkar, "Automatic Logo Detection and Extraction using Singular Value Decomposition," *Proc. IEEE Int'l Conf. on Communication and Signal Processing (ICCSP)*, pp. 787-790, 2016.
- [2] Umesh D. Dixit and M. S. Shirdhonkar, "Logo-based Document Image Retrieval using Singular Value Decomposition Features," *Proc. IEEE Int'l Conf. on Signal and Information Processing (IconSIP)*, pp. 1-4, 2016.
- [3] Umesh D. Dixit and M. S. Shirdhonkar, "Automatic Logo Extraction from Document Images," *Proc. Int'l Conf. CNC 2015*, pp. 221-226, 2015.

Bibliography

- [1] Rangachar Kasturi, Lawrence O’Gorman and Venu Govindraju, “Document image analysis: A primer,” *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002.
- [2] Manesh B. Kokare, M. S. Shirdhonkar, “Document Image Retrieval: An Overview,” *International Journal of Computer Applications*, vol. 1, no. 7, pp. 128-133, 2010.
- [3] J. T. Favata and G. Srikantan, “A Multiple Feature Resolution Approach for Hand/ Printed Digit and Character Recognition,” *International Journal of Imaging Systems and Technology*, vol. 7, no. 4, pp. 304-311, 1996.
- [4] G. Srikantan, S. Lam and S.Srihari, “Gradient Based Contour Encoding for Character Recognition,” *Pattern Recognition*, vol. 29, no. 7, pp. 1147-1160, 1996.
- [5] Hong Liu, Suoqian Feng, Hong bin Zha, Xueping Liu, “Document Image Retrieval Based On Density Distribution Feature and Key Block Feature,” *Proc. Int’l Conf. on Document Analysis and Recognition (ICDAR’05)*, pp. 1040-1044, 2005.
- [6] Shravya Shetty, Harish Srinivasan, Matthew Beal and Sargur Srihari, “Segmentation and Labeling of Documents using Conditional Random Fields,” *Document Recognition and Retrieval XIV*, vol. 6500, pp. 1-5, 2007.
- [7] A. Balasubramanian, Million Meshesha and C.V. Jawahar, “Retrieval from Document Image Collections,” *Int’l Workshop on Document Analysis Systems*, Springer, Berlin, Heidelberg, pp. 1-12, 2006.
- [8] C.V.Jawahar, Million Meshesha and A. Balsubramanium, “Searching in Document Images,” *Proc. 4th Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 622-627, 2004.
- [9] Manesh Kokare, B. N. Chatterji and P. K. Biswas, “Comparison of Similarity Metrics for Texture Image Retrieval,” *Proc. IEEE Conf. on Convergent Technologies for Asia-Pacific Region*, vol. 2, pp. 571-575, 2003.
- [10] Reza Tavoli, “Classification and Evaluation of Document Image Retrieval System,” *WSEAS Trans. on Computers*, vol. 11, no. 10, pp. 329–338, 2012.

-
- [11] Y.Tang, C.D.Ya and C.Y.Suen, "Document Processing for Automatic Knowledge Acquisition," *IEEE Trans. on Knowledge and Data Engg.*, vol. 6, no.1, pp.3-21, 1994
- [12] D. Niyogi and S. Srihari, "The Use of Document Structure Analysis to Retrieve Information from Documents in Digital Libraries," *Proc. SPIE, Document Recognition IV*, vol. 3027, pp.207-218, 1997.
- [13] J. Liu and A.K.Jain, "Imaged-Based Form Document Retrieval," *Pattern Recognition*, vol. 33, no.3, pp.503-513, 2000.
- [14] Y.He, Z. Jiang, B. Liu, and H. Zhao, "Content-Based Indexing and Retrieval Method of Chinese Document Images," *Proc. 5th Int'l Conf. Document Analysis and Recognition (ICDAR '99)*, pp. 685-688, 1999.
- [15] Yue Lu and Chew Lim Tan, "Information Retrieval in Document Image Databases," *IEEE Trans. on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1398-1410, 2004.
- [16] Hong Liu, Suoqian Feng, Hong bin Zha, Xueping Liu, "Document Image Retrieval Based On Density Distribution Feature and Key Block Feature," *Proc. Conf. on Document Analysis and Recognition (ICDAR'05)*, pp. 1040-1044, 2005.
- [17] Tanohiro Nakai, Koichi Kise, Masakazu Iwamura, "Camera Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants," *Proc. 8th Int'l Conf. on Document Analysis and Recognition*, pp. 379-383, 2005.
- [18] L. R. B. Schomaker, "Retrieval of Handwritten Lines in Historical Documents," *Proc. Document Analysis and Recognition, ICDAR 2007*, vol. 2, pp. 594-598, 2007.
- [19] Guillaume Joutel, Veronique Eglin, Stephane Bres, Hubert Emptoz, "Curvelets based Features Extraction of Handwritten Shapes for Ancient Manuscripts classification," *Proc. SPIE-IS&T Electronic Imaging*, vol. 6500, pp. 1-4, 2007.
- [20] Shijian Lu, Linlin Li, Chew Lim Tan, "Document Image Retrieval through Word Shape Coding," *IEEE Trans. on Pattern And Machine Intelligence*, vol. 30, no.11, pp.1913-1918, 2008.

-
- [21] Vikram T.N, Shalini R. Urs, K. Chidananda Gowda, "Person Specific Document Retrieval Using Face Biometrics," *ICADL, Lecture Notes on Computer Science* 5362, Springer-Verlag Berlin Heidelberg, pp. 371–374, 2008.
- [22] Ehtesham Hassan, Santanu Chaudhury, M Gopal, "Shape Descriptor Based Document Image Indexing and Symbol Recognition," *Proc. 10th Int'l Conf. on Document Analysis and Recognition*, pp.206-210, 2009.
- [23] Jilin Li, Zhi-Gang Fan, Yadong Wu and Ning Le, "Document Image Retrieval with Local Features Sequences," *Proc. 10th Int'l Conf. on Document Analysis and Recognition*, pp. 346-350, 2009.
- [24] M. S. Shirdhonkar and Manesh B. Kokare, "Document Image Retrieval using Signature as Query," *Proc. 2nd Int'l Conf. on Computer and Communication Technology*, pp. 66-70, 2011.
- [25] M. S. Shirdhonkar and Manesh B. Kokare, "Writer Based Handwritten Document Image Retrieval Using Contourlet Transform," *Advances in Digital Image Processing and Information Technology Communications in Computer and Information Science*, vol. 205, pp. 108-117, 2011.
- [26] M. Keyvanpour and R. Tavoli, "Feature Weighting for Improving Document Image Retrieval System Performance," *International Journal of Computer Science*, vol. 9, no. 3, pp. 25-130, 2012.
- [27] Giuseppe Pirlo, Michela Chimienti, Michele Dassisti, Donato Impedovo, Angelo Galiano, "Layout Based Document Retrieval System by Radon Transform Using Dynamic Time Warping," *Proc. Image Analysis and Processing –ICIAP 2013, Lecture Notes in Computer Science*, vol. 8156, pp. 61-70, 2013.
- [28] Ravi Shekhar and C.V.Jawahar, "Document Specific Sparse Coding for Word Retrieval," *Proc. Document Analysis and Recognition(ICDAR), 12th Int'l Conf.*, pp. 643-647, 2013.
- [29] Morteza Valizadeh and Ehsanollah Kabir, "An Adaptive Water Flow Model for Binarization of Degraded Document Images," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 16, no.2, pp. 165-176, 2013.
-

-
- [30] Melissa Cote, Alexandra Branzan Albu, “Texture Sparseness for Pixel Classification of Business Document Images,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no.3, pp. 257-273, 2014.
- [31] Nicolás Serrano, Adrià Giménez, Jorge Civera, Alberto Sanchis, Alfons Juan, “Interactive Handwriting Recognition with Limited User Effort,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 1, pp. 47-59, 2014.
- [32] K. Pramod Sankar, R. Manmatha, C. V. Jawahar, “Large Scale Document Image Retrieval by Automatic Word Annotation,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 1, pp. 1-17, 2014.
- [33] Thai V. Hoang, Elisa H. Barney Smith, Salvatore Tabbone, “Sparsity-based Edge Noise Removal from Bilevel Graphical Document Images,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 2, pp. 161-179, 2014.
- [34] Marçal Rusiñol, Volkmar Frinken, Dimosthenis Karatzas, Andrew D. Bagdanov, Josep Lladós, “Multimodal Page Classification in Administrative Document Image Streams,” *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 7, pp. 331-341, 2014.
- [35] M. S. Shirdhonkar and Manesh Kokare, “Automatic Logo Detection in Document Images,” *Proc. IEEE Int’l Conf. on Computational Intelligence and Computer Research*, pp. 905-907, 2010.
- [36] R. Haralick, K. Shanmugam, I. Dinstein, “Textural Features for Image Classification,” *IEEE Trans. on Systems, Man and Cybernetics*, vol. 3, Issue 6, pp. 610–621, 1973.
- [37] Umesh D. Dixit and M. S. Shirdhonkar, “Face-Biometric Based Document Image Retrieval using SVD Features,” *Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing Series*, vol. 556, pp. 481-488, 2017.
- [38] Dalal N. and Triggs B., “Histograms of Oriented Gradients for Human Detection,” *Proc. Int’l Conf. on Computer Vision Pattern Recognition*, pp 886-893, 2005.
-

-
- [39] Hassan T. and Khan H. A., "Handwritten Bangla Numeral Recognition Using Local Binary Pattern," *Proc. IEEE Int'l Conf. on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1-4, 2015.
- [40] Umesh D. Dixit and M. S. Shirdhonkar, "Fingerprint-Based Document Image Retrieval," *International Journal of Image and Graphics (IJIG)*, vol. 19, no. 2, pp. 1-17, 2019.
- [41] G. Zhu and D. Doermann. Tobacco-800 Complex Document Image Database and Ground truth. online, 2008. <http://lampsrv01.umiacs.umd.edu/projdb/edit/project.php?id=52>
- [42] S. Seiden, M. Dillencourt, S. Irani, R. Borrey and T. Murphy, "Logo Detection in Document Images," *Proc. Int'l Conf. on Imaging Science, Systems, and Technology*, pp. 446-449, 1997.
- [43] Pham T., "Unconstrained Logo Detection in Document Images," *Pattern Recognition*, vol. 36: pp. 3023-3025, 2003.
- [44] Zhu, G. and D. Doermann, "Automatic Document Logo Detection," *Proc. Conf. on Document Analysis and Recognition*, pp. 864-868, 2007.
- [45] Sina Hassanzadeh and Hossein Pourghassem, "A Novel Logo Detection and Recognition Framework for Separated Part Logos in Document Images," *Australian Journal of Basic and Applied Sciences*, vol. 5, pp. 936-946, 2011.
- [46] Arash Asef Nejad and Karim Faez, "A Novel Method for Extracting and Recognizing Logos," *International Journal of Electrical and Computer Engineering*, vol. 2, no. 5, pp. 577-588, 2012.
- [47] Hongye Wang and Youbin Chen, "Logo Detection in Document Images Based on Boundary Extension of Feature Rectangles," *Proc. 10th Int'l Conf. on Document Analysis and Recognition*, pp. 1335-1337, 2009.
- [48] Guangyu Zhu and David Doermann, "Logo Matching for Document Image Retrieval," *Proc. 10th Int'l Conf. on Document Analysis and Recognition*, pp. 606-610, 2009.
- [49] Rajiv Jain and David Doermann, "Logo Retrieval in Document Images," *Document Analysis Systems, IAPR International Workshop on, Document Analysis Systems*, pp. 135-139, 2012.
-

-
- [50] V. P. Le, N. Nayef, M. Visani, J. M. Ogier, C. De Tran, "Document Retrieval Based on Logo Spotting Using Key-Point Matching," *Proc. Int'l Conf. on Pattern Recognition (ICPR-2014)*, pp. 3056-3061, 2014.
- [51] Lijie Cao, "Singular Value Decomposition Applied To Digital Image Processing," Division of Computing Studies, Arizona State University Polytechnic Campus, Mesa, Arizona 85212, pp. 1–15.
- [52] S. Djeziri, F. Nouboud, R. Plamondon, "Extraction Of Signatures from Check Background Based On A Filiformity Criterion," *IEEE Trans. on Image Processing*, vol. 7, no. 10, pp. 1425–1438, 1998.
- [53] V. K. Madasu, M. H. M. Yusof M Hanmandilu K. K, "Automatic Extraction of Signatures From Bank Cheques And Other Documents," *Proc. DICTA'03*, pp. 591-600, 2003.
- [54] Abdullah Chalechale and Golshah Naghdy, "Signature Based Document Retrieval," *Proc. 3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 597-600, 2003.
- [55] Sargur N. Srihari, Shravya Shetty, Gady Agam and Ophir Frieder, "Document Image Retrieval Using Signature as Queries," *Proc. 2nd Int'l Conf. on Document Image Analysis for Libraries (DIAL'06)*, pp. 6-10, 2006.
- [56] G. Zhu, Y. Zheng, D. Doermann, S. Jeager, "Multi-scale Structural Saliency for Signature Detection," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.
- [57] G. Zhu, Y. Zheng, D. Doermann, "Signature-Based Document Image Retrieval," *ECCV, Part III, Lecture Notes on Computer Science 5304*, pp.752-765, 2008.
- [58] H. Srinivasan and Sargur Sridhar, "Signature-Based Retrieval Of Scanned Documents Using Conditional Random Fields," *Computational Methods for Counterterrorism*, Springer, pp. 17-32, 2009.
- [59] R. Mandal, P. P. Roy, U. Pal, "Signature Segmentation from Machine Printed Documents using Conditional Random field," *Proc. Int'l Conf. on Document Analysis and Recognition*, pp. 1170-1174, 2011.

-
- [60] Roy P.P, Bhowmick S, Pal U., Ramel J.Y, “Signature Based Document Image Retrieval Using GHT of Background Information,” *Proc. Int’l Conf. on Frontiers in Handwriting Recognition (ICFHR)*, pp. 225-230, 2012.
- [61] Mandal R, Roy P.P, Pal U., Blumenstien M, “Signature Segmentation and Recognition from Scanned Documents,” *Proc. Int’l Conf. on Intelligent Systems Design and Applications (ISDA)*, pp. 80-85, 2013.
- [62] Ilkhan Cuceloglu and Hasan Ogul, “Detecting Hand Written Signatures from Scanned Documents,” *Proc. 19th Computer Vision Winter Workshop*, 2014.
- [63] Thomas Schulz and Robert Sablatnig, “Signature Matching in Document Image Retrieval,” *Proc. 20th Computer Vision Winter Workshop*, Austria, pp. 26-41, 2015.
- [64] Heri Nurdianto, Hermanto Hermanto, “Signature Recognition using Neural Network Probabilistic,” *International Journal of Advances in Intelligent Informatics*, vol. 2, Issue 1, pp. 46-53, 2016.
- [65] Seyyid Ahmed Medjahed, “A Comparative Study of Feature Extraction Methods in Images Classification,” *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol. 7, no. 3, pp.16-23, 2015.
- [66] Lakhdar Belhallouche, Kamel Belloulata, Kidiyo Kpalma, "A New Approach to Region Based Image Retrieval using Shape Adaptive Discrete Wavelet Transform," *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol.8, no.1, pp.1-14, 2016.
- [67] K. P. Soman and K. I. Ramachandran, *Insight into Wavelets from Theory to Practice*, PHI Learning Pvt. Limited, 2010.
- [68] McCune, B. and J. B. Grace, “Analysis of Ecological Communities,” *MjM Software Design*, Gleneden Beach, Oregon <http://www.pcord.com>, 2002.
- [69] Libor Spacek, *Faces Directories*, Faces 94 Directory, <http://cswww.essex.ac.uk/mv/allfaces>
- [70] H. B. Kekre, Sudeep D. Thepade, Akshay Maloo, “Face Recognition Using Texture Features Extracted From Walshlet Pyramid,” *International Journal on Recent Trends in Engineering & Technology*, vol. 5, no. 2, pp. 186-193, 2011.

-
- [71] Anastasios N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," *IEEE Trans. on Neural Networks*, vol. 14, no. 1, pp. 118-126, 2003.
- [72] Majid Ahmadi, Javad Haddadniaa, Karim Faeza, "A Fuzzy Hybrid Learning Algorithm For Radial Basis Function Neural Network With Application In Human Face Recognition," *Pattern Recognition Society*, vol. 36, no. 5, pp. 1187-1202, 2003.
- [73] Ying Weng, Aamer Mohamed, Jianmin Jiang, Stan Ipson, "Face Detection based Neural Networks using Robust Skin Color Segmentation," *Proc. 5th International Multi-Conference on Systems Signals and Devices*, pp. 1-5, 2008.
- [74] Xiaoyang Tan and Bill Triggs, "Fusing Gabor and LBP Feature Sets for Kernel-based Face Recognition," *Proc. 3rd International Workshop on Analysis and Modeling of Faces and Gestures*, vol. 4778, pp. 235-249, 2007.
- [75] Lu Leng, Jiashu Zhang, Muhammad Khurram Khan, Xi Chen, Khaled Alghathbar, "Dynamic Weighted Discrimination Power Analysis: A Novel Approach for Face and Palmprint Recognition In DCT Domain," *International Journal of the Physical Sciences*, vol. 5, no. 17, pp. 2543-2554, 2010.
- [76] Unsang Park and Anil K. Jain, "Face Matching and Retrieval using Soft Biometrics," *IEEE Trans. on Information Forensics and Security*, vol. 5, no. 3, pp.406-415, 2010.
- [77] Bor-Chun Chen, Yan-Ying Chen, Yin-His Kuo, Winston H.Hsu. (2013) "Scalable Face Image Retrieval using Attribute-Enhanced Sparse Codewords," *IEEE Trans. on Multimedia*, vol. 15, no. 5, pp. 1-11, 2013.
- [78] S. Pannirselvam and S. Prasath, "A Novel Technique for Face Recognition and Retrieval using Fiducial Point Features," *Procedia Computer Science*, Elsevier, vol. 47, pp. 301-310, 2015.
- [79] Shiv Ram Dubey, Satish Kumar Singh, Rajat Kumar Singh, "Local SVD based NIR Face Retrieval," *Journal of Visual Communications and Image Representation*, vol. 49, pp. 141-152, 2017.
- [80] Otsu N., "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. on Systems, Man, and Cybernetics* vol. 9, no. 1, pp. 62-66, 1979.
-

-
- [81] Bino Sebastian V, A. Unnikrishnan, Kannan Balakrishnan, "Gray Level Co-Occurrence Matrices: Generalization and Some New Features," *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 2, pp. 151-157, 2012.
- [82] Ross A. and Jain A., "Biometric Sensor Interoperability: A Case Study in Fingerprints," *Maltoni D., Jain A.K. (eds) Biometric Authentication. BioAW 2004. Lecture Notes in Computer Science*, vol. 3087. Springer, Berlin, Heidelberg, 2004
- [83] Anil K. Jain, Yi Chen, Meltem Demirkus, "Pores And Ridges: High-Resolution Fingerprint Matching Using Level 3 Features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no.1, pp. 15-27, 2007.
- [84] S Arivazhagan, L Ganesan, "Texture Classification Using Wavelet Transform," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1513-1521, 2003.
- [85] J. T Overpeck, T Webb, I. C Prentice, "Quantitative Interpretation of Fossil Pollen Spectra: Dissimilarity Coefficients and The Method of Modern Analogs," *Quaternary Research*, vol. 23, no. 1, pp. 87-108, 1985.
- [86] Jiang, Xudong, Manhua Liu, Alex C. Kot, "Fingerprint Retrieval for Identification," *IEEE Trans. on Information Forensics and Security*, vol.1, no.4, pp. 532-542, 2006.
- [87] X. Chen, J. Tian, X. Yang, Y. Zhang, "An Algorithm for Distorted Fingerprint Matching Based on Local Triangle Feature Set," *IEEE Trans. on information forensics and security*, vol.1, no. 2, pp. 169-177, 2006.
- [88] Yuliang He, Jie Tian, Liang Li, Hong Chen, Xin Yang, "Fingerprint Matching Based on Global Comprehensive Similarity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 850-862, 2006.
- [89] Manhua Liu, Xudong Jiang, Alex Chichung Kot, "Efficient Fingerprint Search Based on Database Clustering," *Pattern Recognition*, vol.40, no.6, pp. 1793-1803, 2007.
- [90] Anil K. Jain, Jianjiang Feng, Abhishek Nagar, Karthik Nandakumar, "On Matching Latent Fingerprints," *Proc. Computer Vision and Pattern Recognition Workshops, CVPRW'08. IEEE Computer Society Conference*, pp. 1-8, 2008.
-

-
- [91] Zegarra, Javier A. Montoya, Neucimar J. Leite, Ricardo da Silva Torres, "Wavelet-Based Fingerprint Image Retrieval," *Journal of Computational and Applied Mathematics*, vol. 227, no. 2, pp. 294-307, 2009.
- [92] Anil K. Jain and Jianjiang Feng, "Latent Fingerprint Matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 88-100, 2011.
- [93] Nanni, Loris, Alessandra Lumini, "Local Binary Patterns for a Hybrid Fingerprint Matcher," *Pattern Recognition*, vol.41, no.11, pp.3461-3466, 2008.
- [94] Jung, Hye-Wuk, Jee-Hyong Lee, "Fingerprint Classification Using the Stochastic Approach of Ridge Direction Information," *Proc. FUZZ-IEEE 2009. IEEE Int'l Conf.*, pp. 169-174, 2009.
- [95] Bharkad, Sangita, Manesh Kokare, "Fingerprint Matching Using Discrete Wavelet Packet Transform," *Advance Computing Conference (IACC)*, pp. 1183-1188, 2013.
- [96] Le, Thai Hoang, Hoang Thien Van, "Fingerprint Reference Point Detection for Image Retrieval Based on Symmetry And Variation," *Pattern Recognition*, vol. 45, no. 9, pp. 3360-3372, 2012.
- [97] Cappelli, Raffaele, Matteo Ferrara, "A Fingerprint Retrieval System Based on Level-1 and Level-2 Features," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10465-10478, 2012.
- [98] M.A. Wahby Shalaby, M. Omair Ahmad, "A Multilevel Structural Technique for Fingerprint Representation and Matching," *Signal Processing*, vol.93, no.1, pp. 56-69, 2013.
- [99] Paulino, Alessandra A., Jianjiang Feng, Anil K. Jain, "Latent Fingerprint Matching Using Descriptor-Based Hough Transform," *IEEE Trans. on Information Forensics and Security*, vol.8, no.1, pp.31-45, 2013.
- [100] Arun D. R., C. Christopher Columbus, K. Meena, "Local Binary Patterns and Its Variants for Finger Knuckle Print Recognition in Multi-Resolution Domain," *Circuits and Systems*, vol.7, no.10, pp. 3142-3149, 2016.
- [101] E. O. Rodrigues, T. M. Porcino, Aura Conci, Aristofanes C. Silva, "A Simple Approach for Biometrics: Finger-Knuckle Prints Recognition Based on a Sobel Filter and Similarity Measures," *Proc. International Workshop on Systems, Signals and Image Processing (IWSSIP)*, pp. 1-4, 2016.
-

-
- [102] A. Tzalavra, K. Dalakleidi, Evangelia I. Zacharaki, Nikolaos Tsiaparas, Fotios Con Stantinidis, "Comparison of Multi-resolution Analysis Patterns for Texture Classification of Breast Tumors Based on DCE-MRI," *Proc. International Workshop on Machine Learning in Medical Imaging*, pp. 296-304, 2016.
- [103] Huma Qayyum, MuhammadMajid, Syed Muhammad Anwar, Bilal Khan, "Facial Expression Recognition Using Stationary Wavelet Transform Features," *Mathematical Problems in Engineering*, vol. 2017, pp. 1-9, 2017.
- [104] M. S. Shirdhonkar and Manesh B. Kokare, "Discrimination Between Printed and Handwritten Text in Documents," *IJCA Special Issue on Recent Trends on Image Processing and Pattern Recognition*, pp.131-134, 2010.
- [105] Haralick Robert M., Linda G. Shapiro, "Computer and Robot Vision," vol. 1, *Addison-Wesley*, pp. 28-48., 1992.
- [106] Maenpaa Topi, Matti Pietikäinen, "Texture Analysis With Local Binary Patterns," *Handbook of Pattern Recognition and Computer Vision*, vol.3, pp.197-216, 2005.
- [107] Chaudhury S., Harit G., Madnani S., Shet R. B., "Identification of Scripts of Indian Languages By Combining Trainable Classifiers," *Proc. ICVGIP 2000*, pp. 20-22, 2000.
- [108] Kulkarni A., Upparamani P., Kadkol R., Tergundi P., "Script Identification from Multilingual Text Documents," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 15-19, 2015.
- [109] Padma M. C. and Vijaya P. A., "Script Identification from Trilingual Documents using Profile Based Features," *International Journal of Computer Science and Applications*, vol. 7, no. 4, pp. 16-33, 2010.
- [110] Pal U. and Chaudhuri B. B., "Automatic Identification of English, Chinese, Arabic, Devanagari and Bangla Script Line," *Proc. 6th Int'l Conf. on Document Analysis and Recognition*, pp. 790-794, 2001.
- [111] Rajput G. G. and Anita H. B., "Handwritten Script Recognition Using DCT and Wavelet Features at Block Level," *International Journal of Computer Applications, Special issue on RTIPPR*, vol. 3, pp. 158-163, 2010.
-

-
- [112] Obaidullah S. M., Mondal A., Das N., Roy K., "Script Identification from Printed Indian Document Images and Performance Evaluation Using Different Classifiers," *Applied Computational Intelligence and Soft Computing*, vol. 2014, pp. 1-12, 2014.
- [113] Pardeshi R., Chaudhuri B. B., Hangarge M., Santosh, K. C., "Automatic Handwritten Indian Scripts Identification," *Proc. IEEE 14th international conference on frontiers in handwriting recognition*, pp. 375-380, 2014.
- [114] Tan C. L., Huang W., Sung S. Y., Yu Z., Xu Y., "Text Retrieval from Document Images Based on Word Shape Analysis," *Applied Intelligence*, vol. 18, no. 3, pp. 257-270, 2003.
- [115] Arani S. A. A. A., Kabir E., Ebrahimpour R., "Handwritten Farsi Word Recognition Using NN-Based Fusion of HMM Classifiers With Different Types of Features," *International Journal of Image and Graphics*, vol. 19, no. 01, pp.1-21, 2019.
- [116] Djeddi C., Siddiqi I., Souici-Meslati L, Ennaji A, "Text-Independent Writer Recognition Using Multi-Script Handwritten Texts," *Pattern Recognition Letters*, vol. 34, no. 10, pp. 1196-1202, 2013.
- [117] Roy P. P., Bhunia A. K., Das A., Dey P., Pal U., "HMM-Based Indic Handwritten Word Recognition Using Zone Segmentation," *Pattern Recognition*, vol. 60, pp. 1057-1075, 2016.
- [118] Spitz A. L., "Determination of Script and Language Content of Document Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 235-245, 1997.
- [119] Thanuja C. and Shreedevi G. R., "Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection," *International Journal of Engineering Research and Applications*, vol. 3, pp. 1329-1335, 2013.
- [120] Lakshmi T. V., "Reduction of Features to Identify Characters from Degraded Historical Manuscripts," *Alexandria Engineering Journal*, vol. 57, no. 4, pp. 2393-2399, 2018.

-
- [121] Chandrakala H. T., “A Kannada Document Image Retrieval System based on Correlation Method,” *International Journal of Computer Applications*, vol. 77, no. 3, pp. 39-46, 2013.
- [122] P. B. Patil and M. B. Kokare, “Interactive Semantic Image Retrieval,” *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 349-364, 2013.
- [123] Umesh D. Dixit and M. S. Shirdhonkar, “A Survey on Document Image Analysis and Retrieval System,” *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 2, pp. 259-270, 2015.
- [124] Umesh D. Dixit and M. S. Shirdhonkar, “Automatic Logo Extraction from Document Images,” *International Journal on Cybernetics & Informatics (IJCI)*, vol. 4, no. 2, pp. 221-226, 2015.
- [125] Umesh D. Dixit and M. S. Shirdhonkar, “Automatic Logo Detection and Extraction using Singular Value Decomposition,” *Proc. IEEE Int’l Conf. on Communication and Signal Processing (ICCSP)*, pp. 787-790, 2016.
- [126] Umesh D. Dixit and M. S. Shirdhonkar, “Logo-based Document Image Retrieval using Singular Value Decomposition Features,” *Proc. IEEE Int’l Conf. on Signal and Information Processing (IconSIP)*, pp. 1-4, 2016.
- [127] Umesh D. Dixit and M. S. Shirdhonkar, “Signature based Document Image Retrieval using Multi-level DWT Features,” *International Journal of Image, Graphics & Signal Processing (IJIGSP)*, vol. 9, no. 8, pp. 42-49, 2017.
- [128] Umesh D. Dixit and M. S. Shirdhonkar, “Face-based Document Image Retrieval System,” *Procedia Computer Science, Elsevier*, vol. 132, pp. 659-668, 2018.
- [129] Umesh D. Dixit and M. S. Shirdhonkar, “Preprocessing Framework for Document Image Analysis,” *International Journal of Advanced Networking and Applications (IJANA)*, vol. 10, no.4, pp. 3911-3918, 2019.
- [130] Umesh D. Dixit and M. S. Shirdhonkar, “Language-based Document Images using Hybrid Texture Features,” *Advances in Biometrics – Modern Methods and Implementation Strategies, Springer E-Book. (Accepted)*
- [131] Umesh D. Dixit and M. S. Shirdhonkar, “Language-based Document Image Retrieval System,” *International Journal of Information Technology (IJIT), Springer. (Accepted)*
-