

Ph.D. Thesis Writing

Doctoral Program in “Department Name”

Deep Learning (3D CNN & Holistic CNN) based Content
Based Video Retrieval from Visual & Semantic Features

in Hadoop

by

<Name of the Candidate>

<Reg. No of the Candidate>

<Supervisor Name>

<Date of Submission (DD MM 20YY)>

ABBREVIATIONS

Abbreviation	Acronym
IR	Information Retrieval
CBIR	Content Based Image Retrieval
CBVR	Content Based Video Retrieval
CCV	Color Coherent Vector
RGB	Red Green Blue
HSV	Hue Saturation Value
HSB	Hue Saturation Brightness
CMY	Cyan Magenta Yellow
YIQ	Luminance Inphase Quadrature
HSL	Hue Saturation Lightness/Luminance
GLCM	Gray Level Co-occurrence Matrix
LBP	Local Binary Pattern
WWW	World Wide Web
HDFS	Hadoop Distributed File System
GoF	Group of Frames
DFD	Distance Field Descriptor
XML	Extensible Markup Language
RDF	Resource Description Framework
YARN	Yet Another Resource Negotiator
DCT	Discrete Cosine Transform
BOW	Bag Of Words
SIFT	Salient Invariant Feature Transform
LTtPS	Local terta Patterns
LDPs	Local Derivative Patterns
MIL	Multiple Instance Learning

OOI	Object Of Interest
STAR	Semi-Structured Clustering Technique
NDV	Near Duplicate Video
GPU	Graphical Processing Unit
MICC	Multimedia and Intelligent Computing Cluster
IFI	Inverted File Index
MSI	Markovian Semantic Indexing
ABIR	Annotation Based Image Retrieval
LSI	Latent Semantic Indexing
FCM	Fuzzy C-Means Clustering
HOG	Histogram of Oriented Gradients
FPS	Frame Per Second
bpp	bits per pixel
CDF	Cumulative Distribution Function
CNN	Convolutional Neural Network
3D	Three Dimensional
BoW	Bag of Words
BoVW	Bag of Visual Words
2D	Two Dimensional

ABSTRACT

In this day and age, information retrieval (IR) plays an important role in multimedia, which allows expressing tool capabilities and execution semantics in declarative and well-distinct way. Video retrieval is one of the most important processes in semantic web mining, which became an important research area in recent days. Several researchers have concentrated on both 2D and 3D video retrieval processes with different algorithms and techniques. This thesis provide the detailed review of conventional approaches with its benefits and also major problem, which has been discussed with pivot on following tasks: searching, classification, feature extraction, clustering, mapping, distance metric learning, ranking, selection and information retrieval. Based on these reviews, we clearly observed that semantic based retrieval concept produces more efficient results in multimedia systems. However, conventional techniques are not effectively handled with some problems such as storage complexity, accuracy of retrieval, computation time, etc.

To overwhelm these problems, this research focused on Content Based Video Retrieval (CBVR) process especially 3D in Hadoop environment. In our process, we applied MapReduce framework to reduce the processing time, storage while retrieving 3D videos. Our newfangled approach is mainly focused on scalability and processing speed of large dataset. Our proposed 3D CBVR is comprised with key frame extraction, Bag Of Visual Words (BOVW) generation, codebook generation and similarity matching. Initially, key frame extraction performs with Spatial and Temporally prioritized Gaussian Filter (STPGF), which effectively selects the optimal key frames for further process. Further, Hybrid 3D CNN is applied to extract the local descriptors like shape, texture and color features. For Texture and Color features extraction, SIFT and SURF descriptors are used. For this purpose, we combine the geometric and topological features in shape using

volumetric shape representation. Then MapRedece with decision tree algorithm is introduced for visual words and codebook generation that reduces the outliers and computation time while analyzing the similar visual words. Then, similarity matching is processed using soft weighting and L_2 distance function, which measure the similarity between the two images (query image and dataset frame). Based on the visual vector similarity, ranking is accomplished using K-Nearest Neighbor process which improves the proposed 3D CBVR retrieval results. In second contributions, 3D CBVR is presented with the denoising concepts. Here, Gaussian White Noise is reduced for the video frames using bilateral adaptive median filter and features are extracted Holistic 3D CNN for color, texture and motion features. Further shape features are extracted using 3D Concurrence Matrix. Our proposed novel techniques and algorithms for 3D CBVR in semantic web significantly reduces the processing time and storage complexity when comparing with previous techniques.

Our experimental result shows better accuracy and maximum number of served user requests than the state- of-the art algorithms. Finally, overall research process improves the different significant metrics such as precision, recall, accuracy, searching time, computation complexity and storage.

Table of Contents

S. No.	Content	Page No.
	ABSTRACT	i
	LIST OF FIGURES	v
	LIST OF TABLES	vii
1.	INTRODUCTION	1
	1.1 Introduction	1
	1.2 2D VS. 3D IN Image Processing	7
	1.3 Retrieval Methods in CBVR	16
	1.4 CBVR Overview	25
	1.4.1 Semantic Search Engine	30
	1.4.2 Hadoop MapReduce	32
	1.4.3 Three Dimensional (3D) Image Processing	44
	1.5 Thesis Motivation	47
	1.7 Thesis Research Methodology	50
	1.8 Organization of Thesis	52
2.	LITERATURE SURVEY	30
	2.1 Content based Image Retrieval	56
	2.2 Content based Video Semantic Search	65
3.	OVERVIEW OF CBVR (2D AND 3D)	73
	3.1 Introduction	74
	3.2 Technical Terms	75
	3.3 Key Frame Extraction and Selection	80

3.4	Bag of Visual Word Construction	84
3.5	Feature Extraction	88
3.7	Similarity Matching	94
4.	3D CNN WITH MAPREDUCE PARADIGM & VOLUMETRIC SHAPE REPRESENTATION FOR CBVR	99
4.1	Introduction	100
4.2	Problem Formulation	101
4.3	Research Findings	101
4.4	Proposed Hybrid 3D CNN	103
4.4.1	System Model	103
4.4.2	Key frames Extraction	105
4.4.3	Bag Of Visual Words (BOVW) Generation	108
4.5	Results Discussion	143
4.5.1	Testing Scenarios	144
4.5.2	Dataset Description	145
4.5.3	Comparative Analysis	146
4.5.3.1	Precision	146
4.5.3.2	Recall	149
4.5.3.3	F-measure	151
4.5.3.4	Retrieval Accuracy	154
4.5.3.5	Retrieval Error Rate	157
4.5.3.6	Retrieval Time	159
4.5.3.7	Cluster Purity	162
4.5.3.8	Segmentation Accuracy	164
4.5.3.9	Key Frame Selection	166
4.5.3.10	Positive Results	169
4.5.4	Obtained Output	172

4.6 Conclusion	176
5. 3D HOLISTIC CNN WITH MAP SHUFFLE REDUCE PARADIGM FOR CBVR CONTENT BASED VIDEO RETRIEVAL	177
5.1 Introduction	178
5.2 Research Methods	179
5.2.1 Denoising Methods	179
5.2.2 Feature Clustering Methods	185
5.3 Problem Definition	192
5.4 Proposed System	193
5.4.1 Key frame selection	198
5.4.2 Denoising	201
5.4.3 Feature extraction	203
5.4.4 Similarity matching	213
5.4.5 Hadoop MapReduce framework	216
5.5 Results Discussion	218
5.5.1 Dataset Description	218
5.5.2 Testing Scenarios	219
5.5.3 Comparative Analysis	221
5.5.3.1 Precision	221
5.5.3.2 Recall	224
5.5.3.3 F-measure	226
5.5.3.4 Retrieval Accuracy	228
5.5.3.5 Retrieval Error Rate	230
5.5.3.6 Retrieval Time	233
5.5.3.7 Key Frame Selection	235
5.5.3.8 Noise Filtering Accuracy	237

5.5.3.9	Positive Results	239
5.5.3.10	Processing Time	242
5.5.4	Obtained Output	243
5.6	Chapter Summary	245
6.	CONCLUSION AND FUTURE ENHANCEMENT	248
	REFERENCES	249



List of Figures

Figure No.	Figure Name	Page No.
1.1	Multimedia Growth in Social Networks	3
1.2	Information Retrieval Cycle Process	6
1.3	2D and 3D Animation	9
1.4	Classification of Image Retrieval	9
1.5	Flow of CBIR	12
1.6	Content-based Image Retrieval System	13
1.7	Content based Image Retrieval (e.g. Google)	14
1.8	Content based Image Retrieval Results (e.g. Google)	14
1.9	Content based Image Retrieval (Advanced Google search)	15
1.10	Image Retrieval Example	16
1.11	Retrieval Methods	17
1.12	Visually similar images (but semantic gaps are presented)	18
1.13	RGB color space Cartesian coordinate system	20
1.14	Single-hexcone model of HSV color space	21
1.15	HIS Color Space Model	22
1.16	Shape Representation and several extraction processes	24
1.17	Spatial and Temporal Structure of Videos	23
1.18	3D Image Based CBVR	27
1.19	Pixelwise & Patchwise Image Representation	28

1.20	Key Frame Extraction	30
1.21	MapReduce architecture	33
1.22	HDFS architecture	40
1.23	Job tracker working	42
1.24	Task tracker working	44
1.25	Thesis Organization	53
2.1	Image-to-Image Retrieval Process	56
2.2	Image-to Video Retrieval Process	65
4.1	System Architecture	104
4.2	Key Frames Extraction using STPGF	106
4.3	BOVW procedure	109
4.4	Flow of Feature Extraction Process	110
4.5	Overall Feature Extraction	111
4.6	Shape Feature Extractions with Topological and Geometrical Feature	112
4.7	Voronai Representation for Object	113
4.8	Voronoi Diagram	114
4.9	(a) Original image	117
	(b) Medial surface extraction and segmentation	117
	(c) Re-adjustment of segmentation	117
4.10	Color and Texture Features Extraction	120
4.11	Feature Extraction in CNN	125
4.12	key point location using SIFT	125
4.13	DoG Estimation and Key point identification	126
4.14	key Point descriptor	127
4.15	Second order Gaussian and Box Filter	129
4.16	Features Clustering using MapReduce	137
4.17	Precision vs. Number of Users	148

4.18	Recall vs. Number of Users	151
4.19	F-measure vs. Number of Users	153
4.20	RA vs. Number of Users	156
4.21	Retrieval Error Rate vs. Number of Users	158
4.22	Retrieval Time vs. Number of Users	161
4.23	Cluster Purity vs. Number of Users	164
4.24	Segmentation Accuracy vs. Number of Users	166
4.25	Key Frames Selected vs. Number of Key Frames	168
4.26	Positive Results vs. Number of Users	171
4.27	Obtained Experimental Result	175
5.1 (a)	Denoising Techniques	180
5.1 (b)	Block Diagram for Curvelet Transform	182
5.2	Clustering Algorithms	185
5.3	K-means Clustering	186
5.4	Hierarchical Clustering	187
5.5	Density based Clustering	188
5.6	Grid based Clustering	188
5.7	Model based Clustering	190
5.8	Illustration of DENCLUE density concept	192
5.9	Proposed Architecture	195
5.10	Key Frame Selection	201
5.11	Color Feature Extraction	210
5.12	Texture Feature Extraction	211
5.13	Motion Feature Extraction	212
5.14	Motion Feature	215
5.15	Hadoop Operations	217
5.16	Relevant Results for Action Query Image (Dataset 1)	220

5.17	Relevant Results for Games Query Image (Dataset 2)	220
5.18	Relevant Results for Sports Query Image (Dataset 3)	221
5.19	Performance of Precision	223
5.20	Performance of Recall	227
5.21	Performance of F-measure	227
5.22	Performance of Retrieval Accuracy	230
5.23	Performance of Retrieval Error Rate	232
5.24	Performance of Retrieval Time	235
5.25	Performance of Key Frame Selection	237
5.26	Performance of Noise Filtering Accuracy	239
5.27	Performances of Positive Results	242



List of Tables

Table No.	Table Name	Page No.
1.1	Difference between 2D and 3D Representation	6
1.2	Properties of Gray level Co-occurrence Matrix	12
3.1	Comparison of Various Techniques used in Image-to-Image Retrieval process	53
3.2	Different Conventional Techniques used in Image to Video Retrieval process	60
4.1	Hardware Configuration of our Implementation	93
4.2	Experimental Results of Training for Three Scenarios	97
4.3	Testing results for Three Scenarios	97
4.4	Overall MapReduce Results	98
5.1	Hardware Configurations	128
5.2	MapReduce Processing Time	129



CHAPTER 1

INTRODUCTION



1.1 INTRODUCTION

Multimedia can be defined as a combination of more than one media files with advanced computer based technology. A multimedia includes text, audio, video and graphics that are comprised into a single file format. Multimedia files are actively shared among people over the internet in different social networks and social applications. Multimedia also plays a major role in education by creating digital libraries and live online conferencing with worldwide experienced lecturers. Multimedia has become a common way used by people all over the world for sharing thoughts and ideas. Recently in past years people are attracted towards social networks to stay connected with people all over the world for multimedia information collection and sharing. Social networks over internet include Facebook, Twitter, Youtube, etc. This supported creation of real-life connections of people by sharing daily life's happening. YouTube is said to be the second largest search engine from which individual searches for videos under specified area. This social network also includes Businesses, Organizations, Research companies with emerging technologies, public media, etc. Due to the increased use of online customers in search of videos and images, the concept of search engine and database storages were under detailed research to tolerate huge number of customers. Data mining technology has emerged which discovers the concealed information from the huge scale database.

It is also acknowledged as Knowledge Data Discovery (KDD). Data mining is further defined as the preceding unknown, significant discovery and possibly beneficial information from the huge volumes of the data. It is an integrative research domain, it represented as the combination of numerous research areas, predominantly information theory, machine learning, statistics and database systems. Information Retrieval (IR) is a Problem-Oriented discipline which is concerned with the effective problem and accurate solution to retrieve the desired information between the human user and human generator [Rajam et al., 2013]. IR types and requirements are illustrated in table 1.1.

The proper definition for IR is follows: “*IR is determining the materials from the large collection of database that meets the human or system constraints or requirements (usually information can be any type)*”.

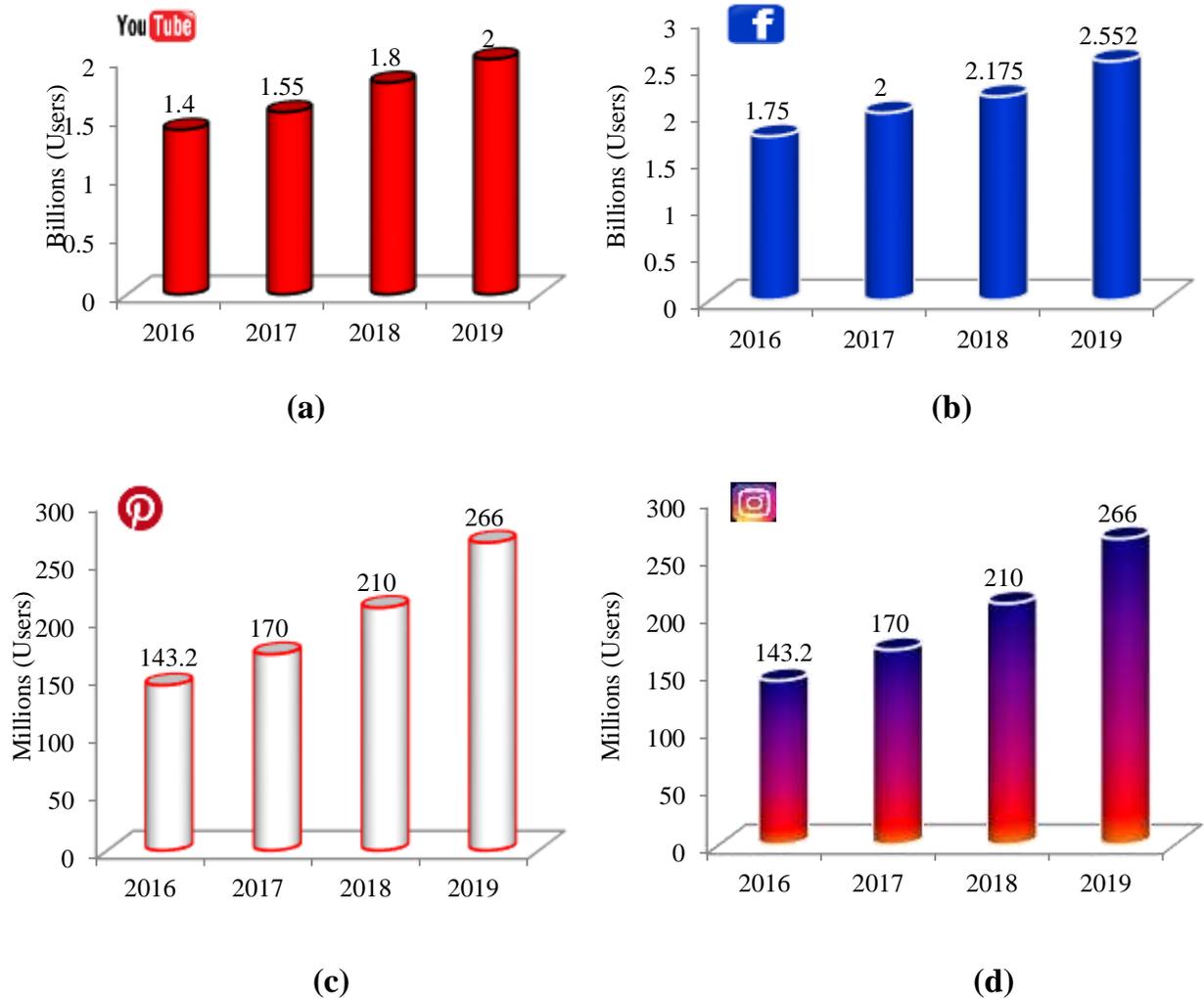


Figure 1.1 Multimedia Growth in Social Networks

With the growth of digital technologies and huge storage capacities in worldwide, a greater dimensions of digital media exist today. In recent years, the count of digital images is exponentially growing since the need of image acquisition increases for numerous reasons (personal use, security purpose, etc.). However, these images are published on the web or social networks via Internet [Ruofei Zhang et al., 2007, Khan et al., 2012, Mansoori et al., 2007, Ben Yossuf et al., 2014].

A digital video camera can be considered as an example for image sensor. Mass storage is required to store the images corresponding to their applications and usages. The principles of storage categories involve with three such as (1) online storage, (2) short-term storage and (3) archival storage.

Table 1.1 Information Retrieval Types and Requirements

Retrieval Type	Requirements
Retrospective	<ul style="list-style-type: none"> • Past information searching • Several queries pretended against static collection • Time variant
Prospective	<ul style="list-style-type: none"> • Future information searching • Static query pretended against dynamic collection • Time dependent

There are several types of information is retrieved within the database that video, audio, text, images, source codes, web services or applications. Figure 1.1 shows the cycle process of IR [Wu et al., 2013, Kamakshaiah et al., 2017, Alfred et al., 2010, Premalatha et al., 2014]. Image is a Matrix or Array of squared pixels organized by Columns and Rows. Image is a two dimensional function and the image function is $f(x, y)$ where x and y is the spatial coordinates and f is the amplitude of (x, y) . In image retrieval, the digital images are classified into thrice that are follows [Turlapaty et al., 2017, Vikhar et al., 2017, Karbil et al., 2017]:

- *Colored Images*

It comprised of three 3 Bands such as RED, GREEN, and BLUE. Each color component consists of 8 bytes of intensity, and intensity range or color distribution is different for pixels.

- *Gray scale Images*

It is referred as the Monochrome Images that are not represented by any color components, but it consists of the one color level of brightness. This is visually represented in black and white color. This type of image consists of 8 bytes (0-255) of brightness.

- *Binary Images*

This is very simplest type of image and considers binary values(0,1). The black value is represented in 1, and the white value is represented in 0. This type of images is well-supported for computer-vision applications.

- *Medical Images*

Medical imaging data is the main source in current bio-medical applications, which acquisition is represented in different color forms (gray scale, RGB, etc.). Different types of medical image modalities are CT, PET, MRI, and Ultrasound.

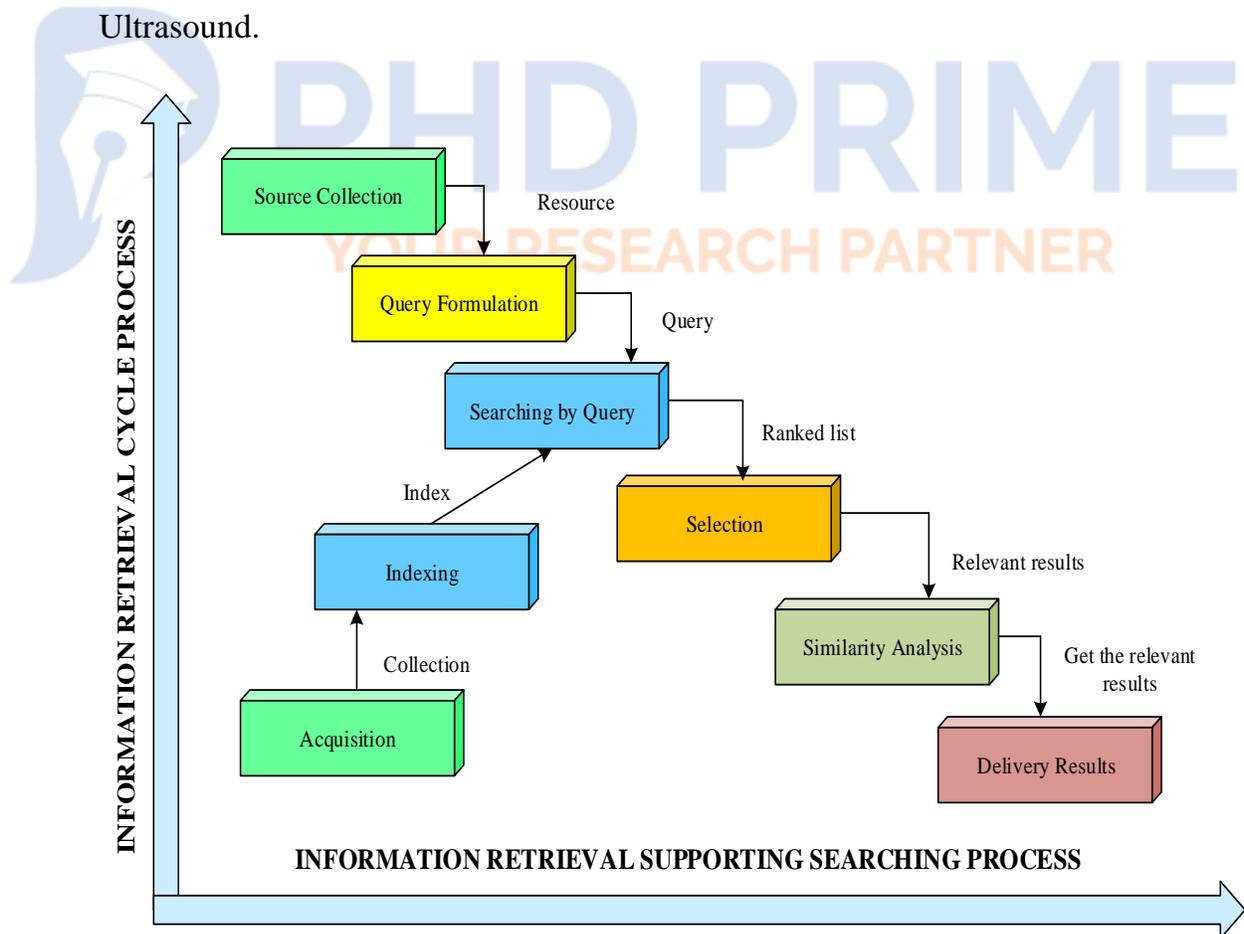


Figure.1.2 Information Retrieval Cycle Process

Processing with images ^[2] includes image acquisition, image pre-processing, storage, image enhancement, restoration and reconstruction, recognition, filtering, image segmentation, enhancement, feature extraction, classification etc. Based upon the concept, type of image is chosen and particular algorithms are involved for achieving effective results. Some of the fundamental steps in image processing are listed along with their requirement,

- **Image Acquisition**

Image Acquisition is the initial step involved to process an image. Image acquisition is defined as creation of photographic images (i.e.) to acquire the action from the source image (real-time camera capture) and consider as input for further processing.

- **Preprocessing**

The input images may contain some errors based on their geometry, brightness, intensities, contrast, blurring, pixel values, etc. such errors are adjusted in this step by utilizing certain mathematical models. With the priori information of the image, pre-processing will overcome image degradation.

- **Segmentation**

It is defined as the process of partitioning the image into sub parts. Segmentation of an image is completed, only when the region of interest in the image is isolated. Image thresholding is one of the commonly used image segmentation technique.

- **Image Enhancement**

It is added only in certain application, it is optional i.e. either it can be used else ignored. Enhancement includes noise removal, highlighting edges, sharpening, focusing the image de-blur regions, improving brightness, color space enhancements in color images, etc.

- **Feature Extraction**

It can be defined from its own name itself i.e. the features present in the image are extracted. Size, shape, composition, location are some of the features that are considered in an image.

- **Image Restoration and Reconstruction**

Degraded images are restored by prior knowledge on performing Image Restoration & Reconstruction. This step is performed either locally or globally over the image area. This process is similar to image enhancement but they are not same.

- **Classification**

It is a method that is being used for separating objects or images with some similarity measures. Classification is performed by supervised classification or unsupervised classification.

The above described steps / process are shown in figure 1.2 that illustrates the fundamental steps involved in processing an image. Images are processed for particular reason either for detecting crime / detecting injurious cells, etc. Likewise sharing and retrieval of multimedia from cloud, Hadoop storages are also becoming popular at present.

1.2 2D VS. 3D IN IMAGE PROCESSING

Two-dimensional (2D) and Three-dimensional (3D) images are the two major classifications of types of images used for image processing. 2D structures are supposed to have only width and height, whereas thickness is not involved. A 3D image consist its third dimension that includes width, height and thickness (i.e.) depth of the object. 2D images are divided into 'N' rows and 'M' columns, in which the intersection of a row and column is defined as pixel. 2D image is defined as two dimensional function ' $f(x, y)$ ' in which the terms ' x ' and ' y ' represents spatial co-ordinates. Then the amplitude in the function of ' f ' at any pair of ' (x, y) ' co-ordinates are called as intensity or gray level. Table 1.1 represents the differences between 2D and 3D representations

Table 1.1 Difference between 2D and 3D Representation

DESCRIPTION	2D MODEL	3D MODEL
Definition	Represents an object with two dimensions (i.e.) Length, Height	Represents an objects with three dimensions (i.e.) Length, Width and Height
Use	Simple 2D objects	Complicated 3D objects
Views Generation	Difficult	Simple
Representation	Flat	Life-like
Co-ordinate System	Cartesian co-ordinate system, Orthogonal co-ordinate system	Spherical co-ordinate system, Polar co-ordinate system
End line Representation	Coordinates : (x, y)	Coordinates : (x, y, z)
Representation of curve edges	Splines, ellipse, circle, etc.	Suitability spaced generators
Representation of complex objects	Sectional views and additional views are required	Complex component is interrupted correctly

CBIR is used for various computer vision and image processing applications

- Art Collections (e.g. fine arts museum of san francisco)
- Medical Imaging (MRI, Ultrasound, CT, PET, etc.)
- Scientific Databases (e.g. earth science, and remote sensing)
- World Wide Web (WWW)

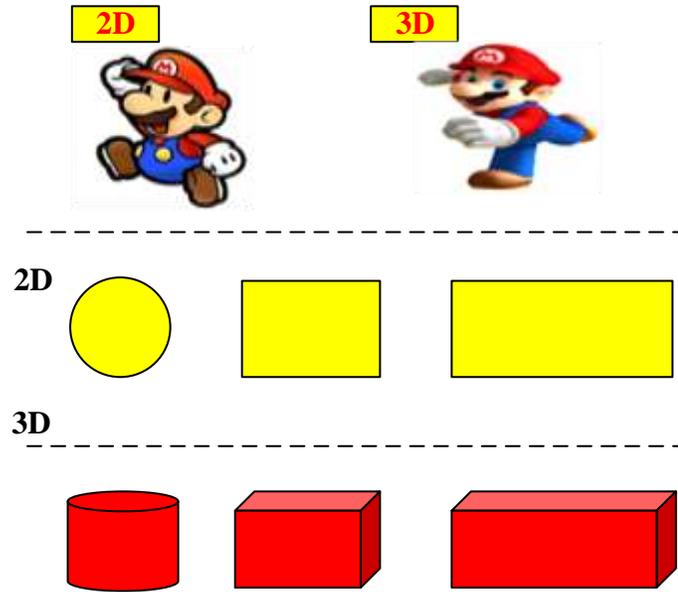


Figure 1.3 2D and 3D Animation

Figure 1.3 depicts the 2D and 3D animation. Image retrieval has become an interesting topic growing its demand rapidly in current years. A large number of images are available in the databases [Syam et al., 2013, Rana et al., 2018]. Image Retrieval classified into two types: (1). Text-based Image Retrieval, and (2). Content-based Image Retrieval.

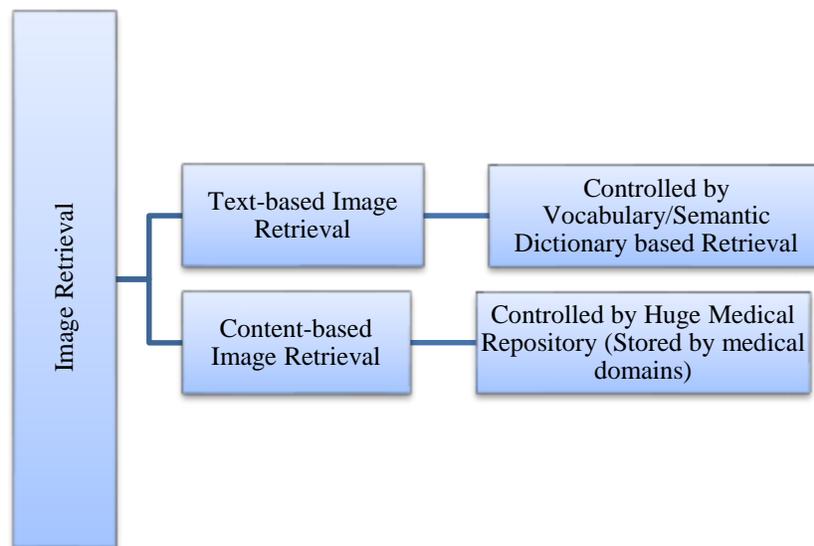


Figure 1.4 Classification of Image Retrieval

Each classification in image retrieval is designated as follows:

- *Text-based Image Retrieval:* In this retrieval, clinical staffs can find the patients current status from the medical database. For this PACS is used, which aids to identify the status from a large medical repository. In text-based image retrieval, the query is “Textual Words”. For textural words, the server response with the relevant images by keyword extraction, query formulation, synonym identification, and similarity computation between the textual query and the medical repository. The major limitation of this type of image retrieval is searching capability is limited to users to view the exact reports for the query words. In addition, the expectation of user is not satisfied because this system retrieves small number of images as relevant so that searching is limited, and lack of scalability. Furthermore, the medical repository ignores the visual and semantic properties for each image.
- *Content-based Image Retrieval:* In this system, users can obtain the more number of accurate images as relevant so it resolves the challenges of scalability, and wide-range of searching support. This retrieval system is based on the Query by Example, in which query image is given to the retrieval system as an example. Then retrieval system performs the matching, and retrieves visually similar images. In retrieval, it match the query image and retrieve the visually similar images by distance measures.

In order to effectively retrieve or index, content-based image retrieval (CBIR) concept is introduced. This term is released in 1992 by T. Kato. CBIR is also called as content-based visual information retrieval (CBVIR) and query by image content (QBIC) [Caicedo et al., 2007, Zhou et al., 2017]. Content refers to the image, text, or video. *The definition of CBIR is given below:*

Content-based Image Retrieval

CBIR aims to retrieve the most top-k images that are visually and semantically similar to the query image from a large collection of images. Hence, this system will automatically index the relevant images in a database by similarity functions or distance measures.

Text-based image retrieval is based on the manual query search by keyword extraction, bag of words collection, and matching with the database. The results of this text-based image retrieval depends on the manually labeling the data by a human, which leads to irrelevant image retrieval, less efficiency, and scalability, and time-consuming operation. The simple image retrieval pipeline is described as follows:

- *Image Preprocessing*

In this step, image quality is improved by removing the unwilling distortions in the image and corrects the geometric transformation of the images (rotation, scaling and translation). Some of the important preprocessing tasks are follows: denoising, contrast enhancement, normalization, and so on.

- *Feature Extraction*

It is the process of converting the given input data into the set of features are known as feature extraction. In medical image processing, feature extraction has started with the initial sets of dependable data and it performs the feature extraction in the borrowed values which is known as features. It makes the classification method to be much satisfactory for prediction. Most of the medical image classifiers are based on the result of feature extraction in CAD.

- *Feature Selection*

Feature selection is the process of reducing the feature space extracted from the whole image, which enhances the detection rate and reduces the execution and response time. It is obtained by neglecting the irrelevant, noisy and redundant

features, and chooses the features subset that able to obtain the good performance with subject all metrics. In this stage, the dimensionality is reduced.

- *Classification*

In this stage the selected features are applied for classification in which each sample is classified to denote as the desired class. However, the classification is a big issue in medical image analysis tasks such as object detection, pattern recognition and so on. For automatic classification, optimum sets of features are used needed to be further improved for better clinical analysis for the human body by disease.

- *Retrieval*

Similar images are retrieved by computing the similarity between the query image and the database images

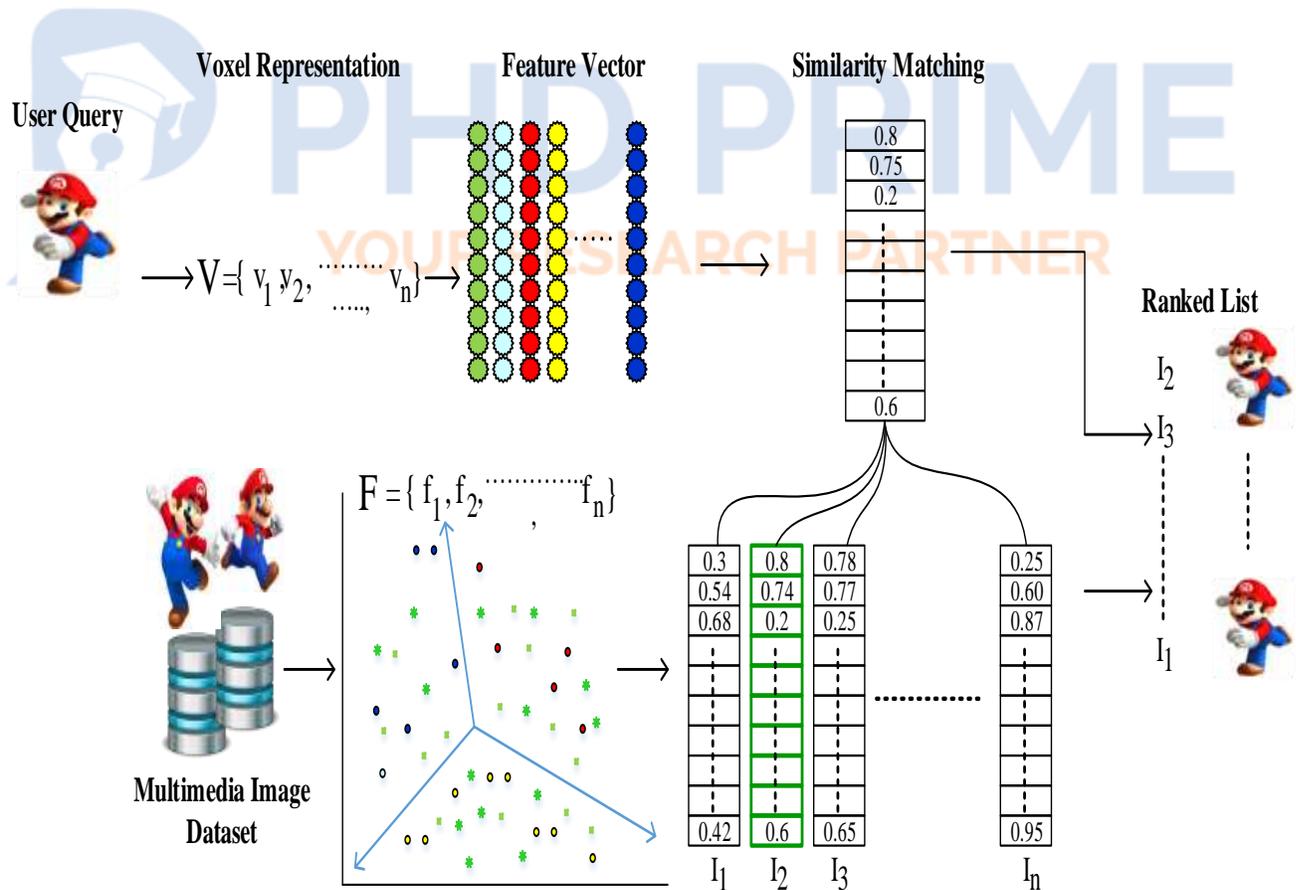


Figure 1.5 Flow of CBIR

Problem Statement in CBIR

- *Assumption* – A query image is given that describes the desired information need of the user. IR system consists of the massive images in the database
- *Task* – Rank all the images in large collection of medical images based on the user demand
- *Require* – User must fulfill the retrieval results and gives feedback regarding the information that acquired.

This research study is considered “Image” is a query. In image, the content can be categorized into three classes: (1). Spatial, (2). Semantic, and (3). Low level content. In spatial content, the presence of object position is analyzed from the image. The synonym of the image is represented as the semantic content, and the visual, color, shape, and texture are considered as the low level content [Mystry et al., 2018]. Figure 1.6 represents the CBIR for the query image, in which the basic flow for retrieving images from the database is clearly depicted.

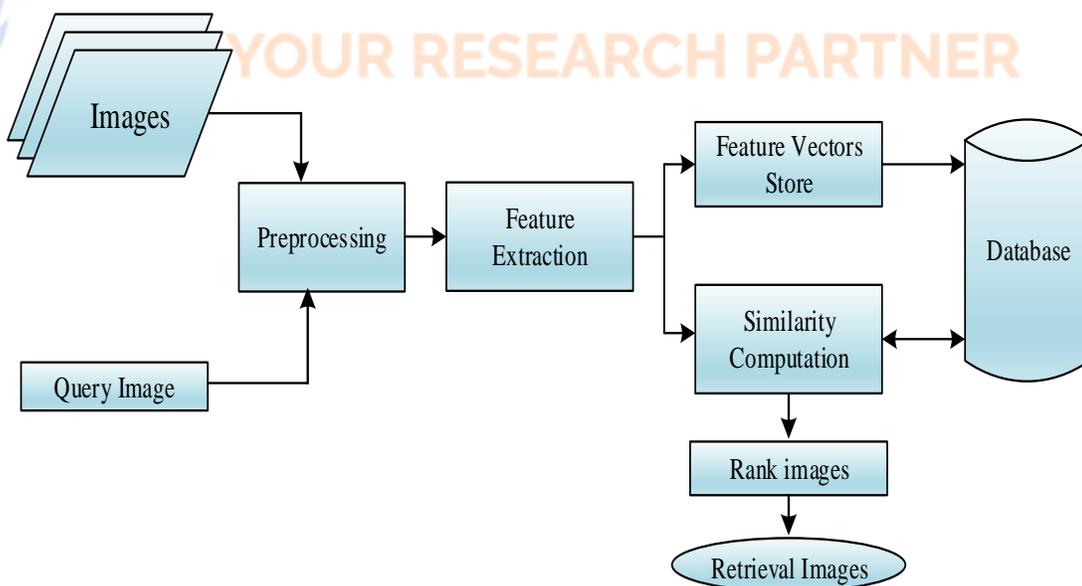


Figure 1.6 Content-based Image Retrieval System

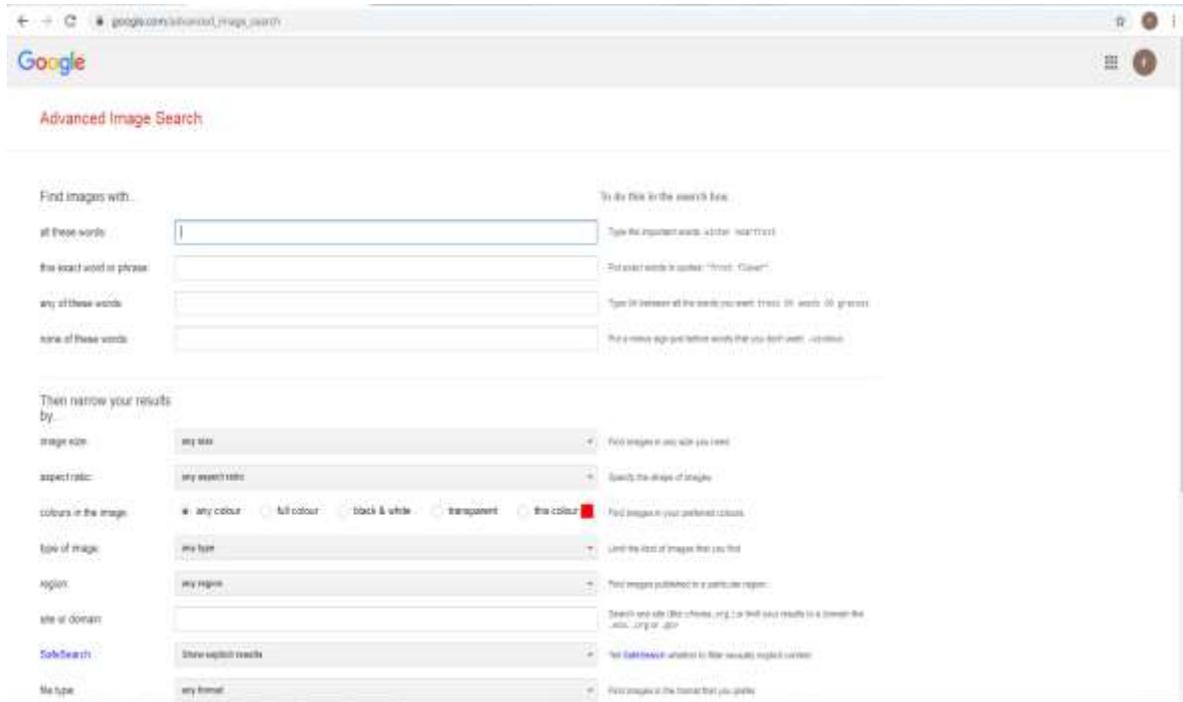


Figure 1.9 Content based Image Retrieval (Advanced Google search)

A consequence of increasing consumer demand for visual information requires sophisticated technology to represent, index, model, and retrieve multimedia information. In particular, robust techniques are efficient to index (retrieve) and compress visual information, new scalable algorithms allowing access to large databases of images and videos. The performance of retrieval process is generally estimated by several significant metrics, which capture the overall efficiency of retrieval process. Most of the researchers have focused on semantic retrieval for improving the efficiency of retrieval procedure and obtain most relevant results in CBVR and CBIR [44],[45]. Semantic retrieval is used, since it provides the most meaning-full information about the input queries. Semantic analysis provides efficient results and as per human perception access to databases, content filtering, summarization, enhanced human and computer interactions, etc. semantic techniques are the promising approach which is used to bridge the gap between low-level features and high-level concepts. Recent computer vision technologies and algorithms are supported for efficient semantic video/image retrieval and analysis. The following sub-sections describe the state-of-the-art image and video retrieval techniques.

1.3 RETRIEVAL METHODS IN CBIR

There are several methods have been presented such as query by example, semantic retrieval, relevance feedback (human interface), machine learning and deep learning. These methods are used for image recognition and classification applications. Figure 1.10 shows the examples for image retrieval. For the query image, set of relevant images and the irrelevant images are given.

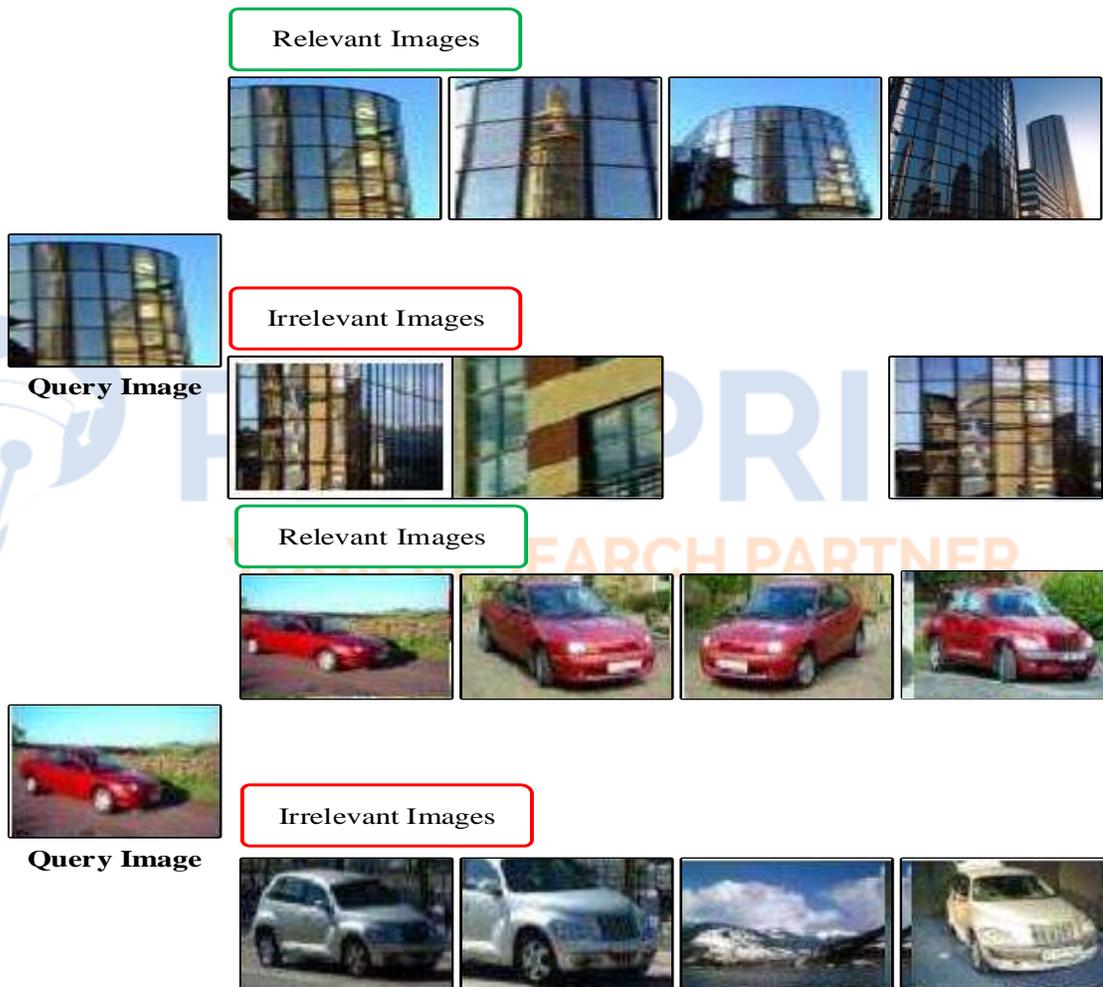


Figure.1.10. Image Retrieval Example

One of the most commonly used method for comparing two images in CBIR is the distance measure.

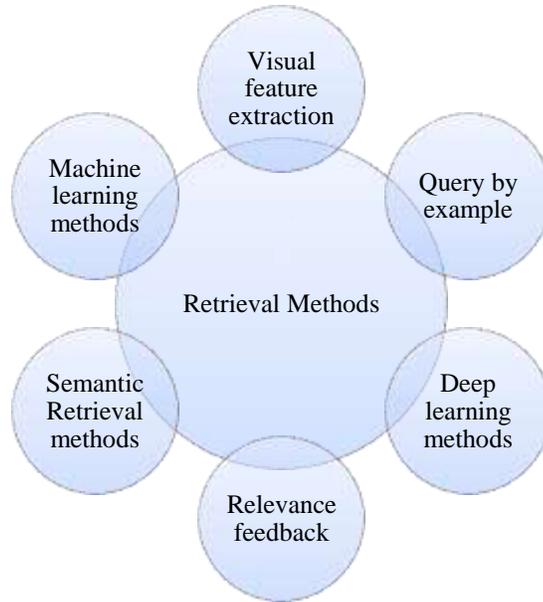


Figure 1.11 Retrieval Methods

Most of CBIR performance is based on the low-level features of the image including Color, Shape, Texture, and Object. The process of CBIR is to retrieve the most similar images that are visually and semantically related to the query image. For effective retrieval of images by similarity, CBIR must depend on appropriate feature set extraction which describing the desired image contents. Furthermore, appropriate query type, matching, indexing, searching, and retrieval methods are required, because images are generated automatically and indexed in the database. In CBIR from large collection of database, query image feature vectors and the database feature vectors are computed with the similarity. A threshold value is specified for database images. The distance values are must be less than the threshold value. Therefore CBIR minimizes the human efforts and runs automatically. In CBIR, there are two challenges are emerging that remains unsolved: visual-gap, and semantic gap.

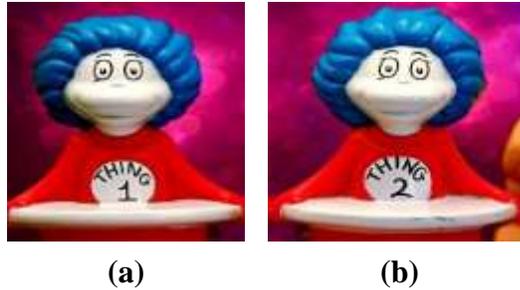


Figure 1.12 Visually similar images (but semantic gaps are presented)

Figure 1.12 (a), and (b) images are visually similar, but they are different due to semantic gaps. Feature extraction is the process of extracting the image features to distinguish the image is visually similar or not. It is a very essential element for understanding the image, particularly for the segmented or preprocessed regions of the image.

CBIR methods are classified into three broad categories ^[1] based on the features such as color features, texture features and shape features. Histogram based methods and statistical features comes under the category of color features, then spectral features and statistical features in texture features and lastly model based, region based and boundary based methods in shape features. CBIR is capable to match digital images with its visual content present in the image. The visual content of an image significantly consists of contents,

- Color feature
- Texture feature
- Shape feature

1) Color Feature

Color feature is one of the significant feature for identifying similarities by measuring global and local points in image. Usage of color layout features will require a refining technique. In retrieval process color features can be measured as color histogram, color

moments, dominant colors and Color Coherent Vector (CCV). Colors have been represented in different color spaces ^[7] as

- Red Green Blue (RGB)
- Hue Saturation Value (HSV)
- Hue Saturation Brightness (HSB)
- Cyan-Magenta-Yellow (CMY)
- Luminance-Inphase-Quadrature (YIQ)
- Hue Saturation, Lightness / Luminance (HSL)

The most important representation of color models are RGB and HSV which are effectively applied in image processing, image analysis, image storage and computer graphics processing

RGB color model is composed with primary colors Red, Green, and Blue, which is mostly used in color CRT monitors and color raster graphics. This model is represented as “additive primitives” since the colors are added together to produce the desired color. Figure 1.13 represents the Cartesian coordinate system; the colors are defined as vectors with diagonal representations, starting from (0, 0, 0) black to (1, 1, 1) white. The color of pixel (p) is a linear combination of the basics vectors R, G, B written as,

$$\Phi_p = r_p \hat{i} + g_p \hat{j} + b_p \hat{k} \quad (1.1)$$

From the above equation, the components of RGB are composed and formulated and this plays a major role in feature extraction for the achievement of accurate results based on the defined processing.

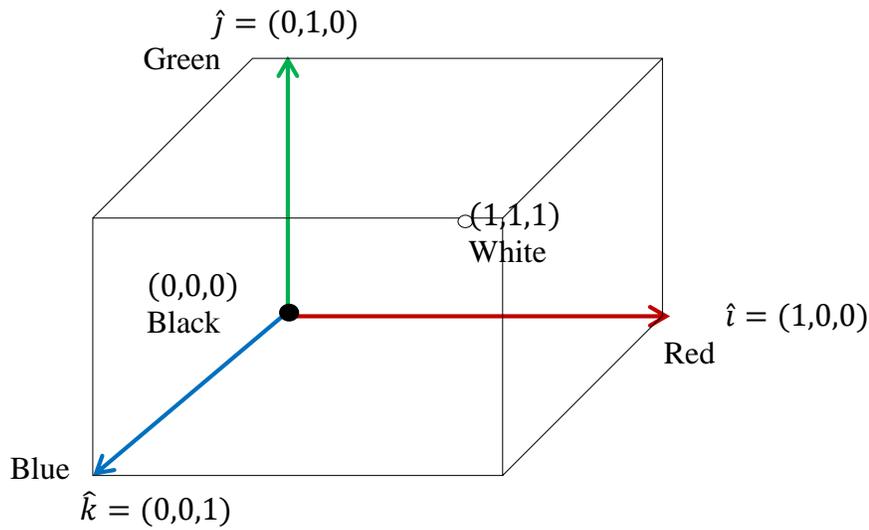


Figure 1.13 RGB color space Cartesian coordinate system

HSV stands for Hue, Saturation and value based, which represents the intensity of color that is decoupled form of color information in the represented image. Hue is referred as the number which identifies the location of corresponding pure color on the color model. The hue fraction is represented between 0 to 1 where '0' is red, '1/6' is yellow, '1/3' is green and forth among color smodel. Saturation refers the position of white color, a pure red is fully saturated i.e. represented as 1, tints of red are saturated with less than 1 and white is saturated at 0. Value represents the lightness of color, dark color has '0' (black) with increasing forwarding away from black.

Figure 1.14 represents the single hexcone of HSV color models, which contains H, S, V models with its representation. The color of pixel (p) in the color space with its elements H, S, V is written as

$$\varphi_p = [h_p + s_p + v_p] \quad (1.2)$$

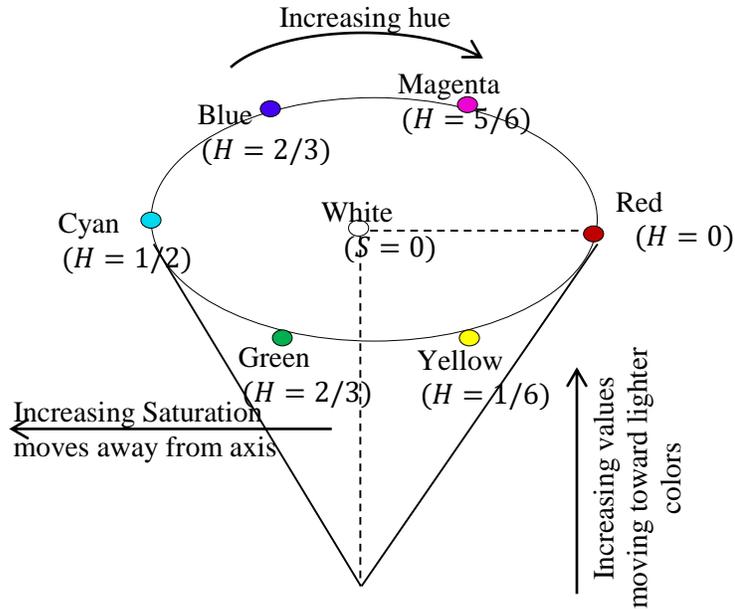


Figure 1.14 Single-hexcone model of HSV color space

The color space model is often used for computing the similarity between two colors. The similarity between two colors i and j is given by,

$$C(i, j) = w_h H(i, j) + w_s S(i, j) + w_i(i, j) \quad (1.3)$$

Where

$$H(i, j) = \min(|H_i - H_j|)$$

$$S(i, j) = |S_i - S_j|$$

$$I(i, j) = |I_i - I_j|$$

The degree of similarity between two colors i and j is given by,

$$S(i, j) = \begin{cases} 0 & \text{if } (H > H_{max}) \\ 1 - \frac{C(i, j)}{C_{max}} & \text{otherwise} \end{cases} \quad (1.4)$$

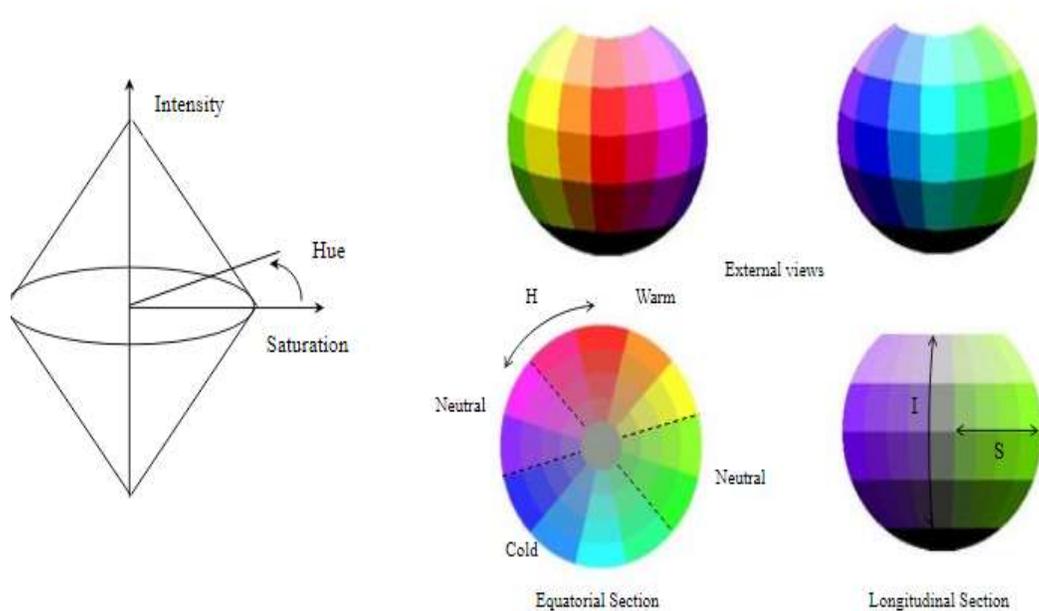


Figure 1.15 HIS Color Space Model

2) Texture Feature

Texture features are mainly used for the purpose of accurate segmentation and classification results. Image textures are defined as quantitative measures of the intensities (roughness), edges and direction in the region. Texture features are referred as surface characteristics i.e. appearance of a particular object. Gray co-occurrence matrix is one of the significant method that is used for feature extraction. This feature is considered in the applications of medical imaging, remote sensing and CBIR.

A Gray level co-occurrence matrix is a statistical method of examining texture that defines the spatial relationships of pixels. It is represented as rows and columns which are equal to number of gray levels. The matrix element $P(i, j|x, y)$ is the relative frequency of two pixels which is separated by pixel distance between neighboring pixels (x, y) , where 'i' and 'j' are the intensity values of 'x' and 'y' respectively. The GLCM has four different properties that are illustrated in table 1.2.

Table 1.2 Properties of Gray level Co-occurrence Matrix

Texture Feature	Definition	Formula
Energy	Energy is defined as sum of squared elements in GLCM, also called as Angular second moment or Uniformity	$\sum_i \sum_j P^2(i, j)$ (1.5)
Correlation	It measures the linear gray level dependency of neighboring pixels	$\sum_i \sum_j P(i, j) \log P(i, j)$ (1.6)
Contrast	The intensity contrast is measured between pixel and neighboring pixel	$\sum_i \sum_j (i - j)^2 P(i, j)$ (1.7)
Homogeneity	Homogeneity is inversely proportional to contrast in terms of equivalent distribution in pixel pair population	$\sum_i \sum_j \frac{P(i, j)}{1 + i - j }$ (1.8)

3) Shape features

Shape feature is also one of the significant visual feature present in an image. Shape is a feature that is difficult to describe the image but partially the object representation could be projected. Shape is determined with the following parameters ^[3] as Mass, Centroid, Mean, Variance, Dispersion, Eccentricity, Convexity, Elliptic Variance, etc. Two different ways are termed to extract shape feature:

- Internal Representation- Use of object boundary and its features (e.g. boundary length)

- External Representation- Description of region occupied by the object on image plane

Figure 1.16 illustrates the shape feature extraction techniques, which is broadly classified into two categories such as region based and contour based methods. Shape-based image retrieval is processed by measuring similarity between shape features.

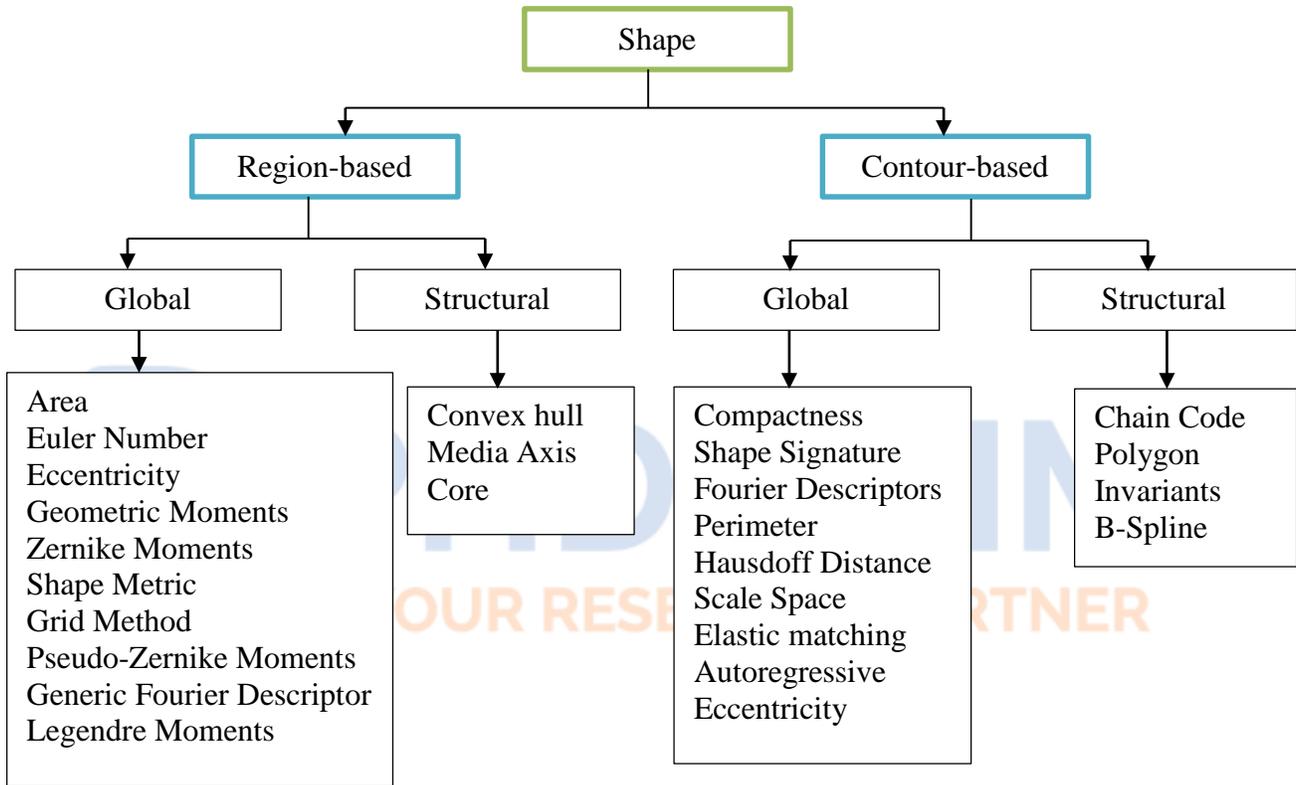


Figure 1.16 Shape Representation and several extraction processes

The other classes of features are described as follows:

- **Semantic or Relational features**

It is extracted by the object motion and moving trajectory. It is due to the camera motion and activity in the semantic descriptor, which describes the motion scene.

- **Statistical Features (pixel level)**

Geometrical features (position, location, and orientation) are considered as the statistical features. In images it is computed by the translation, rotation, and scaling operations.

1.4 CBVR OVERVIEW

CBIR have been gradually moved towards Content Based Video Retrieval (CBVR) with increased users in World Wide Web (WWW). Usually videos include audio, texts, objects and faces of living being.

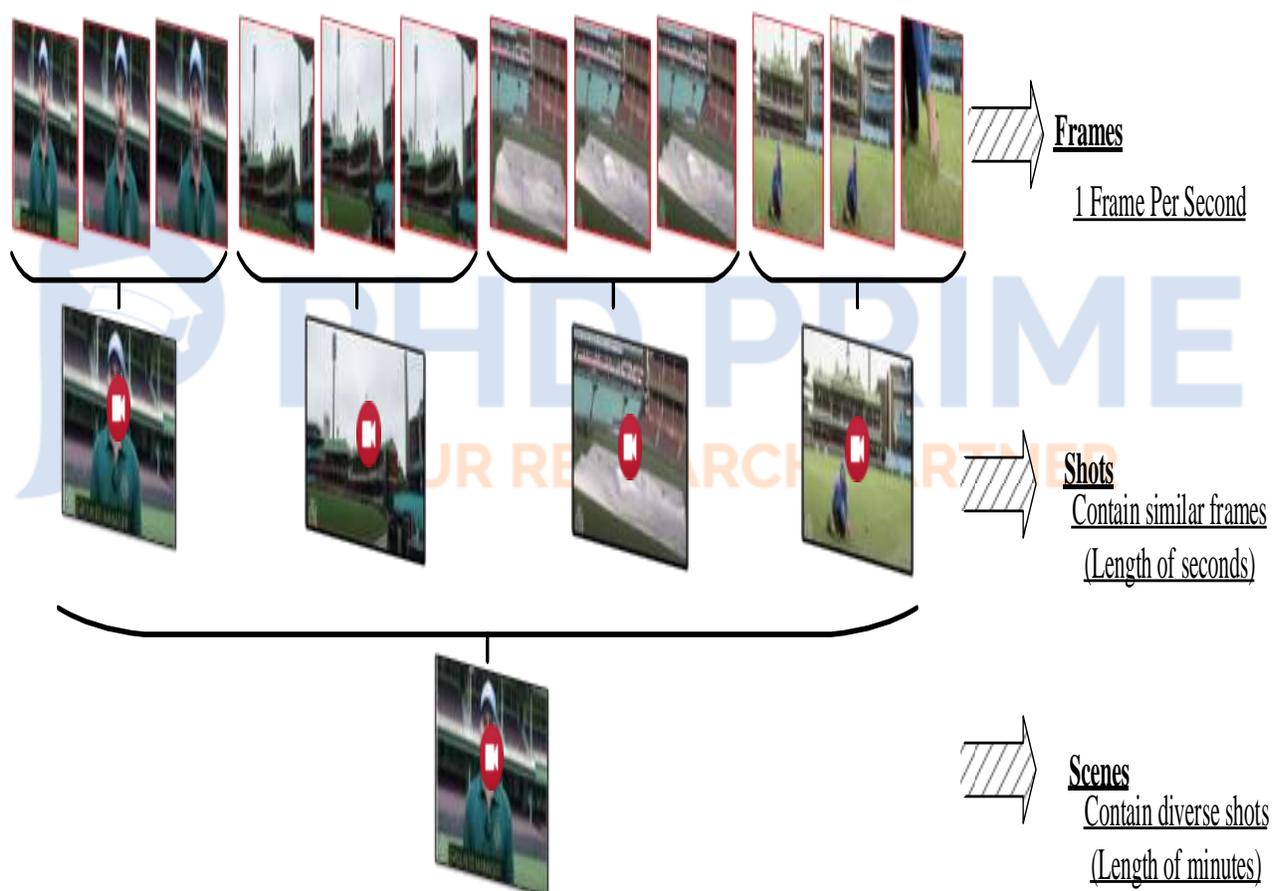


Figure 1.17 Spatial and Temporal Structure of Videos

However, frames are extracted at 1 frame per second (fps). Shots are number of frames that taken without any scene change from the camera. The length of frames is based on the seconds order. Scenes are longer video segments that consist of interrelated shots and represent a semantic unit for the given type of content. For instance, news

content consists of a news story. Due to different angles, objective, subjective, point of view, high angle, low angle and dutch angle. Similarly, different types of shots are presented as follows,

- **Extreme Long Shot** – It represents the vast area from a great distance and also it known as a wide shot
- **Long shot** – It takes the entire action for a scene. A shot that sets up the scene and also provides the common information about the event.
- **Establishing Shot** – This is a camera shot that sets up what is about to take place and this is usually a sequences about the starting point.
- **Medium Shot** – This shot falls between the close up and the long shot.
- **Close Up Shot** – It is also known as narrow shot or tight shot which closure to the particular frame

Main Challenges in Handling Videos

As there are number of challenges while processing videos. These challenges are highlighted in following:

(i). Temporal Redundancy: The same frame is repeated twice and starts a completely new frame from time to time. Keep and re-use some bits of the previous frame or next frame. Determine significant changes and just show the differences

(ii). Motion Difference: Each frame in a video is presented with unique motion features, color histograms, motion histograms, audio features, etc. Motion information includes certain parameters as motion content, motion uniformity, motion panning and motion tilting.

- Motion Content is the measure of total action content present in a video is defined as motion content. Variations in motion content are based on the video that is

considered. For instance a vehicle crash video will have higher motion content whereas a person capturing videos will have minimum motion contents.

- Motion Uniformity parameter is to measure the motion's smoothness in a video with the corresponding function of time.
- Motion Panning is defined as Motion of the camera alters in directions left to right and right to left during capturing.
- Motion Tilting is the vertical motion component are measured under this metric, here if vertical motion is higher in a video, then panning shots will be lower.

Motions are considered to be more significant part in each video, changes in motion is the main difference between image and video. In CBVR concept, videos are converted into frames (i.e.) a shot of videos consists of a set of frames shown in figure 1.17.

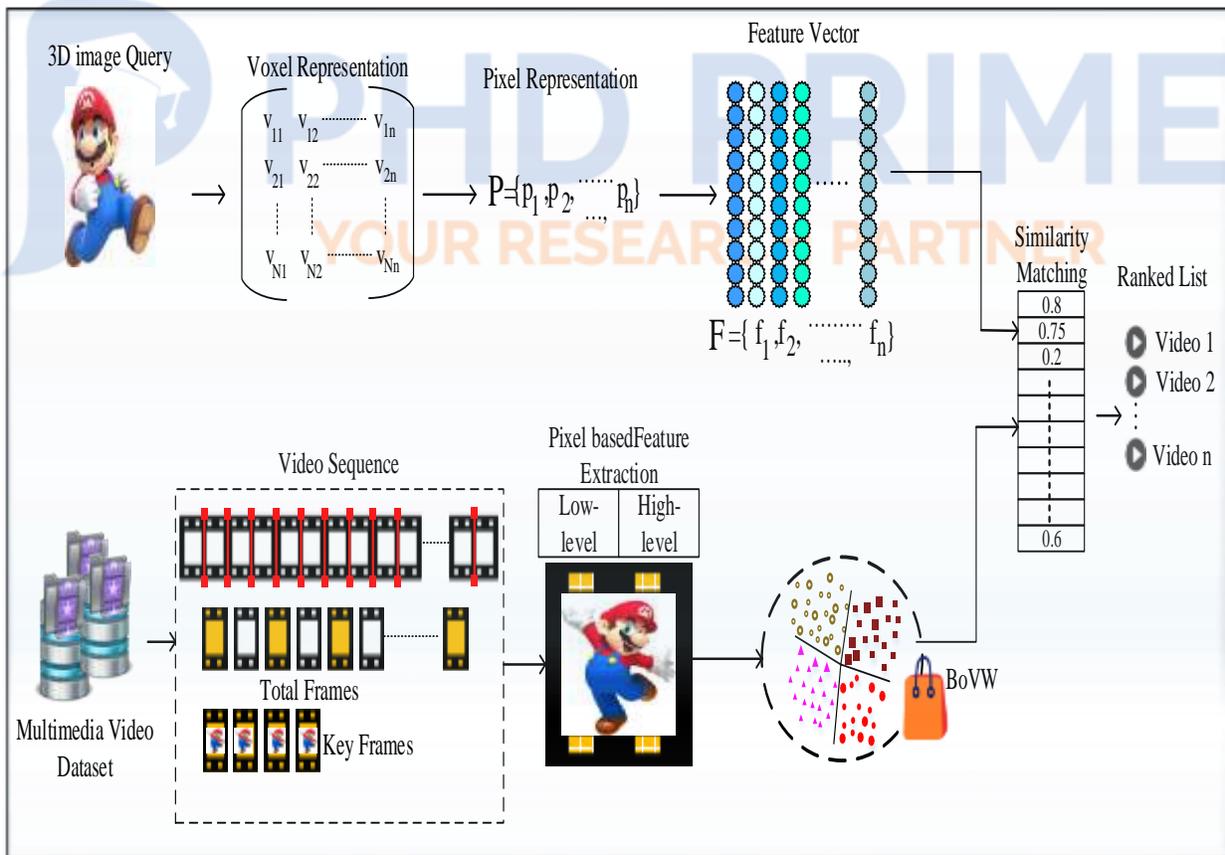


Figure 1.18 3D Image Based CBVR

Based on the size and quality of the video, the number of frames is increased. These converted frames are considered for further processing to retrieve relevant results for the given query. To simplify the process of CBVR with larger number of frames, a process of key-frame extraction is included. Video retrieval approaches mainly focus on spatial and temporal analysis. In CBVR concept video databases are maintained which has shot as the basic unit of data.

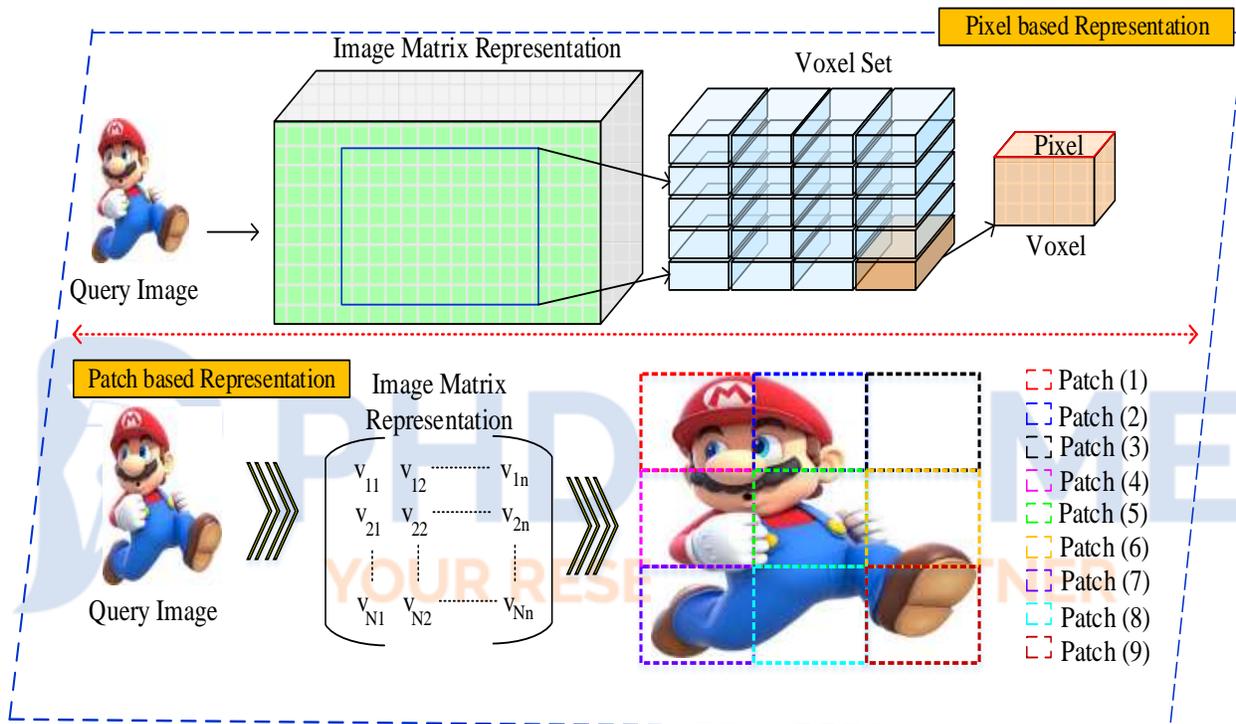


Figure 1.19 Pixelwise & Patchwise Image Representation

Generally, video processing is processed by two methods that are Pixel based Representation and Patch based Representation. In pixel based processing, image is processed by each pixel values and each pixel is compared to the nearby pixels, then only all significant information in the image is controlled. Pixel – A pixel is the key element that builds a complete multimedia content (image, video). Pixel as P_{ij} represents the value of i^{th} color channel at j^{th} image pixel. Pixel-by-Pixel handles pixel-wise process by taking in account of individual pixel surrounded by neighboring pixel. Whereas patch

based processing is also known as block based processing. Here, frames are divided into non-overlapping patches and each patch is compared with its previous patch, then changes in patches are updated. Patch – A patch is defined as a set of pixels that exists in a particular region of the image. Typical patch sizes based on image size - 4×4 , 8×8 , 16×16 , etc. Patch-by-Patch processes are performed in accordance to the individual patch that is surrounded by other patches in the image. For data management five methods are followed, the methods are,

- Metadata-based method
- Text-based method
- Audio-based method
- Content-based method
- Integrated approach

In metadata-based method the videos are indexed and retrieved with respect to the structured metadata information. Some of the examples for metadata information are title, author, producer, director, date, video types, etc. If videos are retrieved based on the associated subtitle then they are called as text-based method. Similarly if videos are retrieved with sound tracks, then it is defined to be audio based method. Next, content based method is performed with two different possibilities they are (i) a video is considered as a collection of images and (ii) video sequences can be divided into groups of frames. For enhancing the flexibility of video retrieval two or more above mentioned methods are combined and used in CBVR.

In ‘Content Based Video Retrieval’ the term ‘content’ refers to shapes, color, textures, geometry, topology or other information present in the image. Video content can also be retrieved without examination of such features (i.e.) by using the video’s metadata such as keywords, captions, sub-titles, etc.

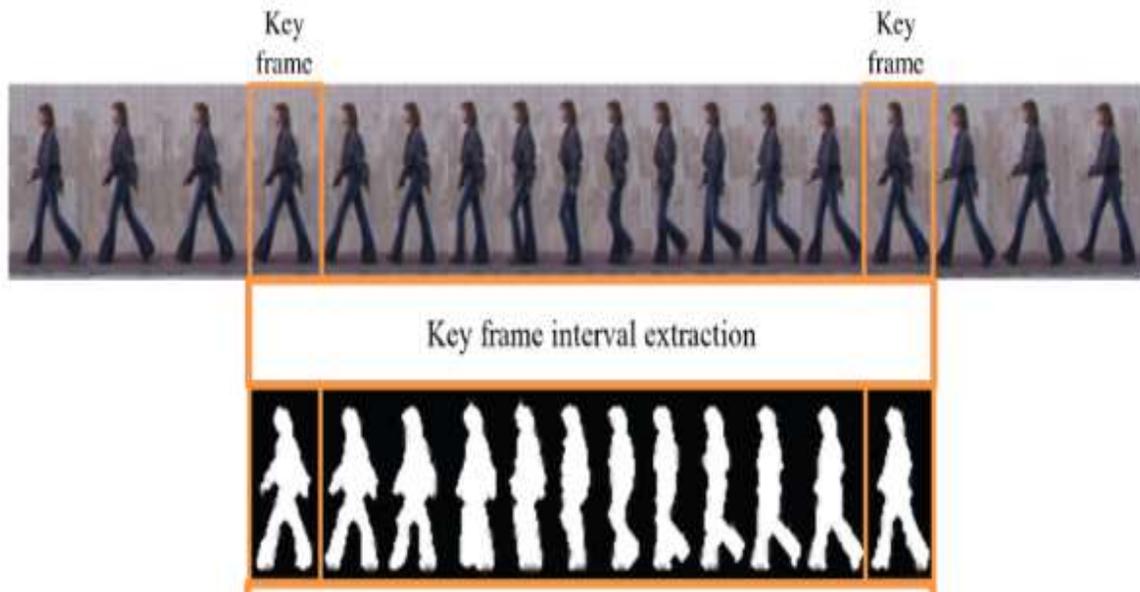


Figure 1.20 Key Frame Extraction

Many of the researchers have considered the features that are present in video frames. Based on the size of the video, number of frames increases so in CBVR, processing with frames becomes complex if the video size is larger. For this purpose, the key frames are extracted from the total number of frames present in the video. Key frame extraction is shown in figure 1.20 that visualizes a video of a human walking.

1.4 BACKGROUND OVERVIEW

1.4.1 Semantic Search Engine

Generally semantic means the study of 'language'. Visual semantics are structural and are communicative. It is defined as a part of semantic that deals with the knowledge about the visual aspects of elements around. Images speak out visually with colors, shape, textures and reactions of characters in the image. Images communicate visually with expression and not with words. Visual semantic includes visual content from surveillance cameras, mobile phones, personal photo collections, news footage, or medical images.

The main properties of semantic analysis can be the following,

- Semantic level provides specific knowledge to objects
- This level bridges the gap between structural and geometric levels by identifying meaningful representation.
- This semantic label is effectively used for analyzing the process of 3D shapes in video retrieval

Semantic web based framework is more popular among people for retrieving images and videos. Google, Yahoo and Bing are the well-known query processing search engines that handles several keywords and produce efficient results. Several search engines are available today for retrieving meaningful information to the corresponding query. Ontology is defined as a significant concept that is used in semantic web infrastructure (in fig 1.9), Resource Description Framework and Web Ontology Language. In simple a vocabulary can be defined as the pillars of semantic web. Semantic web is comprised of resources as webpages, text files images, audios, videos, etc. Available semantic web technologies are Extensible Markup Language (XML), Resource Description Framework (RDF) [4], Metadata, Ontology, Database and Metadata storage technologies, Information management and Knowledge management. Semantic web was developed to solve the following problems as shortage of web content, poor interconnection, shortage in information transfer, knowledge less machines to identify the universal format. The main goals of semantic groups are as follows:

- Semantic web technologies are used for making existing multimedia metadata standards interoperable, therefore previous metadata formats are combined to improve the process.
- Rule based approaches are used for semantic web and its practical applications provide additional functionality to formal semantics
- Semantic web effectively provide the best practices to make multimedia metadata and using multimedia content on web with practical use cases.

Major issues in semantic search engines are lower precision, higher recall, inappropriate queries, irrelevant results, etc. The main aim of semantic search is to enhance the searching accuracy by means of identifying and understanding the user query's intent and contextual meaning appears on the data space for providing most relevant results.

1.4.2 Hadoop MapReduce

The MapReduce is the programming model which has capability to execute the numerous amounts of raw data resolving the number of allocable using comparatively inexpensive product hardware (Srinivasa *et al.* 2015). It is a suitable framework for proficient huge scale data processing environment. It is evolved to overthrow the problems of executing the numerous amounts of data with the reference to the internet oriented applications. These huge size data essential to be indexed, deposited, recovered, examined and also excavated to permit a modest and endures admission to these data and information. In present days, there are four sequential aspects exist in the present business and inventiveness that are handling, storage, picturing and investigating the huge amount of data. The MapReduce can routinely execute the applications on analogous cluster of hardware. Besides, it has the ability to execute the terabytes of data more promptly and proficiently. Henceforth, the reputation of MapReduce process has developed speedily for dissimilar varieties of enterprises in numerous fields. It delivers the extremely operative and proficient framework for the similar implementation of the applications, data distribution in distributed database systems.

The developers who utilize the library of MapReduce must deliberate the two significant processes such as Map and Reduce function (Hashem *et al.* 2016). Here, the Map function receives the key/value pair as input and creates the intermediate key/value pair for further processing. The reduce function combines whole the intermediate key/value pair and then creates the final output. The Map Reduce framework is implemented in the Hadoop open source software. It provides the upcoming

features such as energy efficiency of jobs, scheduling of jobs and tasks, performance, elasticity and load balancing of cluster.

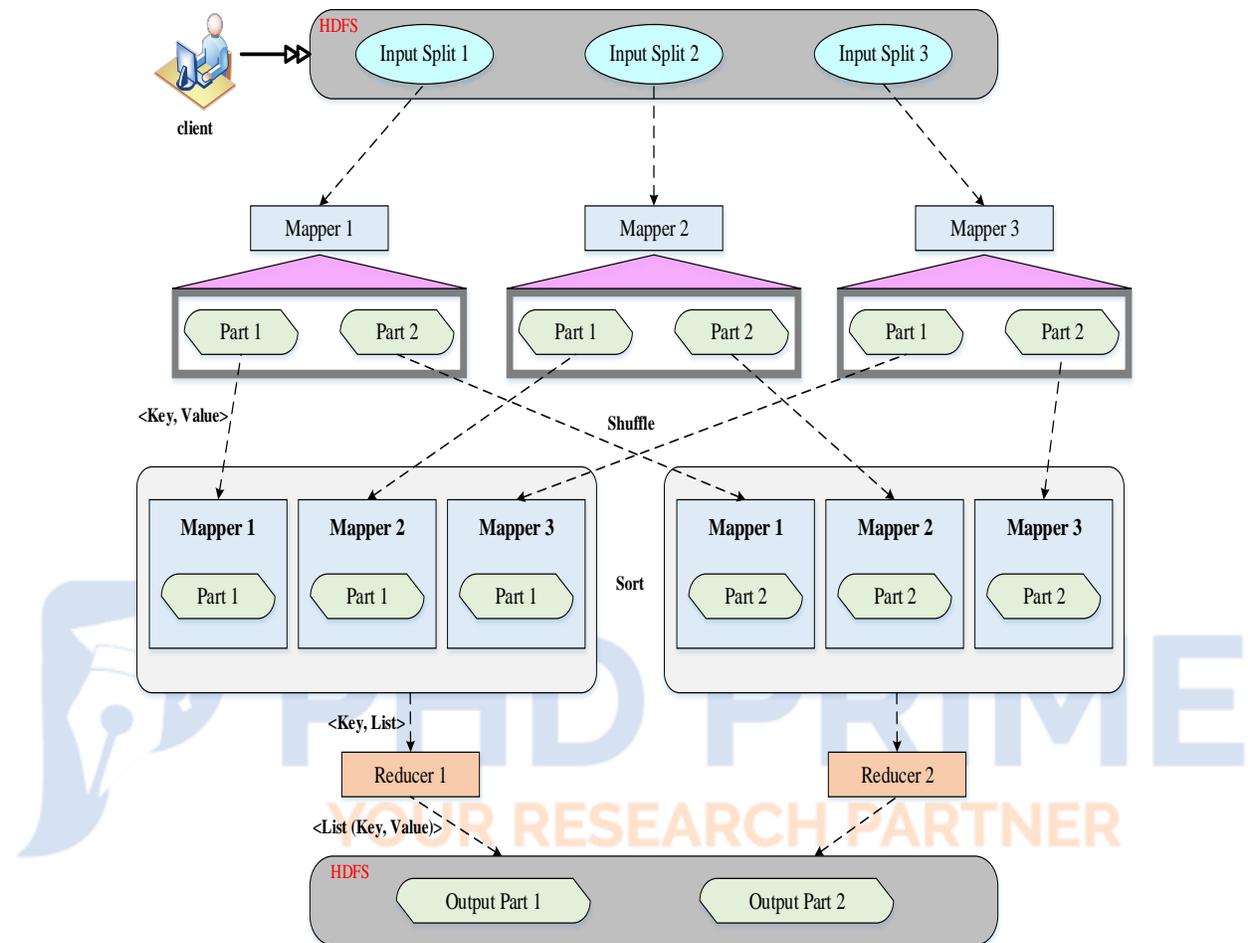


Figure 1.21 MapReduce architecture

Figure 1.21 elucidates the MapReduce architecture implemented in the Hadoop framework with its building blocks (Li *et al.* 2014). It comprises three major processes that are discussed briefly as follows:

Mapper

- The mapper phase is initial phase of MapReduce framework where the provided input is going to segregate into two components such as key and value. Here, the

key is writable and equivalent in the handling stage. Then the provided input is split into the numerous input splits.

- The input splits are the rational splits in environment. Here, the record reader translates these input splits in Key-Value pair. It is the real input data setup for the mapper input for auxiliary handling of the data that exists in the Hadoop system. The input format type changes adaptively for each application. Hence, the programmer has to perceive the input data and code consequently. In this, partition and combiner rationalities are come into the mapper process only to execute the special data process.
- The combiner is also known as the mini reducer. For huge amount of data processing in Hadoop environment requires the high network bandwidth. To overwhelm this issue, the combiner phase is included after completing the mapper process. The partition module in the Hadoop framework plays a significant role in the process of separating the data acquired from either diverse mappers or combiners.

Shuffling and Sorting

- The shuffling and sorting is the intermediate process in the Hadoop system to perform the MapReduce process. After completing the mapper process, there exist huge amounts of middle data to be moved from all the Map nodes to shuffler. It sorts the key of the given input hence the whole pairs with the similar value key is gathered together. Besides, it is obligatory to transfer the sorted output to the reducer nodes for further MapReduce process.

Reducer

- The reducer is the final process in the MapReduce framework. The reducer process obtains the transitional key and group of values of the given key. It gathers all the incoming data to create the smaller set of values i.e. it combines the same

key value pairs to provide the output. In reducer, record writer module writes the data from the reducer to Hadoop environment.

The MapReduce framework is implemented to be fault tolerant since disappointments are a general phenomenon in huge scale disseminated computing. The MapReduce architecture performs to be a better choice for the following reasons:

- The processing of information yields from similar and dispersed architecture with easier software design process of map and reduce methods.
- The MapReduce architecture has the ability to process terabytes of data on the system clusters along with the handling failures.
- The most of the data finding and mining statistics can be occupied into the MapReduce architecture, identical to the pattern based explanation algorithms.

The MapReduce framework has several implementations such as Mars, Phoenix, Hadoop and Google's implementation. In these, Hadoop become the utmost popular software owing to its open source feature. The most standard implementation of the MapReduce model is the Hadoop framework that lets applications to execute on huge clusters. The Hadoop framework qualifies the dispersed, data intensive and analogous applications through investigating an excessive task into the lesser tasks and enormous dataset into the slighter partitions in a manner that each job processes a different partition in parallel. The Hadoop framework permits the dispersed processing of huge datasets through clusters of computers using the particular programming paradigms and models. It is modeled to speed up from a one server to thousands of nodes. It is modeled to estimate the failures at application level rather than depend on hardware for high accessibility. The significant features of the Hadoop system are enlisted as follows:

(i). Distributed Processing

In Hadoop data is processed in analogous way on a cluster of diverse nodes since it is protected in a disseminated manner in Hadoop distributed file system through the cluster of nodes.

(ii). Open Source

The Hadoop framework is called as an open source project. Hence, it can flexible to change based on its business necessities.

(iii). High Availability

Owing to the number of duplicates of data is huge; data can be obtainable and arranged even with the hardware failure. If any hardware smashes of machine is occur, then the data will automatically collect from other pathway.

(iv). Scalability

At every time any new hardware is effortlessly accommodate to the provided nodes, hence the Hadoop framework is known as extremely scalable.

(v). Economy

Hadoop framework is not expensive, since it can be applicable to the generally associated hardware. Hadoop provides the large cost cutting and it has very modest to add numerous machines on that cluster.

(vi). Fault Tolerance

In Hadoop, there exist three copies of each block transversely in the cluster in default manner. Besides, it can be changed based on the provided requirement. Hence, if any node happens to be down, then the data will be acquired from any other nodes in simple way.

(vii). Data Locality

The Hadoop is performed based on the basic principle of data locality. At any time any user provides the MapReduce process, this process will be migrated to data in the cluster instead of transmitting data to the place where the method is succumbed.

The Hadoop provides the most reliable storage layer for the wide spread database to permit the huge bandwidth data coursing to user applications. The distributed file system in the Hadoop environment is modeled to execute the product hardware. By allocating the accommodation and calculation across numerous servers, the resource can scale high and low with the requirement while residual inexpensive. The Hadoop

distributed file system (HDFS) has numerous correspondences with the other distributed file system. However, the alterations are noteworthy. They are discussed as follows:

- It is greatly fault tolerant and it is intended to execute on low rate hardware.
- It provides the large throughput to the stored data; therefore it can be utilized to accumulation and execute large datasets.
- It utilizes the write-one time-read-countless models which satisfies concurrency necessities thus delivers the modest data coherency and permits the high throughput data access.

The HDFS conceits in the belief and offers to be extra effectual when the dispensation is completed adjacent the data instead of moving the data to the applications area. The data engravings are limited to single writer at a moment. The HDFS has various prominent goals that are deliberated as below:

- Confirming the fault tolerance via estimating the liabilities and employing fast rescue methods.
- MapReduce streaming is provided to access the given data.
- Modest and vigorous coherency model.
- Execution is transmitted to the data instead of the data to the execution.
- Scalability in storage and handling huge amounts of data.
- Distributing data and processing across clusters economically.
- Dependability by repeating the data through the nodes and redistributing execution in the event of mistakes.
- Maintaining dissimilar product hardware and operational systems.

Characteristics of Hadoop Distributed File System

In this section, the Hadoop distributed file system characteristics are discussed in detail. The Hadoop distributed file system provides wide range of beneficiary to the pattern mining technology. The characteristics of the Hadoop distributed file system are listed as follows:

- **Large Dataset**

The HDFS based applications are feed by huge scale database. A classic file size exists in the kind of huge gigabytes to short terabytes. It should offer great bandwidth for given data and maintaining millions of files across hundreds of nodes in a one cluster.

- **Hardware Failure**

In general, hardware failure is communal in the clusters. The Hadoop cluster comprises of thousands of machines and each of which stocks a block of data. The HDFS contains a wide range of components; hence there is a better chance of disappointment among them at several point of moment. The identification of faults and the capability to rapidly recover is portion of the essential architecture.

- **Simple Coherency Model**

The write-single-read-numerous contact method of files allows the huge throughput data admission as the data single time written must not be altered. Thus simplifies the data coherency problems. A MapReduce oriented application considers benefit of this model.

- **Streaming Data Access**

The streaming data access is highly significant in the characteristics of the Hadoop distributed file system. The applications that executed on the Hadoop HDFS required being process the flowing data. These applications need not be executed on the common purpose file systems. HDFS is modeled to allow huge scale batch processing which is permitted by the huge throughput data access.

- **Moving Compute Instead of data**

Any calculation is effective if it implements adjacent to the data since it evade the network transfer issues. The transferring the calculation neared to the data is a keystone of HDFS based programming models. HDFS offers entire demanded application interfaces to transfer the calculations nearer to the data afore the implementation.

- **Heterogeneous hardware and software portability**

The HDFS is modeled to implement on the product hardware that hosts the various platforms. This feature supports extensive assumption of this policy for huge scale calculations.

Building Blocks of Hadoop

This portion provides the building blocks of the Hadoop environment (Erraissi *et al.* 2014). The HDFS framework constructed in tandem with the widespread master/slave architecture for both dispersed storage and distributed calculation.

Figure 1.22 depicts the architecture for HDFS environment. This architecture represents the four nodes that are data node, name node, job tracker and task tracker. The explanation for each block in the HDFS architecture is provided as follows:

- **NameNode**

The HDFS cluster comprises of the master server which manages the file system namespace and regulates file access called as the NameNode. The namenode is the master of HDFS framework which guides the slave datanode to execute the small level Input/output tasks. The namenode is defined as the bookkeeper of the HDFS framework; it retains the tracks of how the consumer provided files are fragmented into the file blocks, which nodes stock those blocks, and the whole strength of the distributed file system effectually. The operations of the namenode are listed as the memory and the input/output concentrated. Intrinsically, the server accommodating the namenode generally doesn't stock any consumer data or accomplish any calculations for a MapReduce platform to lesser the load on the engine.

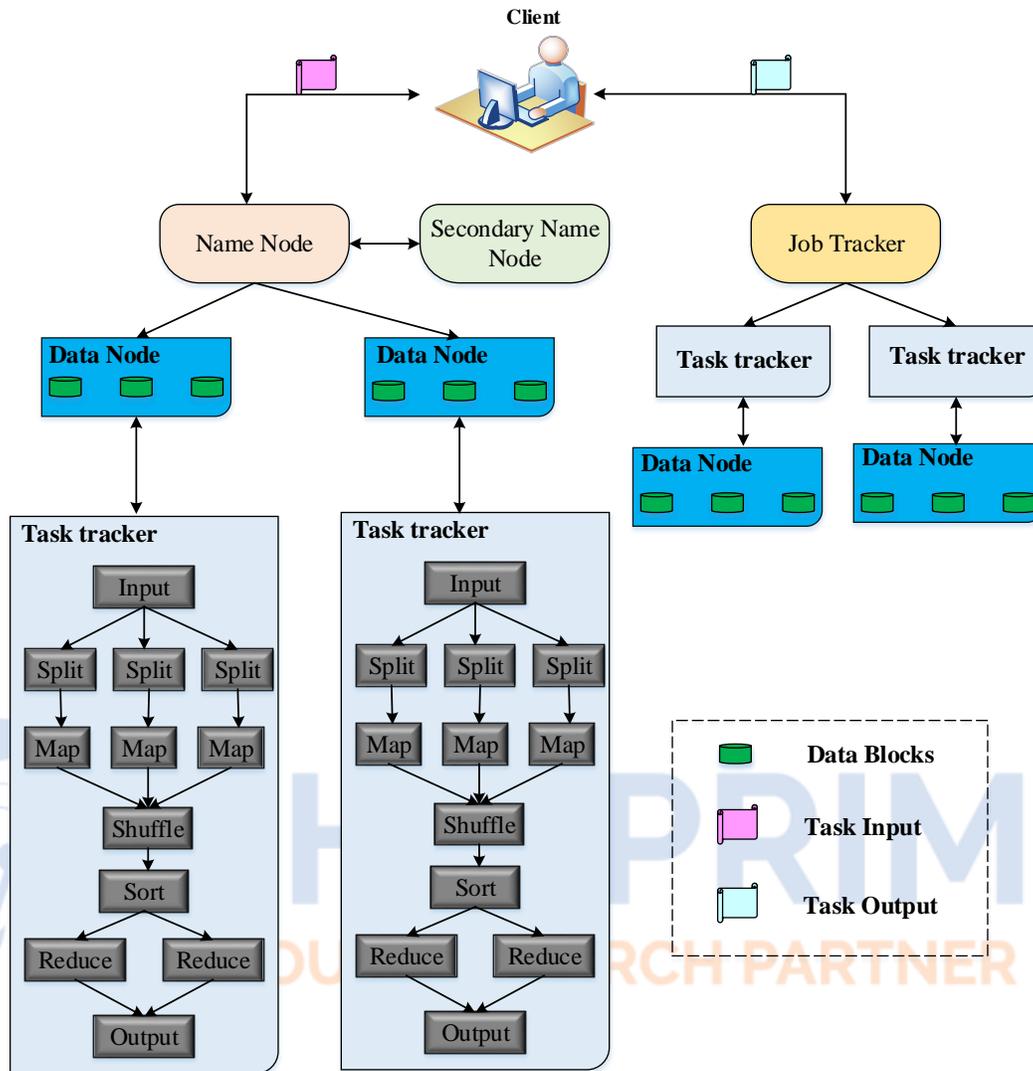


Figure 1.22 HDFS architecture

This signifies the namenode server doesn't increase as a tasktracker or data node in the Hadoop system. There is inappropriately an undesirable feature to the significance of the namenode its specific point of damage of the Hadoop cluster. If the host nodes fault for the software or hardware reasons, then the Hadoop cluster will endure the operations effortlessly or else it can be rapidly restartable.

- **DataNode**

The datanode in the Hadoop cluster is present in each slave machine to execute the process of the distributed file system analysis and inscription HDFS

blocks to actual files on the local file system. The data nodes in the Hadoop system can converse straightly with another data nodes exist in the system. This conversation is used to reduce the redundancy of the data blocks present in each data nodes. The datanodes are frequently comments the namenode regarding its process. Starts with an initialization, each datanodes notify the namenode regarding the blocks it's presently processing. After completion of the mapping process, the datanodes persistently inquire the namenode to afford statistics regarding local deviations. In addition to receive the commands to generate, transfer or remove blocks from the confined disk.

- **Secondary NameNode**

The secondary namenode is an associate daemon to investigate the state of the HDFS cluster. As similar to the namenode, each cluster comprises single secondary namenode and it generally exists on its private machine as well. No other datanode or tasktracker daemons execute on the unique server. The secondary namenode fluctuates from the namenode this process doesn't obtain or store any existent interval modifications to HDFS framework. As an alternative, it interconnects with the namenode to proceeds the portraits of the HDFS metadata at specific intermissions demarcated by the configuration of the clusters. The namenode is the single point of failure for a Hadoop cluster and the secondary name node portraits is used to reduce the downtime and data loss of the HDFS system. However, the namenode disappointment demands the human intervention to restructure the cluster to utilize the secondary name node as the crucial name node. The work of the secondary name node is not designating the subordinate to the namenode. Instead of it, secondary name node occasionally read the archive scheme modifications log and also affords the reserve for the anterior; hence it modernizes the files effectually. In huge number of cluster atmosphere, the secondary namenode typically execute on the various machine than the main namenode as its memorial necessities are on the unique order

as the main namenode. This task permits the namenode to begin the upcoming moments fastly.

- **Job Tracker**

The job tracker is the interface between the application and Hadoop framework. It identifies the implementation plan via defining which archives to be process next, allocates nodes to various tasks and investigates entire tasks as they are processing. If any failure occurs, then the job trackers will spontaneously unveilings the task, probably on a various node upto a predefined bound of revises.

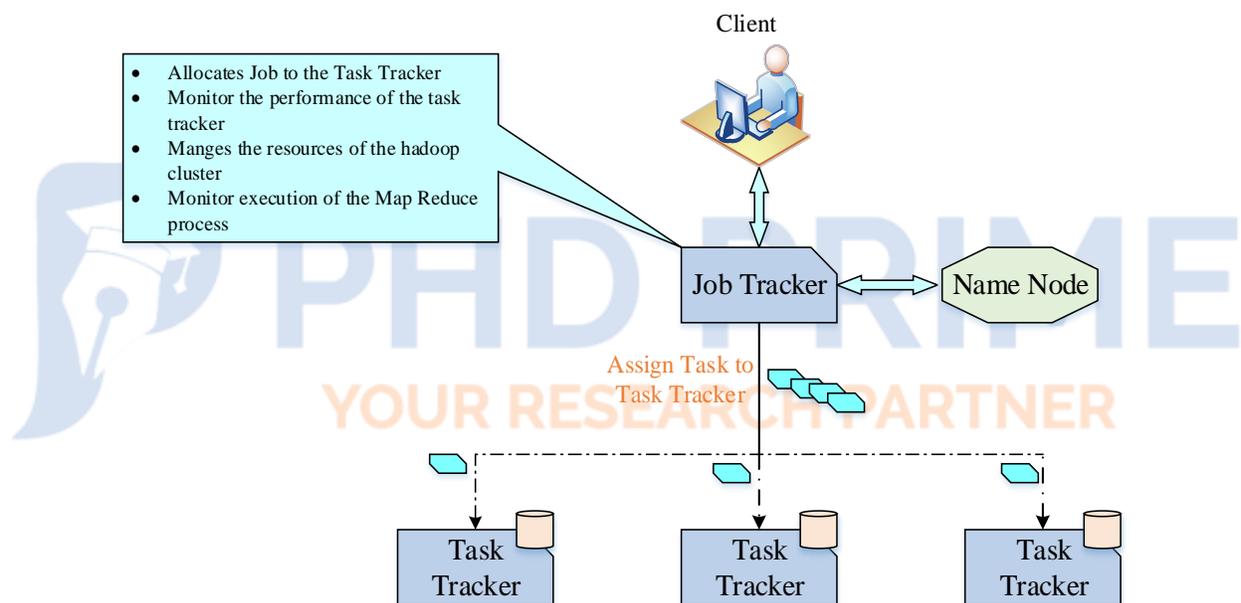


Figure 1.23 Job tracker working

Figure 1.23 depicts the working of the job tracker present in the Hadoop architecture. There is simply single job tracker in each Hadoop cluster which commonly execute on the server as the master node of the Hadoop cluster. The job tracker is the master supervising node which investigates the general implementation of the Map Reduce job. The job tracker is the amenity inside the Hadoop which performs the MapReduce tasks to unique nodes exist in the cluster, presently the nodes which have the data or else nonetheless in the identical rack.

The client provides the jobs to the job tracker which communicate with the namenode to estimate the position of the data. The job tracker provides the job to the selected task tracker nodes. It keeps on investigating the performance of the task tracker. If the task tracker nodes do not provide the heartbeat signal frequently, then the particular task tracker is considered to be a failure node and the provided job is reserved on a dissimilar task tracker. The job tracker keep informed it prominence whenever the job is completed.

- **Task Tracker**

The task tracker monitors the implementation of each task on each slave node. Each task tracker can offspring several virtual machine to manage the numerous map or reduce tasks in analogous manner. Figure 1.24 elucidates the task tracker working process in the Hadoop environment. One fundamental obligation of the task tracker is to continuously converse with the job tracker node. Since if the job tracker doesn't obtain the heart signal from the task tracker node within the specific amount of time, then it consider that specific task tracker node as the failure node. Then, the job tracker node assigns the jobs to another task tracker node. Here, the task tracker breaks the allocated task into the map and reduce tasks where client given process is executed.

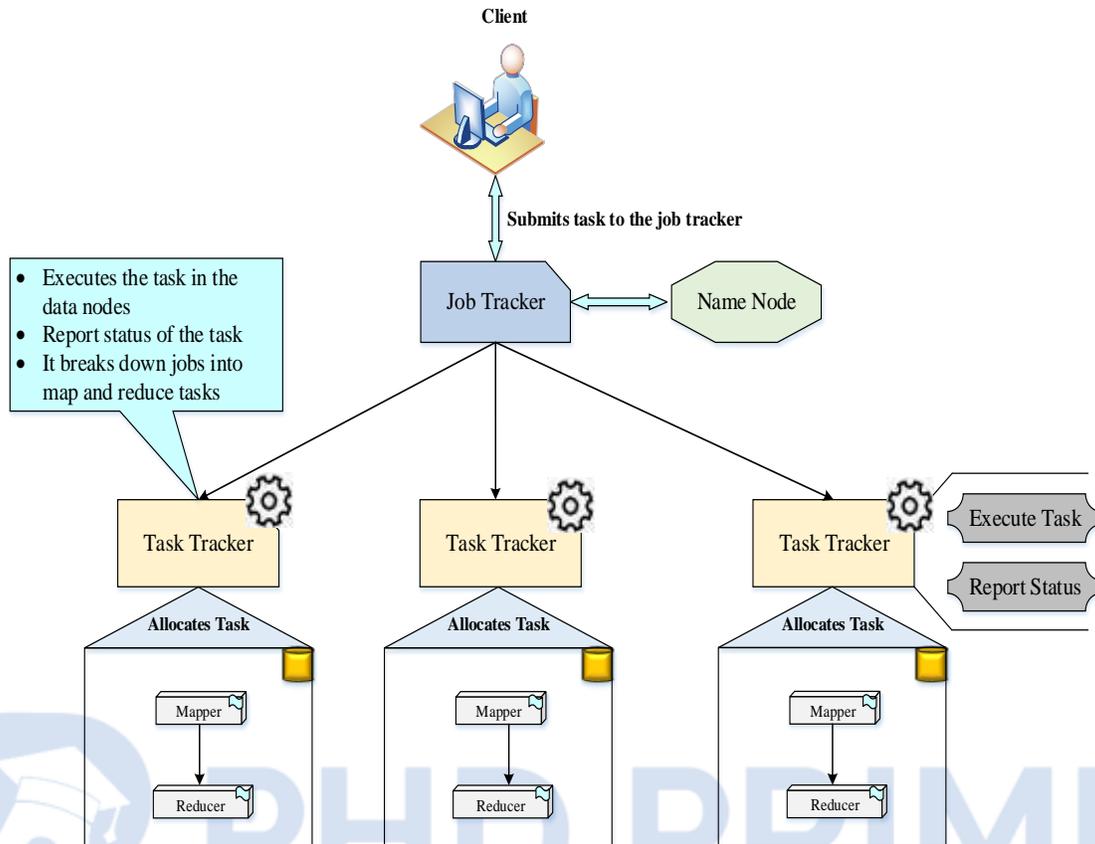


Figure 1.24 Task tracker working

Every task record has the slots regarding the implementation map which progressively decreases and provides the evolution to the job tracker regarding the execution of the allocated task. Every data node managed by the task tracker transmits the heartbeat message to the namenode occasionally. The time difference between the two successive heartbeat intervals is assigned as three seconds.

Three Dimensional (3D) Image Processing

With the recent development of internet and advancements on computer technology, that has been facilitated for propagation of 3D models on the web [9]. The massive impact of 3D models in day-to-day life was already observed in application domains spanning from edu-entertainment to scientific visualization. These 3D models have become widespread in variety of applications including mechanical CAD, computer

graphics, molecular biology, medicine, etc. Due to the success of introducing 3D visual technologies, the demand for wide variety of 3D content such as 3D images, 3D videos and 3D games are increasing significantly. To fulfill the user demands, new 3D video content is created and also existing 2D videos are converted into 3D format by using 3D conversion tool. The main uses of 3D models are as follows,

- Creation of basic objects in games environment
- Computer animated film industry for creating animated characters and objects with advanced technologies
- Medical industry uses 3D models of human objects
- 3D visualize architectural objects in architects fields
- 3D models provide presentation of vehicles, new devices, etc. in engineering fields

In computer systems, 3D describes an image, which provides the perception of depth analysis of images. 3D model displays a picture of objects in the form that physically present with that labeled structure. 3D allows pictures flat picture appearance to human eye that is considered with three dimensions which include depth, weight and height. The result of 3D image process is analyzed and processed to extract information, that supports wide range of applications, which include shape search on the web, video retrieval, object recognition, face recognition for surveillance and security, clinical processing in medicine, mapping of earth surface, etc.

Numerous improvement in computer graphics, multimedia, computer vision and other fields have enabled to develop the advanced types of visual media named 3D video ^[10]. ^[11]. It extends the traditional video technologies by capturing three-dimensional appearance and dynamics of real-world scenes. It provides depth analysis and visual effects about the observed three-dimensional scene structure. The depth-based representations obviously transform the videos are displayed with different types such as volumetric display, stereoscopic display, auto-stereoscopic display, and holographic

display. 3D video captures two essential information about real-world scenes which dynamically changes over time: Shaders and texture. Shaders is commonly used to produce the lighting shades in 3D models, which allows adjustments of opacity, light, reflectivity, shading, etc. Textures provide the realistic representation for 3D objects, which specify high level on details about an object. The digital representation of 3D objects are categorized into three levels of granularity: geometric, Structural and semantic levels.

Figure 1.25 represents the 3D digital object and its relevant characteristics such as name and URL (Simple resources), set of triangles and normal (geometric representation), skeleton of teapot (Structure of shape) and visual representations of shapes (Semantic). The levels of a 3D shapes are categorized as follows,

- Geometric level
- Structure level
- Semantic level

In geometric level, 3D shape is determined by determining its shape using geometric representation methods such as triangle mesh (collection of vertices, edges and faces), a Non Uniform Rational B-Splines (NURBS- represents curves and surfaces of 3D shape). The geometric representation provides the spatial characteristics of 3D object which is capable to interact with human visualization for effective support on different analysis process. Geometric level is effectively used to provide the physical properties of 3D shapes.

Structure view produces the abstraction level, which is used to identify the segment or portion of 3D shape in tubular parts. For structural view, morphological or geometrical analysis representations are required to detect the relevant features of 3D shapes. This structural level describes the same shape into different manner based on the characteristics of object

Semantic level provides specific knowledge to objects, which delivers the semantic labels (visual words) to specific shape of objects. This level bridges the gap between structural and geometric levels by identifying meaningful representation. Semantic models explicitly tag the “semantic labels”, which automatically examine the properties of 3D shapes. This semantic label is effectively used for analyzing the process of 3D shapes in different areas as image retrieval, video retrieval, image storage, image analysis, etc.

1.5 THESIS MOTIVATION

We formulated the key problems of the thesis in following: there is an amazing growth in the amount of digital video data in recent years. There exists a gap between low level features and high level semantic content. In CBIR, the problem is how to find the similarity between the query image and all the images in database. However comparing the image by patch by patch is not feasible because the object/scene in the image may lead to any change. Typically, an optimum set of visual features are extracted in image and changed into a fixed size vectors for representation of the image. Further, the significant characteristic for the image, semantic gap is considered to improve the performance. Consider the visual and semantic contradictions between the query image and large size database is the primary requirement for giving effective query response for users.

- Necessity of Video Database Management System
 - Increase in the amount of video data captured
 - Efficient way to handle multimedia data

As mentioned earlier, we mathematically define the problem to compare the similarity between two images (α, β) as follows:

$$IR(\alpha, \beta) = \sum_{a \in \alpha} \sum_{b \in \beta} k(\alpha, \beta) \quad (1.9)$$

$$= \sum_{a \in \alpha} \sum_{b \in \beta} \phi(\alpha)^T \phi(\beta) \quad (1.10)$$

$$= \psi(\alpha)^T \psi(\beta) \quad (1.11)$$

With the function of Equations (1.9) – (1.11), we generate certain research questions for feature extraction, feature fusion or reduction and indexing images in large database:

- **RQ1 (feature extraction)**

How to define the query image content α by set of visual features $\{a_1, a_2 \dots\}$?

- **RQ2 (feature fusion/reduction)**

How to change the feature sets $\alpha = \{a_1, a_2 \dots\}$ with various sizes to a fixed-length vector $\psi(\alpha)$?

The major reasons for content based video retrieval are discussed further in this section. Video retrieval is recently used by enormous users all over the world. Sharing of texts has become older and videos / images sharing are popular at present is possible to present our ideas realistically which could not be done on texts. Video retrieval concept involved in several previous image processing algorithms and techniques for achieving,

- Accurate results with relevant data
- Minimum processing time
- Maintenance of image / video quality
- Capability to manage large number of user queries
- Reduce space complexity

The motivation behind our proposed research work is to reduce space complexity and also to provide original quality of videos from storage. Quality of video was a major issue which should be solved to satisfy users with their queries. Each minute thousands and thousands of users upload their videos, which requires higher storage capacity. If storage availability is less then the involvement of number of users will be minimized. Hereby our motive is to use an effective storage (i.e.) Hadoop and maintain the video quality by

using best and novel image processing algorithms. We evaluate our proposed work in three different scenarios according to the hardware specifications and the performance metrics measured are listed below,

- Precision and recall
- Overall map reduce time
- Percentage Accuracy
- Positive results
- Processing time
- Key frame detection
- Filtering accuracy

1.6 THESIS OBJECTIVES

Our main objective is to improve accuracy in results by introducing novel algorithms in image processing and to support huge data with suitable storage environment. We list up all our objectives of proposed research work,

- A broad overview on Content Based Image Retrieval and Content Based Image Retrieval is to be discussed with the limitations and advantages of this concept. On having a detailed study over these concepts, is the major reason to define ideas on this area.
- Usually CBIR and CRVR have concentrated only on 2D, our main focus is to process with 3D which is trending at present and 3D is likely attracted by people at all ages. We aim to move up this concept to next level by using 3D images and videos in this work.
- A design in 3D for content based video retrieval is proposed that supporting large number of users with accurate result achievement.
- To resolve the storage issue, novel idea of Hadoop map reduce framework is used which was not focused in other previous works of video retrieval.

- Key frames are selected using Conditional Entropy based Fast Key Frame selection process which is effectively supported to select key frames that are more informative, this process involves mathematical computation of entropy values. 3D frames are involved with the removal of noise that is called denoising performed by the combination of Bilateral Adaptive Median Filtering.
- Newfangled feature extraction is performed by combining topology and geometry when extracting shape feature, since topology and geometry plays a major role in each frame of the video. 3D Hybrid SIFT and SURF features are extracted to mitigate the problem of loss of geometry.
- Threshold Based-Predictive Clustering tree is built for generating visual vocabulary and for matching accuracy we have combined soft weighting scheme with L_2 distance.
- The problem of noise in constructed visual words is minimized by novel numerical semantic analysis, this is enabled to produce accurate retrieval results for the given user query.
- Features like shape, color, texture and motion are extracted and considered for similarity matching scheme which is named as Multi-Featured Matching Scheme for accurate similarity matching based on features. Deep learning based algorithms based
- Our novel algorithms show better performance results by improvising retrieval accuracy, positive results and precision-recall.

All the above listed objectives are elaborated in following sub-sections and further chapters takes CBVR concept to next level with 3D based processing.

1.7 THESIS RESEARCH METHODOLOGY

Video retrieval process in 3D is not discussed by any researchers due to the limitations on storage and other image processing limitations existed in 3D. Our newfangled work is concentrated on faster retrieval of videos with higher positive results

for all the user's queries. The significant methodologies involved in this research are discussed in this section.

For faster retrieval process we have used Map Reduce in Hadoop environment which deals storage problem as well as retrieves faster results. Moving into the core idea, our work includes key frame extraction, feature extraction, codebook generation and similarity matching.

- Extraction of key frames from total number of frames obtained from a video file. Key frames are selected with the execution of temporal maximum occurrence frame (TMOF) in Hadoop Distributed File System (HDFS). Based on the pixels, histogram is constructed for group of frames (GoF). Histograms for 3D frames are constructed using width, height and depth. Optimal frames are determined and chosen as key frames. Conditional Entropy based Fast Key Frame selection algorithm selects optimal frames from shots which are converted from a complete video. This algorithm is enabled to eliminate redundant frames from the estimated relative entropy values. Hence key frame extraction process is significant in video retrieval concepts for minimizing cost and resource utilization.
- For denoising 3D frames and incoming 3D query images, Bilateral Adaptive Median Filtering is proposed for removing noises. The initial step of this process is defining 3D frame or image in Taylor series. Each and every sub bands are taken in account and the presence of noise is eliminated. After elimination of noise a new frame or image is reconstructed. This approach also leads to increase in accurate identification of 3D videos for the given 3D image query
- Feature extraction is the first step in the generation of Bag of Visual Words, which is comprised with a set of visual words. Feature extraction involves with the consideration of three significant features as shape, color and texture. Extraction of shape feature is comprised with four sequential steps. They are Medial surface extraction and segmentation, Re-adjustment of segmentation, super ellipsoid approximation and 3D distance field descriptor (3D DFD). In this shape feature

both topology and geometry are combined for better efficiency. Color feature is processed by estimating Gray-level Co-occurrence matrix and the texture features include Angular Second Momentum, Contrast, Entropy, Correlation, Mean and Variance. In our second work we have additionally include motion feature, since a video is analyzed and matched so motion feature is also considered along with other three features.

- For the construction of visual vocabulary, we use Threshold Based-Predictive Clustering Tree which requires asset of local descriptors. Predictive Clustering Tree is adapted by means of top-down induction of decision tree (TDIDT) algorithm. The constructed visual vocabulary induces noise which is resolved by using numerical semantic analysis. Construction of large number of visual words is managed by Map and Reduce functions.
- Soft-weighting scheme is introduced for matching similarity between two images to efficiently retrieve relevant videos to the corresponding query of user. In this matching scheme, top-k nearest visual words is selected and then weights are assigned for visual words. Nearest visual word's similarity is estimated, and then L_2 distance function based similarity is determined. With this result k-nearest neighboring visual vector images are listed and their corresponding videos are accurately listed.
- Multi-features light weight similarity matching scheme is proposed for accurate matching of image with the videos. Based on the features the similarity is estimated in MapReduce framework. Due to this similarity matching, the performance results with accurate 3D video retrieval.

1.8 ORGANIZATION OF THESIS

Overall structure of thesis is segregated into five chapters. Each chapter is devoted to specific topic. The chapters involved in this thesis are as follows;

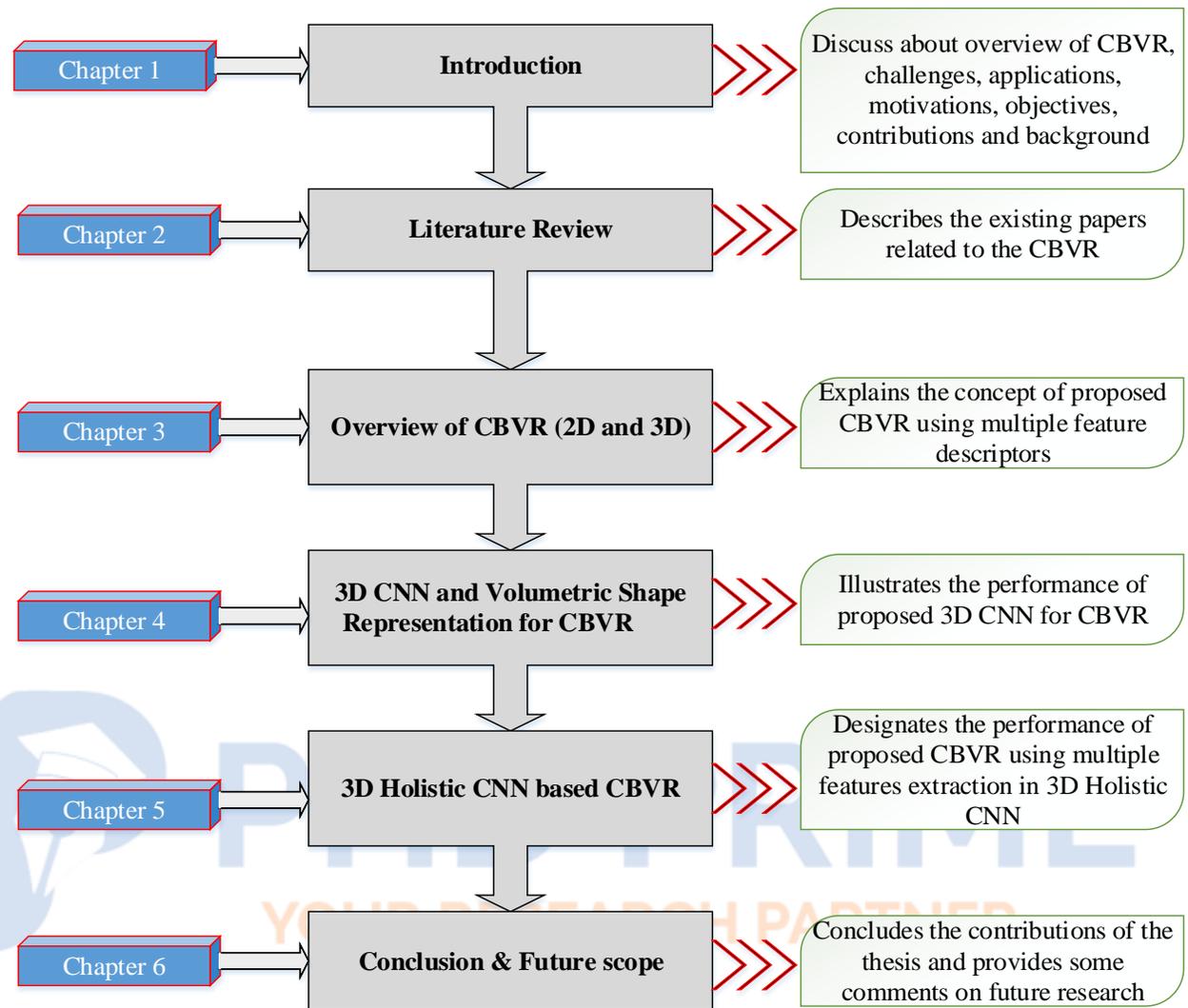
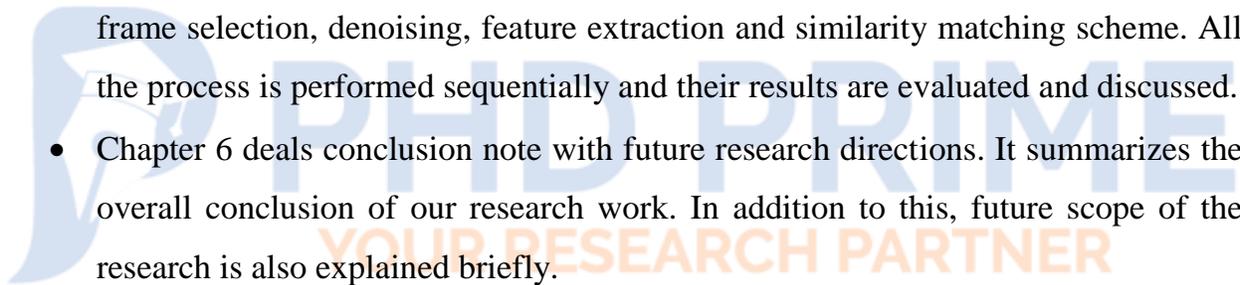


Figure 1.25 Thesis Organization

- Chapter 1 illustrated with introductory note of the overall work. This emphasized on the methodology used and its purpose. Section 1.1 explains the motivation of the thesis. Section 1.2 deals about the Objective of this proposed work. Then Section 1.3 deals with the novel Methodologies used in our research work. Section 1.4 gives the background theory involved for this entire thesis.
- Chapter 2 explains literature survey of our proposed work. In this chapter we have focused nearly 45 papers and explained about the concepts. It focused on main papers and analyzed to define the proposed objective.

- Chapter 3 discusses overview and state-of-the art on image and video retrieval techniques. In this chapter, we have analyzed several image to image retrieval techniques and image to video retrieval techniques. This complete analysis supports to identify the problems, challenges and issues existed previously.
- Chapter 4 focus on 3D CBVR in CNN. This mechanism provides key frame extraction, BOVW construction, feature extraction, visual codebook generation and matching. Each process is employed in this research framework. SIFT and SURF feature descriptors are combined in 3D CNN for feature extraction. It extracts the features in a parallel way. Finally 3D CBVR in Hadoop Map Reduce is evaluated with significant performance metrics.
- Chapter 5 details the lightweight 3D video retrieval over Hadoop using Holistic Feature Extraction and MapReduce model. This chapter includes process of key frame selection, denoising, feature extraction and similarity matching scheme. All the process is performed sequentially and their results are evaluated and discussed.
- Chapter 6 deals conclusion note with future research directions. It summarizes the overall conclusion of our research work. In addition to this, future scope of the research is also explained briefly.



CHAPTER 2



LITERATURE SURVEY

PHD PRIME

YOUR RESEARCH PARTNER

In the past years, several approaches have been proposed for efficient video retrieval and image retrieval, but it provides less effective results [12]. Several multimedia applications for 3D shape feature extraction, classification and matching processes are discussed in [13]. The following sections describe the recent related works on 3D video retrieving process with respect to key frame selection, bag of visual words, feature extraction and similarity matching.

2.1 CONTENT BASED IMAGE RETRIEVAL

Image retrieval is the process of localizing (searching and retrieving) images from large number of database, which improves the requirements of specified images [46]. It is one of the active research field and researchers have proposed several efficient algorithms and techniques. Image retrieval was classified into two techniques such as (i) Text based retrieval and (ii) Content based retrieval. Text based retrieval requires textual (keyword) annotations for searching and retrieving images from the database. Content based image retrieval is also known as visual image retrieval, which represents the unique description like color, shape and texture of the images. Image-to-image retrieval process is demonstrated in figure 2.1.

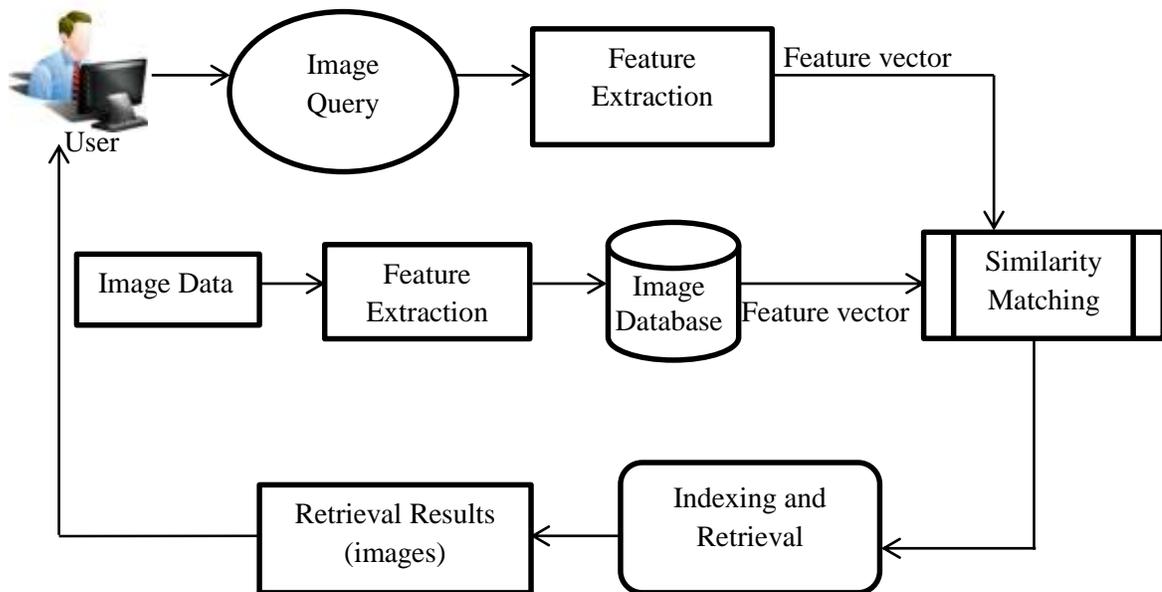


Figure 2.1 Image-to-Image Retrieval Process

A recent technology allows indexing, storing, transmitting, retrieving and manipulating of huge collections of images in databases. The required images are retrieved with respect to similarity of features. The features of query images are compared with features from image database for extracting the most similar images. Huge number of image retrieval algorithms/ techniques are developed to improve the retrieval process such as Cross model multimedia retrieval, content based image retrieval, collaborative image retrieval, etc. However, these techniques have some additional benefits and drawbacks for retrieving images from large databases. In this sub-section, we discuss about previous research works with its algorithms, benefits and drawbacks.

Raftopoulos et al ^[47] introduced a Markovian semantic indexing (MSI) approach for online image retrieval. MSI was adaptable for the Annotation Based Image Retrieval (ABIR) system, which provides the most efficient content for both text-based queries and image caption. The main scope of this process was improving the performance of retrieval over the online image processing systems. Stochastic approach was utilized because the user has possibility to unselect the relevant images, for this aggregate Markovian chain was interpreted. During training phase, the images were not annotated. After the submission of user query, the annotation was generated for relevance image and also user defined queries are in the combination of keywords. In testing phase, annotation data was collected from training phase and retrieve the results based upon keyword relevance probability weights in order to offer with more relevant images. This proposed MSI approach was compared with Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI), from that result MSI achieves great result. The main advantages of this process were to provide more relevant images and also perform automatic annotation and indexing process in ABIR system. However, this process consumed high computation time and also it does not fully satisfy the user requirements.

Kenneth et al ^[48] defined a new approach for providing personalized query suggestion to the user in order to enhance the semantic query searching. Most of the user's queries are too short or it was ambiguous. In order to tackle this problem, personalized concept

based clustering strategy was proposed. This approach combined graph relationship and concept based clustering. For clustering, four significant steps were processed: (i) to extract the query information and its relationship with the support of web-snippets, (ii) select the user preferences, (iii) perform agglomerative clustering algorithm for finding queries, which was conceptually similar to one another, (iv) to suggest the most relevant queries to users. The main benefits of this personalized concept based clustering was increasing the prediction accuracy at the same time reduces the computational cost. However, this process was not concentrated on indexing and ranking process which leads to redundancy problem.

Shaoting et al ^[49] proposed unsupervised graph-based fusion of image retrieval sets which improves the retrieval quality; it provides image representations and reduces the overlapping top results. In this process, two different retrieval algorithms were involved: (i) local features indexed by vocabulary tree and (ii) holistic features indexed by compact hashing codes. The retrieval quality of candidate images were measured online by consistency of neighborhoods of top candidate images, which was specific to individual queries. Then, the retrieved results for queries were considered into graphs to provide the sorted list of candidates images based on similarity scores. The graph fusion method was used to improve the efficiency of retrieval by considering ranked (page re ranking) lists of similarity scores. Hence, the multiple times of re-ranking is not necessary to improve the quality of retrieval process.

Steven et al ^[50] proposed a novel semi-supervised learning framework for image retrieval based on labeled and unlabeled data. This proposed process was comprised with semi-supervised active learning and support vector machines. Initially, the number of input labeled images was entered into semi-supervised active learner in which SVM was performed to produce a rough decision boundary based on labeled data instances. Then, gaussian fields and harmonic functions based semi-supervised learning techniques were applied to smooth the labels for improving the classification performance. In semi-supervised learner, the unlabeled data was separated from the labeled data, which was

minimizing the risk during top-k image retrieval. This retrieval process gives the balanced classification performances; at the same time it provides possibility to retrieve repeated or unwanted data.

Premkumar and Prassenna et al ^[51] proposed a new methodology named decision Markov Semantic Indexing (MSI) for online image retrieval system. It performs automatic semantic annotation and indexing process to improve the retrieval information. Firstly, the user's query forwarded to system, which constructed AN mixture Markov chain (AMC) model for detecting the keywords from image. Based on keywords, the automatic annotation process was performed by hidden markov model which represents the semantic relationship between keywords. Then, the low-level feature vectors such as color, shape and texture was extracted using stochastically generated hidden markov model. Further, MSI approach was measured with similarity distance, which examined a statistical model with hybrid text/visual characteristics based aspects. After completion of similarity measurement, multimodal hyper graph learning based sparse coding method was performed for image retrieval. Then, image re-ranking method was suggested to enhance the retrieval results through collection of weights. However, this process consumes high computation time as well as it does not able to satisfy user requirements.

Murthy et al ^[52] developed a content based image retrieval using hierarchal and K-Means clustering technique. In this paper, image was given as input query and retrieves images based on image content. Content based image retrieval was an approach for retrieving semantically-relevant images from an image database which automatically derives image features. This proposed technique consist of two clustering algorithms (hierarchal and K-Means), which was used to group the images into clusters based on color content of images. Because, the color content of image was considered to be unique for each users. Initially, the input image performs hierarchal clustering, which takes in account of every point on the image into own cluster and then find most similar pairs of clusters. Then, K-Means clustering was performed to determine the similarity based distance for clustering. This two clustering process provided better performance than

individual clustering algorithms. Hierarchical clustering improved the searching time whereas K-Means clustering improved the clustering quality and accuracy of retrieval. However, this clustering process is complex and it was difficult to determine the number of clusters due to improper feature extraction.

Valiollahzadeh et al ^[53] introduced a novel algorithm for face recognition system; this algorithm was combination of Adaboost and SVM classifier. In first stage, the given input color was converted into gray scale image. Further, feature extraction process was processed in offline phase using 2D haar algorithm, which classifies the various statistical features based on the initialization of feature weight. Based on the features of images, classification was performed using combination of Adaboost and SVM classifier. Finally, the face images were identified after preprocessing stage is completed. The main advantage of this paper was providence of accurate classification result, which was higher than other classifiers such as neural networks and decision trees. This paper has some advantages hence the face recognition process has concentrated on classification and not for feature extraction. This process also reduces the region of interest when selecting specific color.

Kannan et al ^[54] developed content based image retrieval for coarse content image classification which was used to reduce the searching time of images. The coarse content of images was collected into three categories such as high-texture detailed image, average-texture detailed image and low-texture detailed image. Initially, preprocessing was performed to enhance image quality and suppress the distortions of images. Then, mean values of RGB images were estimated with respect to separation of R, G and B values. Further, texture value of both query image and database images were classified by entropy classification, which provides the energy content of image. Histogram features were estimated to extract the features from input image, which increases accuracy of retrieved result. After that Fuzzy C-Means (FCM) clustering was performed, this was used to cluster the similar groups based on relevant features. Then, threshold value was added with entropy classification, which was compared with concerned cluster and target

images for retrieving relevant images. The main advantage of this process was improving the retrieval process since it proceeds with effective preprocessing and feature extraction. However, it consumes larger amount of time for preprocessing and extraction process.

Table 2.1 Comparison of Various Techniques used in Image-Image Retrieval Process

Techniques	Algorithms/ Methods	Benefits	Drawbacks
Cross Model Multimedia Retrieval	<ul style="list-style-type: none"> • Correlation and annotation hypothesis 	<ul style="list-style-type: none"> • It provides efficient retrieval result than unimodal • It solves semantic analysis problems 	<ul style="list-style-type: none"> • Accuracy is low when handle the large number of datasets.
Search based face Annotation	<ul style="list-style-type: none"> • Unsupervised face alignment • GIST feature extraction • Locality Sensitive Hashing (LSH) algorithm 	<ul style="list-style-type: none"> • Scalability is increased due to clustering process • Improve the searching efficiency. • Reduces the computation time 	<ul style="list-style-type: none"> • It is not efficient for real time applications.
Robust Transfer Video Indexing (RTVI)	<ul style="list-style-type: none"> • Multiple kernel Learning (MKL) • Maximum Mean Discrepancy(MMD) 	<ul style="list-style-type: none"> • Reduce the noise content in image • Enhance the result due to metadata classification 	<ul style="list-style-type: none"> • It does not satisfy the user requirements • It is difficult to determine the metadata since it was not handle the

			semantic analysis
Automatic Figure-Ground Segmentation	<ul style="list-style-type: none"> • K-nearest algorithm • SVM classifier • Markov Random Field 	<ul style="list-style-type: none"> • Automatically segment the images and increase the segment accuracy. 	<ul style="list-style-type: none"> • It is only suitable for small scale dataset • High Computation time for large database
CBIR with Clustering	<ul style="list-style-type: none"> • Naïve Random Scan(NRS) • Local Neighboring Moment (LNM) • Neighboring Divide-Conquer (NDC) • Global Divide-Conquer (GDC) 	<ul style="list-style-type: none"> • It reduces search space and number of iterations by voronoi diagram • It improves the retrieval effectiveness and efficiency • It is easy to implement and provides efficient result even large number of datasets 	<ul style="list-style-type: none"> • Difficult to identify the correct number of clusters • Dynamically updation process is not effectively performed in clusters
Binary Sift Codes	<ul style="list-style-type: none"> • Cross indexing strategy • Flexible Binarization strategy • SIFT descriptor 	<ul style="list-style-type: none"> • Cross indexing improves the retrieval efficiency • Binary code reduces storage memory and computation time 	<ul style="list-style-type: none"> • Computationally expensive • SIFT feature into binary code is complex since the code length is large
Spatially-Constrained	<ul style="list-style-type: none"> • Gaussian smoothing • K-nearest re- 	<ul style="list-style-type: none"> • It effectively removes the distraction of images 	<ul style="list-style-type: none"> • High computation cost • High computation

Similarity Measure (SCSM)	ranking		time for gaussian smoothing
CBIR using Image Retrieval	<ul style="list-style-type: none"> • Supervised active learning algorithm • SVM classifier 	<ul style="list-style-type: none"> • It gives balanced classification performance 	<ul style="list-style-type: none"> • It is possible to retrieve the unwanted data
Query Specific Rank Fusion	<ul style="list-style-type: none"> • Vocabulary features • Holistic features 	<ul style="list-style-type: none"> • It preserves efficiency and scalability for both vocabulary and holistic features 	<ul style="list-style-type: none"> • It is not practically suitable • High computation time • Accuracy is degraded by presence of noisy or irrelevant features
Cascade Category-Aware Visual Search	<ul style="list-style-type: none"> • K-Means Clustering • BOVW generation 	<ul style="list-style-type: none"> • High accuracy due to visual descriptor, category label and contextual clues 	<ul style="list-style-type: none"> • High repeatability that leads to appear many noisy features in background
Collaborative Image Retrieval	<ul style="list-style-type: none"> • Laplacian Regularized Metric Learning • Graph Regularization Method 	<ul style="list-style-type: none"> • It reduces the semantic gap issues by considering relevance feedback information 	<ul style="list-style-type: none"> • Storage maintenance problem while storing huge amount of data • Lower accuracy than other methods
Linear Distance Metric	<ul style="list-style-type: none"> • Distance metric Learning 	<ul style="list-style-type: none"> • Improve the accuracy of result by 	<ul style="list-style-type: none"> • Time Complexity

Learning (LDML)	<ul style="list-style-type: none"> • Classification • Clustering 	considering multiple features	
Geometric Optimum Experimental design	<ul style="list-style-type: none"> • GOED • Reproducing Kernel Hilbert Space (RKHS) 	<ul style="list-style-type: none"> • Efficient result when compared with SVM active approach • Overcome the diverse of potential issues • Label independent approach 	<ul style="list-style-type: none"> • It supports only small number of databases • It does not provide spatial information about images • GEOD analysis is difficult and expensive
Wavelet based CBIR	<ul style="list-style-type: none"> • Wavelet Filter coefficient • Sub sampling 	<ul style="list-style-type: none"> • Effectively retrieve the image by decomposing process • Improves the spatio-temporal information of images 	<ul style="list-style-type: none"> • Longer compression time • Difficult to relate coefficients to position in input image
Parallel Processing using Map Reduce	<ul style="list-style-type: none"> • Image Indexing • Distributed Processing • Hadoop Map Reduce Processing 	<ul style="list-style-type: none"> • This process increase the performance of data insertion and query processing • It reduces the computation time due to parallel processing 	<ul style="list-style-type: none"> • It is unusable for real time applications • It requires lot of time to perform the task thereby increasing latency

3.2 CONTENT BASED VIDEO SEMANTIC SEARCH

Video retrieval is most interesting research topic in both multimedia and real life applications. There are vast amount of video archives including broadcast news, movies, meeting videos, documentary videos, etc. For video retrieval, information considered they are visual content, audio information and text information. Video retrieval techniques have been classified into two different types such as (1) Text-based video retrieval and (2) Content-based video retrieval. In first type, the name itself suggests with the method of retrieving videos with the help of text information ^{[55],[56]} present in the videos. Then, in second type the videos are retrieved based on the contents of the query given by the user. Video retrieval contains content analysis, content modeling, feature extraction, indexing and querying.

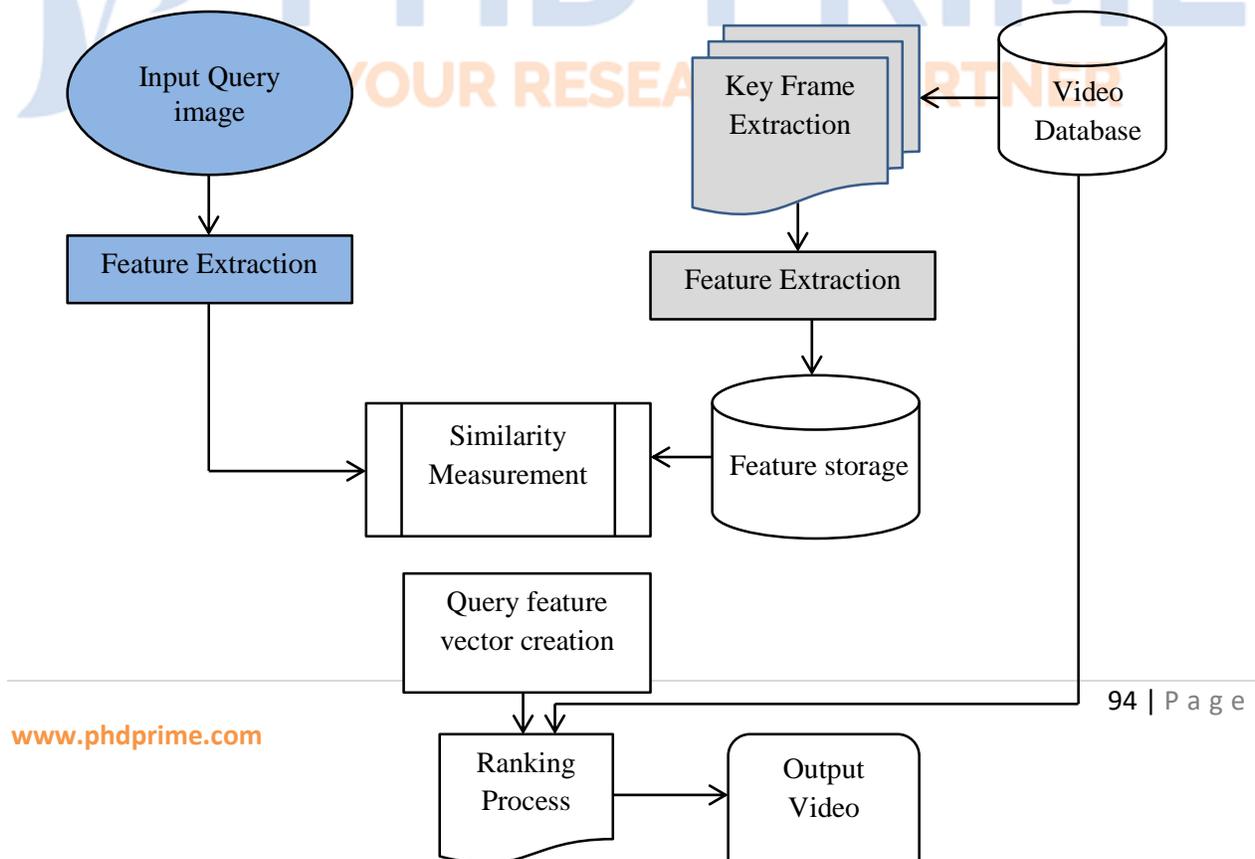


Figure 2.2 Image to Video Retrieval Process

In this section, we discuss with a comprehensive literature survey on video retrieval which enables image and video retrieval by content. Figure 2.2 represents the image to video retrieval process; the user query providing query is the initial step. Further, feature extraction is processed with respect to different extraction techniques. In video database, the frame extraction is performed and key frame are selected for further processing and provides efficient result. Based on the key frames, feature extraction is processed and then similarity measurement is performed with input feature extraction. Finally, ranking process is applied to each similar result and then detects the top most relevant result videos to user. One of the important issues in the multimedia video retrieval process is that efficient similarity measurement between user's submitted query and video stored on the database by which retrieve the relevant data from it. A number of video retrieval techniques such as indexing clustering, and feature extraction are reviewed in this section. It evaluates the various image to video retrieval systems with respect to different conventional techniques.

Fan et al ^[57] proposed an efficient classification and automatic annotation of large scale video. In this paper, ontology concept was integrated to boost hierarchal video classifier and multimodal feature selection. The main objective of this paper was tackling the semantic problems by using automatic video detection with semantic classification.

Initially, video segmentation was performed using multimodal boosting algorithm, which reduces the storage size in feature extraction process. Further, ontology was constructed to define the vocabulary of domain-dependent videos and contextual relationships. Then, hierarchal boosting algorithm was used to perform similarity matching between user input and video database, this algorithm provides retrieval result based on semantic results. This process provides more institutive solution for query specification and evaluation. The main advantages of this paper were reduction of computation complexity and storage space. However, this process has drawbacks; since it was not able to satisfy user requirements due to the use of ontology concept.

Susu Shan et al ^[58] exploited a textual saliency detection method with random walk model to spatiotemporal textual saliency by visual features, local video region graphs, spatio and temporal maps. In this paper, two main components were involved such as (i) Efficient feature extraction model by combining random forest and CNN, (ii) random walk with restart model for both spatial and temporal saliency information. Initially, video frames were extracted by using top-down feature extraction scheme, which generates confidential scores for each sliding window on video frames. The confidential score was considered as initial and second scores, initial score was computed from Histogram of Oriented Gradients (HOG) and LBP features. Further, CNN was used to obtain second confidential score. Then, final confidential score was calculated based on the result of initial and second scores. After that spatiotemporal salient regions were detected in video sequence based on random walk with restart graph model, which effectively performs over confidence and visual features. The main benefits of this process was enhancing the speed of confidence computation and achieving comparable

accuracy. However, the construction and updation of graph consumes high response time during extraction process.

Noel et al^[59] proposed video event classification and retrieval over semantic analysis using a scalable and efficient modeling scheme. This proposed process contains three main components such as semantic representation for video frames, efficient method for generating temporal features from semantics and event model generation. In semantic classifier, video frames were represented in the form of sigmoid normalized semantic model vector, which constructs low-level features as GIST, color histogram, SIFT, edge histogram, Fabor filters and LBF. On completion of this, semantic relationship between features was estimated from linear temporal pyramid, which aggregates the dynamic boundaries based on distance of semantics between adjacent frames. Further, Hadoop Map-Reduce environment was implemented for mapping the set of features and reducing the features as per event by kernel selection and data scoring. Map-Reduce provide significant improvements in performance over temporal information on event classification.

Ahanger et al^[60] initiated the retrieval process with annotated metadata that involves segment and structure analysis. Based upon these kinds of multiple metadata set, video information was differentiated from composition. The first step in this system was to use the conventional technique for differentiating video segments from data using segmenting metadata. This means that the users given queries was matched with the annotated and unstructured metadata associated with segments and retrieves the relevant data for improving the recall. In second process the retrieval segments were clustered using vector based clustering approach for retrieving potentially value added data. In third stage the

transitive technique was applied to improve the recall process of retrieval. Finally, the creation time relationship was expanded to final candidate set of relationship for increasing recall. The main advantage of this approach was that it increases the recall rate of the data using four step hybrid approaches. However, the vector based clustering allows the incorrect segments that lead to degrade the overall performance.

Shuirchi et al ^[61] proposed an efficient video retrieval system using 3D shape in large number of digital video contents. The main objectives of this system are listed as follows: automatic extraction of video frames from video sequence, arrangement of video frames in virtual 3D space and provide search interface to increase search efficiency. Initially, video frames were extracted using Chi-Squared test of HIS histograms. Further, features were extracted automatically using color, texture and shape. Then, the frames are arranged in 3D vector space, so the images were located close to each other. For measuring similarity, neural network based unsupervised learning was designed, which provides topological relationship between images in high dimensional space. This process effectively retrieves videos by arranging cut frame images which are close to each other. However, this process uses conventional techniques for retrieving, which was not optimal to perform for 3-Dimensional.

Table 2.2 Different Conventional Techniques used in Image to Video Retrieval

Techniques	Algorithms/Methods	Benefits	Drawbacks
------------	--------------------	----------	-----------

Content based Video Retrieval	<ul style="list-style-type: none"> • Visual-word based semantic signature • Fisher based approach • Distance Learning 	<ul style="list-style-type: none"> • It improves the classification/recognition 	<ul style="list-style-type: none"> • Semantic problems occur due to Visual content is not correlated with semantic annotations.
Trajectory and Appearance using Video Retrieval	<ul style="list-style-type: none"> • K-Means Clustering • Gaussian Mixture Model • Adaboost classifier 	<ul style="list-style-type: none"> • It enhance the accuracy and performance in searching stage at run time 	<ul style="list-style-type: none"> • It produces inappropriate result due to environmental noises • Loss of visual features
CBIR using Multimodal and Multimedia Video Ranking Model	<ul style="list-style-type: none"> • Multilevel Learning • Graph Representation • Nearest Neighbor Ranking 	<ul style="list-style-type: none"> • It provides Efficient result when compared with SVM 	<ul style="list-style-type: none"> • High algorithmic complexity • It requires extensive memory
Video Retrieval	<ul style="list-style-type: none"> • Automatic Segmentation 	<ul style="list-style-type: none"> • Effectively identifies the distance between two 	<ul style="list-style-type: none"> • Re-Ranking increases the

based on Segmentation	<ul style="list-style-type: none"> • Structured Visual Retrieval 	image pairs	<p>computation time</p> <ul style="list-style-type: none"> • Clustering process is complex and difficult to determine number of segments
Multiple Instance Learning	<ul style="list-style-type: none"> • 3D Shape Retrieval • Classification • Graph based Matching 	<ul style="list-style-type: none"> • It robust against noise and outliers • It supports highly effective user annotation and intuitive visualization results 	<ul style="list-style-type: none"> • The selected frames are not optimal in representing corresponding video frames
CBIR with MapReduce	<ul style="list-style-type: none"> • Auto Color Correlogram Coefficient 	<ul style="list-style-type: none"> • It reduce the processing time by parallel process • It provides efficient result even large datasets 	<ul style="list-style-type: none"> • Redundant mapping increase time when an update to database is required
Krisch Descriptor based Video Retrieval	<ul style="list-style-type: none"> • Key frame Extraction • Graph based Representation 	<ul style="list-style-type: none"> • It is easy to classify the videos based on characteristics like color, shape, etc. 	<ul style="list-style-type: none"> • It takes more time for both graph construction an key frame extraction • It does not provide spatial information about frame representation

Content Based Video Retrieval (CBVR) using Ontology	<ul style="list-style-type: none"> • Object ontology • Machine learning (NN and fuzzy logic) • Semantic template 	<ul style="list-style-type: none"> • It improves the retrieval accuracy compared with traditional methods using color histogram and texture features 	<ul style="list-style-type: none"> • The response time of system is slow with the increase in number of rules • Accuracy is highly depend on the knowledge and human experts
Hashing for Large-scale Video Retrieval	<ul style="list-style-type: none"> • Deterministic Quantization • Dynamic Temporal Quantization 	<ul style="list-style-type: none"> • Similarity computation is efficient than Euclidean distance 	<ul style="list-style-type: none"> • It is difficult to obtain the spatio-temporal features due to fixed size of hashing
Machine learning tools for Video Retrieval	<ul style="list-style-type: none"> • Binary Bayesian Classifier • Low-level features • Artificial Intelligence 	<ul style="list-style-type: none"> • This process reduce the semantic gap 	<ul style="list-style-type: none"> • It requires large amount of training samples • The training set is fixed during leaning and application stages
CBVR using Clustering	<ul style="list-style-type: none"> • Fuzzy c-Means • PSO clustering 	<ul style="list-style-type: none"> • It improves the accuracy by achieving higher retrieval rate 	<ul style="list-style-type: none"> • This method is easily suffers from partial optimism,

			which causes less exact regulation of speed
Relevance Feedback performance	<ul style="list-style-type: none"> • Multiple-feedback • Motion-related icons 	<ul style="list-style-type: none"> • It enhance the relevance feedback information • It minimizing the semantic gap problem • This process improves the interaction between system and human 	<ul style="list-style-type: none"> • Motion analysis is difficult and computationally expensive
Map Reduce Parallel Computing	<ul style="list-style-type: none"> • Vector Space Model • Automatic large scale clustering 	<ul style="list-style-type: none"> • It provides high scalable and good fault-tolerant process • It improves the quality of retrieval 	<ul style="list-style-type: none"> • The performance of this process is not satisfy the practical requirements
3D Mesh Video Retrieval	<ul style="list-style-type: none"> • Landmark Tracking • Critical Point Tracking • 3D Facial Model descriptor 	<ul style="list-style-type: none"> • It increases the accuracy of video retrieval • It effectively estimate the both head motion and facial deformations 	<ul style="list-style-type: none"> • Storage maintenance problem while storing huge amount of data • High Complexity
Semantic data for	<ul style="list-style-type: none"> • Minimum Spanning Tree 	<ul style="list-style-type: none"> • Increases the recall rate of data 	<ul style="list-style-type: none"> • It does not effectively handle

Improving Video Retrieval	<ul style="list-style-type: none">• Clustering	<ul style="list-style-type: none">• It provides high accuracy due to efficient tree construction	the unstructured metadata
---------------------------	--	--	---------------------------



CHAPTER 3

OVERVIEW OF CBVR (2D AND 3D)



3.1 INTRODUCTION

Image processing is the most significant and wide research area that is used for several beneficial applications all over the world. At present, people share their thoughts, opinions and ideas through visual images and videos. This is more preferable by modern generation with smart phones and internet. This was awaited for the past 10 years of research to produce effective results in 3D video retrieval concepts ^[62]. Our research area in image processing is initiated to support worldwide multimedia services for people using smart phones in future and also at present ^[63]. Working with multimedia requires large amount of storage since the size of multimedia files are larger. Hence sharing large number of videos and images are possible only with larger storage capacity. So undergoing a research in multimedia requires analyzing the storage availability. Retrieval of images and videos from storage is performed based on visual semantics.

3D images and videos are highly enhanced with their pictorial/graphical representation by involving novel techniques of image processing, which magnetizes people towards 3D based images and videos retrieval. 2D has become older due to the introduction of 3D that has overcome the limitations in 2D and also 3D is used in real-time application. 2D is mostly focused for medicinal and industrial based applications. Research on 3D based video retrieval is necessary, due to the growth of recent technologies and active involvement of people have grown up along with latest trends, Soon in future, 3D will be used by people with their handy smart phones. Processing with 3D was challenging task in image processing research area. 3D plays major role, since it enhances user experience in watching videos and images. So our research work is completely focused on 3D video retrieval from Hadoop storage, which is enabled with storage of some Giga Bytes (GB) to Tera Bytes (TB). Hadoop is a framework that is

capable to store and process with enormous datasets using Map Reduce ^[64]. Storage issues and 3D videos was found to be most required concept in the domain of image processing which is not discussed by any researchers. Hereby, the main aim of our research work is to succeed by retrieving higher positive results in minimum processing time.

3.2 TECHNICAL TERMS

1) Video Frames

In Digital video, frames are the streams of captured images which compose the complete moving picture at regular time intervals. These images are considered as digitalized samples which comprises of visual (intensity and color) information at each spatial and temporal location. Usually, the visual information on each sample point is represented as the values of RGB color component space. Frame rate is defined as the number of individual frames that are projected per second, also known as Frame Per Second (FPS). The most common frame rates in video are 24, 25 and 30 frames per second. The frame size is represented with respect to Width (W) pixels and Height (H) pixels as W*H.

2) Gray level Images

The gray level image provides 256 levels of luminance per pixel of possible intensity to each pixel, hence this type of images denote 8 bits per pixel (bpp). The typical RGB color images, with 8 bits for Red, 8 bits for Green, and 8 bits for Blue, hence the intensity of RGB images is represented as

$$I = (R + G + B) \quad (3.1)$$

The intensity ranges of gray level are represented in a theoretical way as a range from 0 (black-weak intensity) to 1 (white-strong intensity) which carries only black-and-white images. The brightness of R, G, and B components is represented as decimal from 0 to 255 or binary 00000000 to 11111111. The intensity of gray level is directly proportional to the number representing brightness level of RGB colors. In gray level, black is represented as R=G=B is 0 or 00000000 whereas white is represented as R=G=B is 1 or 11111111.

3) Color Histogram

An image histogram is defined as the probability mass function of image intensities, which is formulated by counting the number of pixels belonging to each color. Generally, color histogram depends on certain color space such as RGB and HSV. It captures the joint probabilities of the intensities of three color channels (RGB or HSV). The color histogram is formulated as,

$$h_{A,B,C}(a, b, c) = N \cdot \text{Prob}(A = a, B = b, C = c) \quad (3.2)$$

Here, A, B and C denotes three color channels and N is the number of image pixels. Histograms have been widely used in many video retrieval based techniques, which provides the motion-invariant representation about frame and extract possible key frames.

The histogram of digital image with total possibility intensity levels in the range [0,G] is represented as the discrete function:

$$h(r_k) = n_k \quad (3.3)$$

Where r_k is the k^{th} intensity level in the interval $[0,G]$, n_k is number of pixels in image.

Further histogram equalization is performed to process with the image. It is the method of increasing the dynamic range of gray-level value in low-contrast image, which is generally achieved by transformation function that includes Cumulative Distribution Function (CDF). The following transformation represents the CDF as

$$S = T(r) = \int_0^r P_r(\omega) d\omega \quad (3.4)$$

Where, $P_r(\omega)$ is the probability density function of intensity levels, the normalized range of intensity from 0 to 1. The probability value is obtained by dividing all elements of $h(r_k)$ by the total number of pixels in image.

4) Geometry based features

Geometric features are the features of objects; it constructs digitized representations by a set of geometric elements like lines, curves, points, surfaces with the emphasis of precision and accuracy. Several geometric properties are found in objects, some of them include identifying edges that represents the boundary of shapes, determining shapes, curvatures estimation, computing tangents at boundaries, angles, etc. By using these properties, it effectively reconstructs the real objects from digital images. Based on the properties of geometric elements, it can be categorized into two types such as primitive and compound features they are,

- Primitive features
- Compound features

Primitive Features are very simple and significant features of object, whose main objective is to alter the appearance of the part of objects. It can be defined as the basic geometric entities that represent particular corners, blobs, edges, ridges, image texture and salient points. The Compound Features are a collection of several primitive entities, which contains more than two primitive features for extracting geometric elements. There are two types of compound features such as geometric composition and Boolean composition.

5) Topological Descriptors

Topological features provide sufficient global information about an object, which is commonly represented as “study of qualitative properties of certain objects”. Topology is defined into following sub-fields as,

- General topology
- Algebraic topology
- Differential topology
- Geometric topology

General Topology is also called as Point-set topology, in which the properties of topological space including compactness and connectedness are investigated. Algebraic topology uses algebraic formulations, it measures the degrees of connectivity which include homology and homotopy groups. Differential topology is similar to geometrical

entities, which involves the field of differentiable functions on differentiable smooth structure. The Geometric topology mainly focuses on low-dimensional main folds (sphere, torus, cross-caps, etc.). Geometric topologies are of orient ability, local flatness and handle decompositions.

6) Euclidean Distance

Euclidean distance is defined as a straight line distance between two pixels in Euclidean space; this is evaluated using Euclidean norm. It estimates the minimum distance between each pixel and the nearest non zero pixels to binary images. For instance, two points are considered as X and Y in two dimensional Euclidean space. Here, 'X' is coordinated with $(x_1, x_2 \dots x_n)$ and 'Y' is coordinated with $(y_1, y_2 \dots y_n)$. The distance between two end points are defined as the square root of the sum of the squares of differences between corresponding coordinates of points. The Euclidean distance between two points is formulated as,

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (3.5)$$

In video retrieving process, Euclidean distance algorithm uses the mapping of distance between two frames. The map labels of each pixel with distance to nearest boundary pixel in binary image. It is used to calculate the distance between two consecutive frames. The input video frames and videos from datasets are extracted and distance between frames is mapped to determine the videos for retrieval.

7) Mean Squared Error (MSE)

MSE is a significant criterion which is utilized in order to measure the performance of an estimator. MSE is widely used to measure the degree of image distortion since it represents the overall gray scale value error in entire image. The MSE in statistical form is calculated as,

$$MSE = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} [f(x, y) - \hat{f}(x, y)]^2 \quad (3.6)$$

Where, 'f' and \hat{f} represents the original and distorted images respectively, M, N is the dimension of the images. A lower value of MSE provides less error whereas the higher value of MSE gives high error. The characteristics of MSE are illustrated in the following,

- If the MSE is zero, it will be considered as perfect accuracy by the estimator. However, this condition is not practically possible
- The value of MSE is used to create comparison between two statistical models
- MSE is utilized to determine the number of predictors, which involves a model of given reflection set
- The variance analysis evaluates MSE as a part of statistical analysis

8) Voxel

Voxel is a unit of graphical information which defines a point (picture element) on a regularly spaced 3D grid. In 3D space, each element is represented with X, Y and Z coordinates which defines its color, density and position. Voxel is the process of adding

depth to an image using set of cross-sectional images (pixels) known as volumetric dataset. In the voxel method, the voxel surrounds the central grid point and the data value is constant in the voxel. Voxel images are significantly used in the field of Computed Axial Tomography (CAT) scans, medicine, Magnetic Resonance Imaging (MRI) scans, computer games, etc. Voxels are categorized into various approaches such as,

- Slice based - Volume is sliced into one or more axes, it stores color information in voxel
- Sculpture - User stores density of information in voxel
- Building blocks - Users can insert and delete the blocks like construction set toy

3.3 KEY FRAME EXTRACTION & SELECTION

Calic and Thomas^[14] proposed an efficient key-frame selection algorithm, which analyses the behavior of spatial-temporal regions of video categorization. Initially, segmentation process was performed by applying anisotropic diffusion filtering, which removed the noise and irrelevant image background. After this, two stage k-Means clustering procedure was applied to separate the 3D videos into set of regions. In first stage, color similarity was calculated based on pixel values. In second stage, clustering was processed for finding the average color values based on Euclidean distance. Further, key-frame extraction was done by using heuristic rules, which improved the spatial analysis. In key-frame extraction, temporal complexity reduction algorithm named shot detection module was applied, for identification of boundary and frame-frame difference. In addition to the complexity of frame reduction, low-resolution representation was also

applied for reducing the multi frames, where extraction of key frame process. Short type classifier was used to identify the region features and region relations; these are taken as input to the key-frame extraction. The following equation (3.7) provides frame selection based on linear combinations ($\Omega(i)$) of relations and features.

$$\Omega(i) = \sum_{VR_i} \alpha(i, T) \cdot \varphi(i) + \beta(i, T) \cdot \rho(i) \quad (3.7)$$

Where, $\rho(i)$ is denoted as region relations, $\varphi(i)$ represented as region features, $\alpha(i, T)$ and $\beta(i, T)$ are defined as heuristic rules for type T. The frame ($\Omega(i)$) which has maximum value was considered as key-frame.

Kin-Wi et al ^[15] proposed an optimal representation for video shots, which transmits both spatial and global information about frames in the video shots. The main aim of this process was to achieve a high video retrieval performance. In this process, histogram representation was used to provide motion-invariant representation of a frame which was commonly used in many works. Histogram representation was important for extracting key frames in efficient manner. The average histogram (AH(j)), median histogram (MH(j)) and alpha-trimmed average histogram (TrimHist(j, α)) values were estimated for selecting the pixel values from all frames. The equations are as follows:

$$AH(j) = \frac{1}{M} \sum_{i=1}^M H_i(j), \text{ for } j = 1 \dots B \quad (3.8)$$

$$MH(j) = \text{median}\{H_1(j), H_2(j), \dots, H_{M-1}(j)\} \quad (3.9)$$

$$\text{TrimHist}(j, \alpha) = \frac{1}{M-2 \cdot \lfloor \alpha M \rfloor} \sum_{m=\lfloor \alpha M \rfloor+1}^{M-\lfloor \alpha M \rfloor} h_j(m) \quad (3.10)$$

Where, H_i denotes the histogram of i^{th} frame, 'B' is represented as number of bins in histograms, 'M' is the number of frames, $h_j(m)$ is denoted as ordered array of bin values, ' α ' ($0 \leq \alpha \leq 0.5$) is trimming parameter. Further, Temporally Maximum Occurrence Frame (TMOF) was constructed for improving the optimal representation of video slots. In this process, TMOF represented the key frames based on content based video retrieval which captures significant visual representation in video slots. Further, TMOF was enabled to identify the k most frequent occurring values and k highest peaks of probability distribution at each of its pixel positions. Finally, alpha-trimmed histogram based key extraction was compared with TMOF, which outperforms the key extraction process.

Qiang Zhang et al ^[16] suggested a novel key frame extraction method with respect to motion capture data. This paper was focused on automatic key frame extraction by adaptive threshold. Initially, unsupervised clustering algorithm was performed for motion sequence, in which process the similarity based on distances of adjacent frames were detected. Further, improved ISODATA clustering method was designed to identify the improper selection of frames and cluster all proper frames. Improved ISODATA method was based on two rules such as split and merge data. In splitting process, the frames were splitted in accordance to standard deviation (between two frames). In merging process, two categories of frames were merged with respect to similarity measures of adjacent frames. Finally, this methodology was compared with two other previous methods which includes reconstruction motion and mean absolute error.

Peng Huang et al ^[17] designed a method of 3D video summarization which detects the set of key frames automatically without considering motion analysis and shot detection. The key frame extraction process was based on shortest path computation in graph with the support of self-similarity measures. The excellence of 3D video summarization depends on rate (representative cost) and distortion (cost accuracy), rate was considered as entropy of key frames and distortion was defined from the information loss between key frames and original sequence. Based on the rate and distortion, conciseness was estimated which was defined as weighted sum of rate and distortion. Then, conciseness cost matrix (C) was computed for constructing the graph, the equation (3.11) is given as follows:

$$C := (c_{i',j'})_{N_s} \times \left\lfloor \frac{N_s+1}{2} \right\rfloor \quad (3.11)$$

$$(c_{i',j'}) = (1 - \beta) \cdot d_{i',j'} + \beta \quad (3.12)$$

Where, 'β' represents the weight of rate and distortion, $d_{i',j'}$ is the distortion cost matrix that can be derived from self-similarity. Then, location optimization was performed by considering each key frame (node) and number of adjacent frames (neighbor node) to estimate shortest path (Dijkstra's algorithm) in graph. The results of short sequences improve 3D video sequences of movement with several dynamics.

Tong yee et al ^[18] designed an animating mesh representation based on key frame extraction. 3D animating meshes are wide application used in computer graphics and video game industries. The main aim of this paper was to simplify the deforming mesh under minimized animation distortion constraint. In this proposed process, deformation

analysis of animating mesh was used for detecting the animated key frames based on geometric features and motion characteristics of video sequence. In this approach, binary encoding method named bit string was designed for encoding each frame as single bit. Here, 1 represents as key frame and 0 is non-key frame respectively. This approach was capable to provide compact representation of animation in spatial as well as temporal domain.

Maria and Athanasassios ^[19] proposed content based video retrieval by key frame extraction. In this paper, sequence search algorithm was designed for key frame extraction. This method was performed for processing video segmentation based on temporal characteristics. The main objective of this process was to model a fully automatic key frame extraction process from the video sequences that was enabled to be supported over real-time videos. The proposed algorithm was comprised of two significant parts i.e. (i) DCT domain feature extraction and (ii) key frame extraction. In DCT, each frame was compressed up-to the stage of inverse quantization which results DCT coefficients. Based on DCT coefficients, features are extracted from I-frames and similarity values were computed for extracting feature vectors. In turn feature vectors are involved for decision making in key frame selection. Larger numbers of key frames are extracted from video with respect to the threshold value. This paper work was supported to minimize the laborious task or offline video processing. This significant reduction was achieved by exploiting DCT coefficient in feature extraction.

Zhonghua Sun et al ^[20] proposed a method to extract key frames within video sequence by considering spatial-temporal color distribution with respect to time. Two main processes were considered for key frame selection: (i) construction of temporarily

maximum occurrence frame throughout video and (ii) estimation of weighted distance between frames. The salient visual information was preserved, since the changes in visual content and its characteristics are taken in account. TMOF algorithm was used for each video slot; the color histogram of each frame was compared with color histogram of TMOF. By comparing pixel values at same position in frames present within a shot, a reference frame was constructed from color pixel value throughout the frames in same position with maximum occurrence. The preferred TMOF frames of each slot provide descriptive frame for both color and temporal information of video slots. Similarly, distance between histogram and TMOF was estimated for each frame with its neighboring frames. Distribution of features was described by comparing distance between each frame in the shot with TMOF. Then, key frames were extracted with respect to peak of key frame shot.

3.4 BAG OF VISUAL WORD (BOVW) CONSTRUCTION

Chin-Fong ^[21] reviewed related works based on content-based image retrieval system by applying widely used feature representation method, Bag-of-words (BOW). It was mainly designed for image analysis that provides visual analogue of visual words. The following steps are performed to extract the BOW in images: (i) interest point detection, (ii) local descriptors, (iii) visual vector quantization, and (iv) learning methods. Further, the extension methods of BOW representation are feature representation, vector quantization, image segmentation, vocabulary construction, etc. BOW features were applied for other applications as face detection, 3D video retrieval, 3D image retrieval, medical image analysis and so forth.

Ke Ding ^[22] proposed a 3D model descriptor-BOVW descriptor which was represented by histogram that calculates occurrences of words. The main aim of this BOVW was to reduce the storage space and computational complexity while performing 3D model retrieval. The process of constructing BOVW was based on the following steps: (i) Initially, Light Field Descriptor was used to represent the 3D models into a set of projected views. (ii) Then, Fourier descriptor was applied to quantize the projected views into Fourier coefficients. (iii) Gaussian means (G-means) clustering was designed to group similar projected views of each 3D model. (iv) Finally, the codebook was constructed based on obtained clusters of 3D model by using K-means Clustering algorithm. After completion, multi-resolution histogram was composed of number of BOVW descriptors, which reduces the difficulty of k-means clustering (selection of optimal K). Then, novel pyramid matching based similarity was defined for 3D model comparison, which was determined with respect to the size of histogram's bins.

Junaid et al ^[23] proposed a BOVW model for video segmentations into scenes, which was used to compute key descriptors from video shots. Video sequences were divided into shots as several frames initially. Then the similarity value was calculated for each frame to merge the most similar frames. The visual words are represented for each frame from large and sparse histogram of visual words. This proposed BOVW model was applied for feature quantization; each image was represented by sparse histogram of visual words which supports faster and reliable image matching from large databases.

In BOVW model, each frame (f_i) was separated into set of local key point descriptors as $f_i^p = \{p_{i,1}, p_{i,2}, p_{i,3}, \dots, p_{i,m}\}$. The function of BOVW was defined as follows:

$$B: R^d \rightarrow [1, N] \quad (3.13)$$

$$P_{i,j} \rightarrow B(p_{i,j}) \quad (3.14)$$

Where, $(p_{i,j}) \in R^d$ is a map descriptor which was used to produce integer index. After construction of BOVW, scene detection was computed based on distance between visual words histogram and closer shots. Finally, sliding window was used to measure the similarity of each shot and its neighboring shots.

Zhouhui Lian et al ^[24] developed a visual similarity based 3D object retrieval, which supports bag-of-features and efficient shape matching process. The main objective of this paper was to determine the 3D model view as histogram words. It was obtained by vector quantization of salient local features and shape matching. In this process, SIFT features were used to extract the depth-buffer image into histogram, whose representation was given in distinctive way for constructing codebook. Finally, efficient shape matching process was carried out to compute the dissimilarity between two objects by estimating minimum distance for all possible matching pairs. Then, effective experiments also investigated the number of views, codebook, training data and distance function.

Konstantinos et al ^[25] designed a complete 3D models based on set of panoramic views and BOVW models. Initially, 3D models were represented by set of panoramic view (cylindrical projection) by using PANORAMA projection methodology. Panoramic object representation for accurate model attributing (PANORAMA) was employed in panoramic views to capture the position of model's surface information and its orientation as 3D model descriptor. Simultaneously, SIFT algorithm was applied to

hierarchically separate the spatial area of views. These SIFT description was used to form the BOVW model, which contains local features of image. Then, codebook was generated by considering the visual features of images with respect to clustering process.

After construction of codebook, the closest features were computed and matched with visual features by using K-means clustering algorithm. Then, spatial histogram was computed for each result of image, which improves the representation of codeword matching process within codebook. *Takahiko and Ryutarou* ^[26] proposed an improved and more accurate 3D model retrieval algorithm using local visual features. Improved Bag of Features- Scale Invariant Feature Transform (BF-SIFT) algorithm was designed for feature sampling and feature encoding process. BF-SIFT algorithm was comprised with significant process such as review of the below mentioned procedures,

- BF-SIFT algorithm
- Dense sampling
- Fast encoding

1) Review of the BF-SIFT algorithm

This process was executed by the following processes (i) Pose normalization with respect to position and scale, (ii) Multi-view rendering, (iii) SIFT feature extraction, (iv) Feature Encoding, (v) Histogram generation and (vi) Distance computation.

2) Dense Sampling

A pyramid SIFT algorithm was used to concentrate on image samples, to identify whether it is situated near 3D object or not. Based on intensity value, sampling rate was assigned for each pixel.

3) Fast Encoding

In fast encoding, local features are involved in two different steps: Codebook learning- clustering of similar features and Encoding- it was a vector quantization process which produced the closest vector in high dimension space. In this process, SIFT feature based algorithm was introduced for increasing the depth analysis of images. Then, dense sampling process enhanced the number of local features per model according to SIFT feature extraction and vector quantization steps. Clustering process was used to reduce the computational cost and also speedily extracts SIFT features for considering the closest vector with respect to neighboring process.

Yue Gao et al ^[27] introduced a 3D multi-view representation method named bag-of-region words method for extracting visual features in region level. In feature extraction, four steps were included such as SIFT feature extraction, SIFT feature quantization, region splitting as well as representation and region clustering. Initially, grid point were chosen in each image and SIFT local features were extracted from these grid points.

Each object was assigned into set of views, which was used to select a set of uniform distributed points in the object region. Further, each local feature was coded into visual words with a pre-trained visual vocabulary. Then, each view was split into a set of regions and selected region was represented by bag-of- words feature. All the achieved regions with normalized bag-of-words feature were further grouped into clusters using

hierarchical agglomerative clustering method. Then, one feature was selected as representative from each cluster with corresponding distance weight value.

Ivica et al ^[28] proposed an improvement of BOVW approach for efficient and simple image retrieval process. The main aim of this paper was to reduce the sizes of image databases and improve the stability of retrieval systems. In this process, the BOVW features not only construct codebook but also it concurrently builds a complete indexing structure. The indexing structure was effectively used for retrieving the relevant images with respect to ranking process. Random forest of Predictive Clustering Tree (PCT) was applied to construct codebook and index structured through the methods of decision tree and decision rules. In codebook generation, the system consists of off-line and on-line phases. Off-line phase generates the subset of SIFT local features from input images, then PCT was constructed based on descriptive attributes (unique index/identifier) and clustering attributes. Further, TF-IDF weighting scheme was applied for creating index structure, which easily discounts the frequency of visual words. On-line phase was used to extract the local SIFT descriptors from query image, which produces the ranked list of images. On measurement of similarity between two local SIFT descriptors of images were calculated over a random forest of PCT.

3.6 FEATURE EXTRACTION

Xianheng et al ^[29] focused on robust color-feature model of video objects, which was processed by converting RGB pixels to hue color circle. The main objective of this proposed process was to improve the robustness and accuracy of video object retrieval. Two stages are involved for performing video object retrieval such as color feature

extraction and object retrieval. A histogram of 13 perceived colors of human was accumulated for each tracked object, which was presented with RGB values.

Kalman-filter was applied for selecting the tracked object within video object sequence. Then, the RGB pixels were converted into Hue-Saturation-Value (HSV), which was efficient to human perception in color view. The following equations were used for converting the RGB values into HSV.

$$h^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} h_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq h^* \leq 360) \quad (3.15)$$

$$s^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} s_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq s^* \leq 360) \quad (3.16)$$

$$v^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} v_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq v^* \leq 360) \quad (3.17)$$

Where ' τ_{enter} ' is denoted as number of frames that are entered into tracked object view, ' τ_{leave} ' is the number of frames that are leaved into tracked object view, (h_i, s_i, v_i) are the i^{th} frame of hue, saturation and value.

Conrado et al ^[30] proposed a combination of shape and color features for searching and retrieval of 3D models. 2D shape distribution was focused on shape based measures whereas color feature vector was considered for color based measurements. 2D shape descriptor was invariant for 3D model translation and rotation. Here, the shape features are considered as areas, distances, 2D projections and angles. In order to compute the best shape descriptor, two simplest and most effective mean and distance based measurement was made between two target points. In color measures, global color

characteristics of the model was represented, which was 159-bin color histogram. Here, each bin represents the percentage of color in model, then, the similarity of color between query model and 3D model was estimated using similarity matrix. Further, multi-feature similarity measure was computed based on shape and color based measures. The similarity measure was performed with respect to different threshold levels and different shape and color weight values. Based on this process, the color and shape features improve the performance (precision and recall) of retrieval system.

Hinge et al ^[31] developed an automated feature extraction for content based retrieval using map reduce. This paper was mainly proposed to overcome the problem of conventional process such as low accuracy and failure of handling huge amount of datasets. In CBIR, the primitive features (color, texture and shape) were considered for analyzing the features of images.

1) Color features

The dimensional color features were defined which include RGB, HSV and HSB, in which color information was stored as color histogram.

2) Texture Features

Further, texture features were extracted with respect to degree of contrast, coarseness, directionality as well as regularity and randomness.

3) Shape features

Shape was considered as low-level features, which was recognized by their shapes such as circularity, Lake Factor, convexity, direction, eccentricity and relative size.

Further, these features were extracted in parallel form by using map reduce and the extracted features are stored in HDFS file system. This parallel process was efficient in reducing computation time since it uses map reduce. Then, similarity matching was performed between two images based on Euclidean distance measurement. In this process, the low distance images were considered as similar images, which were retrieved to user for the given query.

Federico et al ^[32] proposed a novel framework for partial and global 3D shape matching and retrieval process. This novel framework was combined with topological (graph representation) and highly discriminative geometric features of 3D object. Initially, the object segmentation and extraction was performed using medical surface-segmentation method. Then, the readjustment technique was designed to reduce the noise surface present in object. After readjusting the medical surface, meaningful parts assignment criteria was developed for object decomposition which was used to segment the meaningful parts. Then, super-quadratics have been selected for producing very compact representation of 3D segments since the segments of objects were expected to have shape which was approximated with a super ellipsoid. Matching process was performed by using attributes local geometric features. Then, cost function was utilized to define optimal matches, by minimizing the distances between corresponding local features. Finally, dissimilarity metric was used to effectively combine the results of graph-matching procedure and distance between complex geometric features.

Zechao *et al* ^[33] proposed a novel technique of unsupervised feature extraction using high dimensional data. For unsupervised feature selection process, combined version of structural analysis and cluster analysis named as clustering-guided sparse structural learning (Cgssl). Then, non negative spectral analysis was proposed, which was used to detect the more accurate cluster indicators for discriminative feature selection. The cluster indicators were predicted by original features together with features in the low-dimensional subspace. Then, latent structure analysis was exploited by different features to detect the cluster indicators. Sparse feature selection models were exerted to facilitate the feature selection based on the regularization term.

Yujie *et al* ^[34] designed a curvature-based feature extraction method for 3D model retrieval, which was based on geometrical and topological features. In curvature feature extraction, Voronoi area of the 1-ring neighborhood triangles was computed on each vertex. Voronoi area was depending on mesh triangles, the voronoi area was formulated below:

$$A_{Voronoi} = \frac{1}{8} \sum_{j \in N_1(i)} (\cot \alpha_{ij} + \cot \beta_{ij}) \|x_i - x_j\|^2 \quad (3.18)$$

Where, $N_1(i)$ is the neighbor triangle of vertex x_i , α_{ij} and β_{ij} are represented as opposite angles of edges x_i and x_j respectively. Further, mean curvature vector was computed for each triangle mesh surface. After identifying the features, Earth Mover's Distance (EMD) was designed for evaluating dissimilarity between two multi-dimensional distributions in some features space where distance was measured between single features. The dissimilarity was computed from distance of different dimensional features and its weights.

Yu-Chi et al ^[35] developed a novel 3D shape descriptor and provided efficient method for indexing and matching 3D models. This proposed 3D model described combines both geometric and topological characteristics of surface-based model. Initially, 3D features were extracted for all 3D models and these features were stored in feature database, this process was considered as off-line work. In on-line work, system computes the corresponding features based on user's query and then matching features into database for obtaining results. The feature extraction process was handled with three models such as

- Global features of 3D models
- Selecting sampling points
- Feature matching

In this global feature of 3D models process, 3D histogram equalization was estimated to represent the significant shape characteristics which have geometric transformations such as scaling, rotation and translation. These extracted shape features were capable to maintain the feature invariants. In order to measure the similarity of features, Euclidean metrics, topological metrics and adaptive shape feature was considered.

Selecting Sampling points is present for choosing a sample point. To obtain a sampling point, stochastic triangle method was used, which extracted sampling points from the shape feature descriptor. These sampling points were further used for extraction process.

Then Feature Matching is performed by the authors in this work. After constructing the features, quadratic form distance function was applied to find the similarity between two feature vectors. The equation is as follows

$$d_A^2(p, q) = (p - q) \cdot A \cdot (p - q)^T = \sum_{i=1}^N \sum_{j=1}^N a_{ij} (p_i - q_i) (p_j - q_j) \quad (3.19)$$

Here, p_i, q_i are the N dimensional vector spaces with respect to 'i'. a_{ij} represents the element of weighted matrix A. *Murala et al* ^[36] developed an algorithm for indexing and retrieving images for content based image retrieval using second order Local tetra Patterns (LTtPs). In this process, the combination of local binary patterns (LBPs), Local Derivative Patterns (LDPs) and Local Ternary Patterns (LTP) were estimated by referenced pixels and surrounding neighbors calculates difference in gray level. LTtP was applied to describe the spatial structure of local texture and the magnitude of binary patterns was collected from magnitude of derivatives. This method provides relationship between referenced pixel and neighbor pixel with respect to direction that are computed by derivatives in first order vertical and horizontal direction. In this process, the LTtPs method performance was compared with LBPs, LDPs and LTP on gray scale images for improving the retrieving process.

Hua Zhang et al ^[37] proposed a novel technique of shot boundary detection based on color feature aiming to obtain accurate detection. This proposed technique was enabled to detect shot boundaries changes through the analysis of color histogram distortion differences and an adaptive threshold value, which was based on sliding window content. This sliding window process was improving feature extraction with respect to computation time. For gradual transitions such as fades and dissolves, a preprocessing

has been introduced, and local histogram differences are quantified to binary values by selecting a threshold automatically with reference to the variation of histogram differences.

3.7 SIMILARITY MATCHING

Ting-Chu et al ^[38] proposed a novel query-adaptive Multiple Instance Learning (q-MIL) framework for enhancing video instance search using query image and annotation frames. Initially, video frames were selected based on query image which consists of Object Of Interest (OOI), that provides improved retrieval performance. In this process, the video instance retrieval was performed based on significant processing steps such as feature representation, query adaptive multiple instance learning and object search in videos. For feature representation, the integration of features were taken in account for each segment of frames which include SIFT, texture, shape and color information. After completion of feature extraction, q-MIL algorithm was applied for selecting positive and negative frames in both query input and video sequence. In order to find the positive and negative image, q-MIL algorithm utilizes a window detector with respect to OOI. Further, ranking list of video frames was detected and video frames containing at least one positive window proposal were considered as positive frame with OOI was presented.

Shekar et al ^[39] developed an accurate approach for video retrieval concept with clustering of the kirsch local descriptors and matching of query frame, which was performed by k-nearest neighbor searching algorithm. In the given video, the shot

boundary was detected using Gabor moment features, which obtained the mean and standard deviation values from the video frames. Then, spatio-temporal color distributions were applied to extract the key frames from every shot. Once the key frames are extracted, the local descriptors were estimated based on kirsch features, which was considered as directional features for vertical, horizontal, left and right diagonal directions of video frames. Further, each local descriptor was compared with k-clusters using k-nearest neighbor algorithm. The best match key point was considered as the highest key points and that was assigned to the matching key frame. Finally, the given query frame was compared with each key frame of the video database and retrieve the best matching for user's query.

Steven et al ^[40] proposed a novel multimodal and multilevel ranking framework for content based video retrieval. The main objective of this paper was to represent the videos by graphical structure and multimodal resources were obtained using harmonic ranking functions. Initially, the video sequence was represented by hierarchal structure which consists of video, video stories (semantic story of video frame) and video shots (key frame). With respect to graph representation, the ranking task was formulated over the graph. In ranking function, the weight value was estimated between labeled (relevant video) and unlabeled (irrelevant) examples from graph. Here, the weight between two examples was estimated as below:

$$w_{ij} = \exp\left(\sum_{k=1}^d \frac{(x_{i,k} - x_{j,k})^2}{\sigma_k^2}\right) \quad (3.20)$$

Where, $x_{i,k}$ is the k^{th} component of node x_i , σ_k^2 represented as length of parameter at each dimension. Further, harmonic ranking function was performed based on hitting probability interpretation and multimodal fusion. This proposed multi ranking process enhanced the retrieval performance and computational efficiency. Firstly, the top ranked video stories were obtained based on textual approach. Secondly, visual information and nearest neighbor strategy was applied for retrieving ranked video slots. Thirdly, Support Vector Machine was employed for ranking and finally, semi-supervised ranking function was performed to retrieve top video slots.

Ranjit kumar and nandini ^[41] proposed a Semi-Structured Clustering Technique (STAR) for video retrieval, which was performed by content based image retrieval and distance functions. The main objective of this STAR technique was to achieve effective video retrieval and reduce the computational complexity with respect to time. In STAR, the combination of image boosting and distance clustering was performed. Initially, the database videos were converted into frames. User queries enters into system, and then it was forwarded to image boost cluster calculations, which calculates the distance between query image and frames by using point-point Bregman distance function.

Then, semi-supervised clustering was performed to group the labeled data and unlabeled data separately. The clustering process was processed with respect to agglomerative method and divisive method which improves the accuracy and efficiency of retrieval. Finally, the matched frames were clustered with given input query image and system retrieves required video based on cluster frames.

Zechao et al ^[42] discussed about semantic based data representation, which uses the techniques of robust structured subspace learning to guarantee the subspace using L2, 1-norm. The learned subspace was considered as an intermediate space, which was used to reduce semantic gap between low-level visual features and high-level semantics. It also considers local and global structural consistencies and was robust against noise and outliers. This proposed process was included with several image processing tasks such as image tagging, clustering and classification. Hence, this proposed subspace learning algorithm was slightly effective to uncover the latent subspace, which was used for the semantic retrieval of images in 3D models.

Hanli et al ^[43] presented a new method for Near Duplicate Videos (NDVs) retrieval with respect to parallel processing techniques which include Graphical Processing Unit (GPU) and Map Reduce. In this framework, multimedia and Intelligent Computing Cluster (MICC) was designed to retrieve large-scale NDV. The proposed MICC was show improvements in following aspects: Data loading and transmission, partition and sort procedure and computing power. In MICC, key frames were extracted from video sequences and uploaded into HDFS for constructing training dataset. To obtain the feature vectors, local feature descriptor was applied. Further, k-means clustering algorithm was performed to generate the visual words from feature vectors based on spatial and temporal information. Inverted File Index (IFI) was constructed to represent the bag of words using map reduce process. From the IFI, the similarity score was computed between query and reference video (frame) in database. Two-dimensional Hough transform was employed to identify the similarity scores of each frame levels into video levels.

According to this survey, it is clear that many authors have focused on image and video retrieval concepts. The limitations in previous works are provided with more attention in this proposed work which is detail discussed in further chapters.

Table 3.1. Deficiencies of Previous Works

Previous works	Contributions	Limitations
Markov Chains: Application to Online Image Retrieval	Markovian Semantic indexing works for online retrieval. This performs automatic annotation and indexing process	Larger time consumption and not able to meet with user satisfaction
Video Instance Retrieval	Query-Adaptive Multiple Instance Learning Algorithm	Applicable only for 2-Dimensions
Search-Based Face Annotation	Unsupervised Label Refinement and Locality Sensitive Hashing	Duplicate identity (names) images are present in real-time
Cross-Modal Multimedia Retrieval	Correlation Hypothesis, Annotation Hypothesis. Increasing availability of multimodal information demands novel representations	Not able to produce efficient result with huge amount of dataset

	for content based retrieval.	
Rank Fusion for Image Retrieval	Vocabulary and holistic feature strength	Re-ranking is not sufficient, since it leads in reduction in quality
A Semi-Supervised Active Learning Framework for Image Retrieval	Semi Supervised Learning and Support Vector Machine	Repetition in data retrieved
Cascade category aware visual search	Bag of Words model query distance metric learning and supervised visual vocabulary selection	For large scale of dataset, a larger memory is required

<p>Ordinal Distance Metric Learning for Image Ranking</p>	<p>Distance metric learning, Multiple features are considered to provide accurate result.</p>	<p>Higher time consumption</p>
<p>Hierarchical Video Classification, Annotation and Visualization</p>	<p>Ontology with Hierarchical Boosting Multi-model Boosting algorithm</p>	<p>Using a single ontology concept was not able to meet all the needs</p>



CHAPTER 4

3D CNN WITH MAPREDUCE PARADIGM & VOLUMETRIC SHAPE REPRESENTATION FOR CBVR



4.1 INTRODUCTION

A semantic based CBVR is an important challenge for retrieval of complex queries. A crucial challenge for CBVR is to provide a reasonable accuracy for the given query. For any query relevant results, contents must be optimum i.e. visually similar and semantically relevant to the given query image. However, this will not be possible and easy for the real-world noisy and complex content based videos and the queries are complex structures which contains complex shapes and objects.

In this case, how to design an intelligent algorithm for obtaining the state-of-the-art accuracy is a challenging task. Processing video requires computationally expensive memory, energy and power. The huge volume of video processing in a traditional database cause severe issues and it bring to the key issue in this CBVR. Hence, an effective retrieval model is requested for searching the videos with the maximum number of positive results. Further, waiting time of the user must be less than 1 second and also all positive results must reach to the maximum accuracy. Semantic indexing is the main process of CBVR since it extracts the semantic features from the video. Further, it must be adjusted for arranging, but how to index videos by semantic features is a challenging task. In this work, a modern video retrieval system can be existed already for real-world applications. The main reason is that semantic concepts are quite different from the image features and the semantic concepts are indexed is still an unknown problem. Particularly, concepts are automatically extracted using feature descriptors such as SIFT and SURF with limited amount of accuracy. The inconsistent representation leads to the inaccurate search results if not properly handled.

Searching by semantic queries is more consistent with human's understanding and reasoning about the task where a relevant video is differentiated by the presence of certain semantic concepts. There are sub-problems as Semantic Query Generation and Multimodal Search are existed in CBVR. The multimodal search component was used to retrieve the ranked semantic results.

4.2 PROBLEM FORMULATION

We contribute 3D CBVR in Map Reduce as our novel concept due to problems identified in previous works. Temporally Maximum Occurrence Frame was focused on key frame extraction for video retrieval ^[15] by smoothed histogram using Gaussian filter. The STPGF for key frame extraction was represented as

$$TMOF(i, j) = b_{opt}, 0 \leq i \leq W' - 1 \text{ and } 0 \leq j \leq H' - 1 \quad (4.7)$$

From the equation (4.7) the key frames are extracted however it was only suitable for 2D based retrieval system. Further, shape extraction was concentrated on both topology and geometric features. This process was initiated with Median Surface Extraction ^[68] using Average Outward Flux (AOF). It was estimated with ' $\nabla D'$ ' (i.e.) Euclidean Distance Transform's Gradient field and 'n' neighboring voxels. AOF was evaluated based on the voxel 'x' value given as,

$$AOF(x) = \frac{1}{n} \sum_{i=1}^n (\nabla D(x_i), n_i) \quad (4.1)$$

This evaluated AOF performs effectively in Medial Surface Extraction hence it was not capable to guarantee geometric features. On the completion of feature extraction it

has been continued with codebook generation for the construction of BOVW. This extracted BOVW consists of repeated visual words and so they were considered to be error and also they consume larger amount of time. Hereby we finalize to overwhelm all such problems previously existed and produce effective results.

4.3 RESEARCH FINDINGS

The main ideal of our research is to design an efficient novel methodology for content based video retrieval application in Map Reduce framework using BOVW. The main aim of our research is to achieve,

- Enhance user experience by 3D video retrieval with reduced memory complexity and retrieval
- Storage and processing in Hadoop
- Storage complexity is minimized using Hadoop Map Reduce framework
- Error free BOVW construction
- Effectively extract the shape features with topology and geometry
- Reduced outliers in visual codebook generation
- Relevant results retrieval from combinational similarity matching algorithm
- Larger the query with larger positive results

- Increase the accuracy rate for efficient video retrieval results

All the achievements listed above are attained in “3D CBVR using Map Reduce” research work. To succeed with designing an efficient “3D CBVR”, this work designed novel mechanisms to solve the challenges existed [65], [66], [67], [69]. In general, videos are large sized files which consist of ‘N’ number of frames in it. As per the length of the video increases, the number of frames also increases, hence processing with large number of frames and analyzing each frame was complex and also consumes longer time to retrieve relevant videos from database (storage).

Hence, in video retrieval concept storage remains as an unbroken problem, this problem is recognized and solved by Hadoop. It supports processing of Big Data (TB), storing a diverse set of data and parallel data processing. Hadoop with these provisions is considered to be more favorable to store data with larger sized as 3D images and videos. Hadoop is also a scalable storage platform since, it stores and distributes huge amount of data. In our research, a complete survey is held over image to image retrieval and image to video retrieval which was supportable for identifying the problems, challenges and issues existed previously. Hereby, this work further continues with brief discussions about the key findings of 3D video retrieval.

4.4 PROPOSED HYBRID 3D CNN

4.4.1 System Model

Content Based Video Retrieval (CBVR) plays an imperative role in the field of multimedia retrieval applications. Most of the previous process ignored its concentration

over reduction of processing time. Since video retrieval takes larger amount of time for processing. Most of the researchers focused on BOVW and 3D model individually. Our proposed framework of 3D CBVR is comprised of key frame extraction, shape feature extraction, codebook generation, construction of BOVW and finally matching process. This overall process improves our proposed research results in terms of accuracy, searching time, storage and computation complexity.

Figure 4.1 illustrates the overall architecture of proposed work, which works on offline mode and online mode. All admin side processes are performed in offline mode (O_f) while user side processes is in online mode (O_n). In order to effectively handle larger number of videos and result efficacious, Hadoop HDFS and Map Reduce is involved. The main functionality of HDFS is to automatically manage all the multimedia data to avoid data loss. Overall working procedure of Map Reduce in retrieval application is processed along with key frame extraction, feature extraction, building up BOVW and matching. Initially, videos are uploaded by admin that are converted into frames through frame conversion tool and converted frames are allocated into dataset. From the entire frame sequence, we select a particular key frame which contains all visual important information within the video shot. For key frame extraction, STPGF is proposed, which is an optimal representation of all the frames in video shot. After that BOVW is applied to extract the local descriptors (shape, color and texture) on the selected key frames.

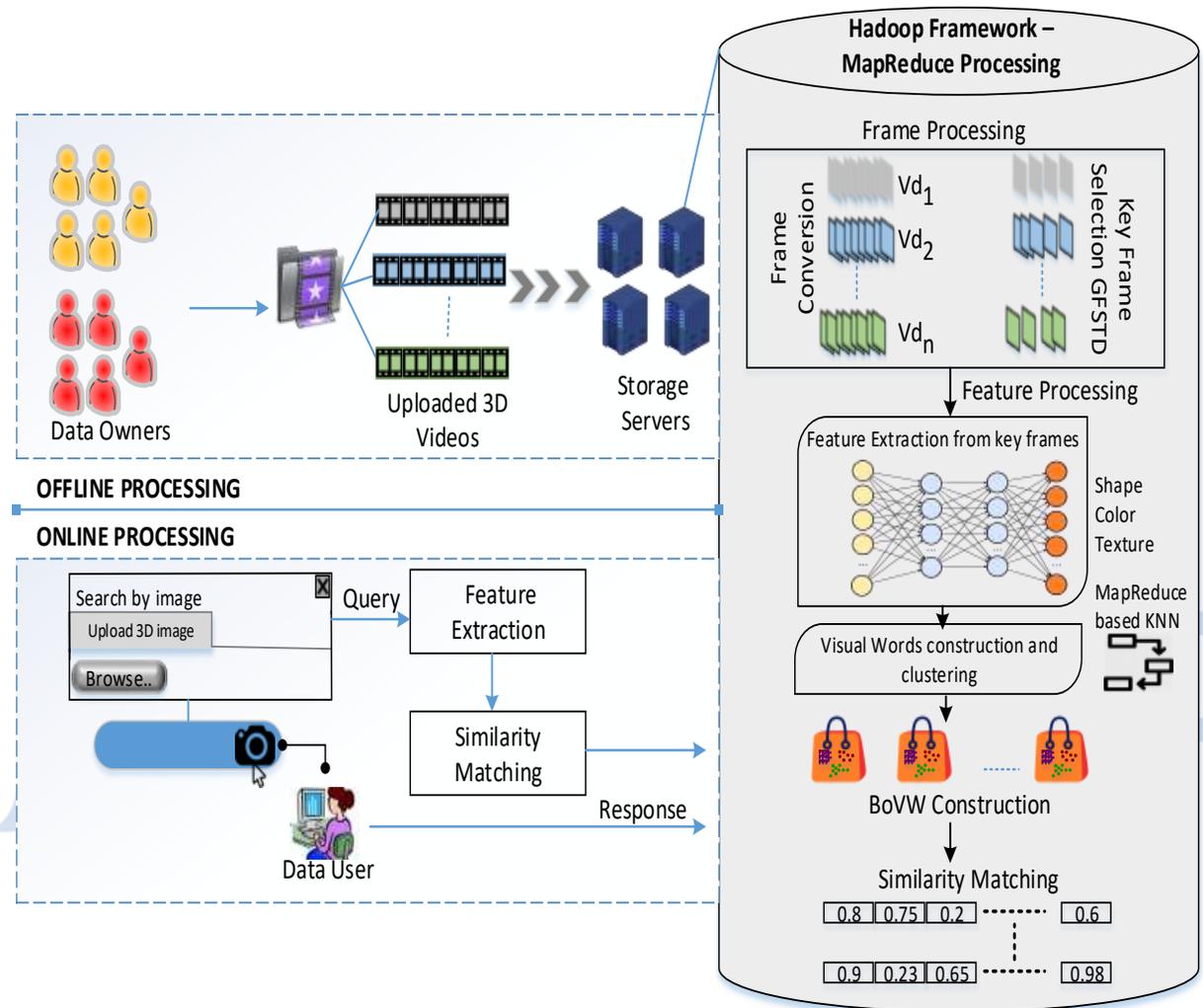


Figure 4.1 System Architecture

Then, geometric and topological features are integrated for shape descriptor and then Hybrid 3D CNN is applied to extract the color and texture descriptors. A new video

retrieval system is designed to increase the positive results from real-time large datasets. This system focuses more on

- Structure and Semantic retrieval
- High Dimensional datasets
- Avoid mismatches in 3D-CBVR

Next the construction of visual vocabulary is performed according to the meaning of words obtained from the key frames. To the resultant, Threshold Based-Predictive Clustering Tree (CT) is applied for visual codebook generation. Then, count the number of occurrence of visual word from the vocabulary and compute the occurrence value. As a final step to the above process, histogram is generated, which act as a graphical representation of the tonal distribution. On the user side, user provides a query image in the online mode and BOVW is computed. With this completion of BOVW, the processed query reaches matching process in which similarity between query image and resultant dataset is obtained using soft weighting scheme with L_2 distance function. Finally, index value is used to rank the matched image and is returned to the user. Further this proposed 3D CBVR process is detailed in following subsections.

4.4.2 Key Frames Extraction

Key frames are constructed from a set of relevant frames selected in the video shots i.e. obtained from original video. Key frame selection plays an initial role in video retrieval applications as it represents the whole video content. Different state-of-the-art

methods were discussed for key frame extraction and their performances were evaluated. However, previous approaches overlook some temporal information. To tackle this problem, we have suggested optimal key frame selection procedure by Spatial-Temporal Prioritized Gaussian Filter (STPGF) in our proposed work, which carries both spatial and global information about the frames in video shots for achieving high retrieval performance. In first step, the given input video is converted into frames using frame conversion tool. Further, key frames are extracted from the frames using STPGF to corresponding Map task. Then, the output key frames are stored in HDFS.

is constructed based on the probability of occurrence of pixel value at each position for all the frames. The outcome from the STPGF is the maximum occurred frames which have multiple entities. From the Group of Frames (GoF), histogram is formed based on the pixel values at each corresponding pixel position. Gaussian function is applied to smooth the histogram.

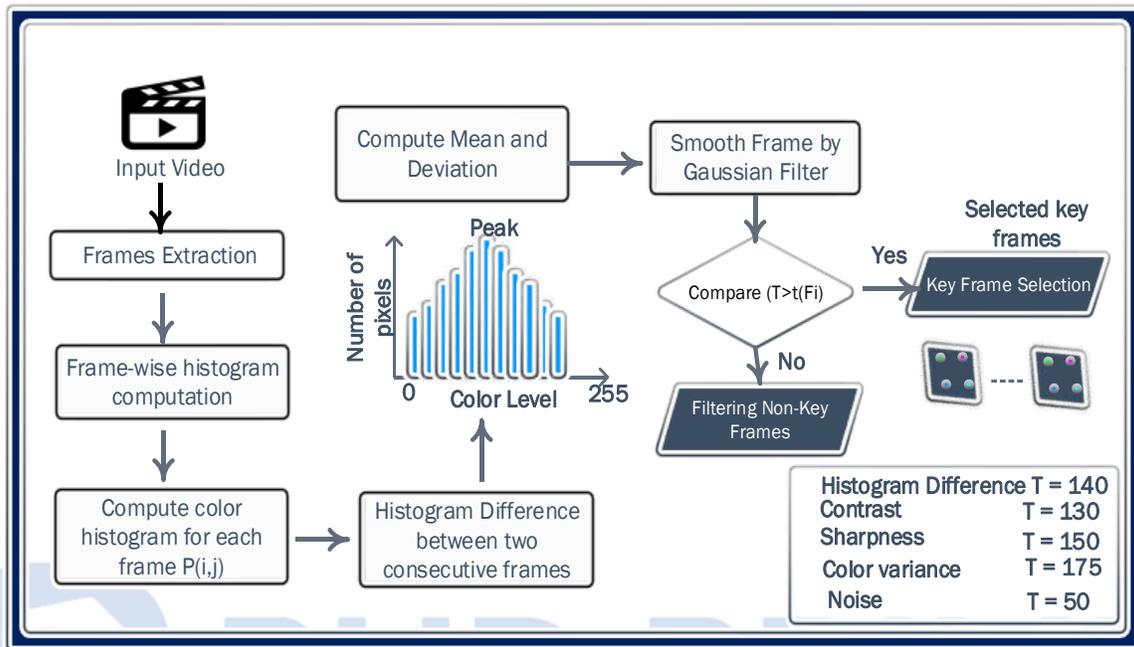


Figure 4.2 Key Frames Extraction using STPGF

The following equation represents the computation process of pixels of the histograms

$$TMOF(a, b, c) = b_{opt}, \quad 0 \leq a \leq I' - 1 \text{ and } 0 \leq b \leq J' - 1 \text{ and } 0 \leq c \leq K' - 1 \quad (4.2)$$

Where $I' * J' * K'$ is the size of a STPGF, I, J, K represents the Width, Height, Depth and b_{opt} is chosen as follows,

$$b_{opt} = \arg \max \{K'_{a,b,c}(b)\}, \text{ for } 0 \leq b \leq B \quad (4.2)$$

By using Gaussian filter, smoothed histogram is determined as follows,

$$K'_{a,b,c}(b) = K_{a,b,c}(b) * G(\sigma, b) \quad (4.3)$$

Where $G(\sigma, b)$ is a Gaussian function with variance σ

$$K_{a,b,c}(b) = \sum_{n=0}^{N-1} \delta(f_n(a, b, c) - b), \quad \text{for } 0 \leq b \leq B$$

$$(m, n) = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{for } m \neq n \end{cases} \quad (4.4)$$

$K_{a,b,c}$ represents the histogram formed by the corresponding pixel at the position of pixels in (a, b, c) , $f_n(a, b, c)$ which represents the pixel level at coordinates (a, b, c) in frame n , total number of frames in GoF is represented as N and number of bins in the histogram is B . Generally, the intensity level of the pixel is equal to the number of bins in a histogram. Final values are the selected key frames which are represented as *MAX Frame* (a, b, c) . 3D video are uploaded as Vd_1, Vd_2, \dots . Videos converted into 'n' number of frames in accordance to video size.

STPGF Functions

- Find pixel value in between two different frames
- Smoothened histogram is determined
- Gaussian filter is used to improve the image quality (reduce contrast, sharp blur edges and remove noise)

Pseudocodes for STPGF

Input: Video Vd

Output: $\{Kf_1, Kf_2, Kf_3 \dots, Kf_n\}$

1. Begin
2. Convert Vd into $\{F_1, F_2, F_3, \dots, F_N\}$
3. For each F_N do
4. Compute HD, C, S, CV, and N
5. Compute μ and σ // Mean and Standard deviation
6. Perform Smoothing
7. If $(T > t(F_N))$
 - {
 - Select key frame
 - else
 - goto next frame
 - }
8. Return $\{Kf_1, Kf_2, Kf_3 \dots, Kf_n\}$
9. End

4.4.3 Bag of Visual Words (BOVW) Generation

Bag of Visual Word is an extension of Bag of Words model from the field of text classification. BOVW provides 3D model by measuring the occurrences of its projected views and quantizes each projected view into different descriptors. It is an efficient, potent and scalable approach for performing CBVR. Generally, BOVW is used to extract set of visual words from an enormous visual vocabulary. The BOVW model represents a sentence by histogram, which is constructed by counting the occurrences of words in sentences. The Basic steps of BOVW are represented as follows,

- Feature Extraction
- Construction of visual vocabulary
- Quantization of each image features as discrete visual words
- Construction of inverted-index using visual words and based on it matching performance is executed.

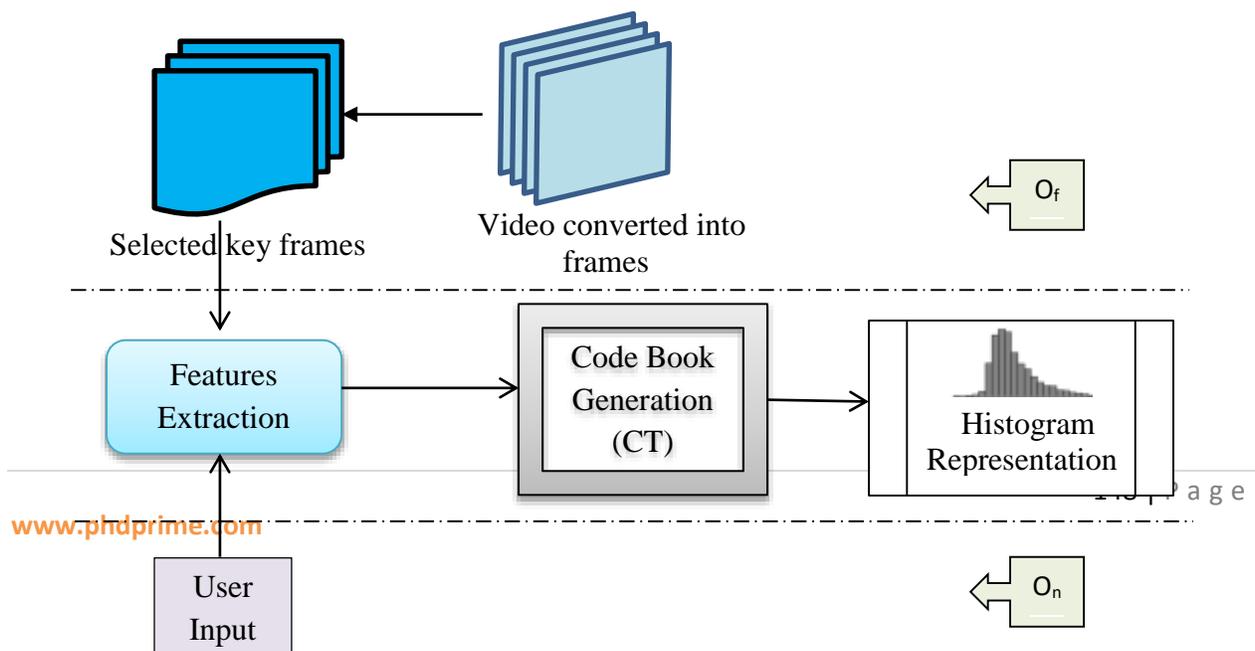


Figure 4.3 BOVW Procedure

The overall architecture of BOVW is presented in Figure 4.3, which envisions the sequence of work performed by BOVW. Commonly, we use anyone local descriptor to extract image feature in 3D model. But, it is the first time we take in account of three different image features (shape, color & texture) for feature extraction. In particular, we have considered combined feature (topological and geometric) for shape. Map Reduce is used as a background for BOVW in implementation. The following subsection covers the above mentioned four steps.

1) Feature Extraction

Initial stage of BOVW is feature extraction from the largest dataset. For this motive, several local descriptors were used in the conventional system ^{[62], [70], [71]}. In most of the previous work, SIFT features are used a local descriptor. Feature extraction processes are described in sub chapters. Feature extraction is one of the most important processes for most multimedia data retrieval tasks. For feature extraction, we consider three different features of the image. Feature extraction proceeds with map reduce in which each task is handled in parallel manner. So it makes feature extraction much simpler as one image does not depend on another and we can easily break down the tasks. Therefore, individual image is an input to the map task and extracted local feature of that image is the output. Map and Reduce functions are stated as follows,

$$Map(fl^{(n)}, I^n) \rightarrow [\langle fl^{(n)}, ft^{1\dots f^{(I^n)}} \rangle] \quad (4.13)$$

$$Reducer_{null}(fl^{(n)}, ft^{1\dots f^{(I^n)}}) \rightarrow [\langle fl^{(n)}, ft^{1\dots f^{(I^n)}} \rangle] \quad (4.14)$$

Where $fl^{(n)}$ is the key which represents the image filename or unique identifier, I^n is n^{th} image of N images in the dataset. Key value assigned to the map tasks are unique identifiers. Each map tasks handles one image pair (i.e.) $\langle fl^{(n)}, I^n \rangle$. Single map task will emit all local features of $(F(I^n))$ as $ft^{1\dots f^{(I^n)}}$ of I^n image. To note, here the reducer is null reducer as the resultant output $\langle key, value \rangle$ pair is same as the input given to mapper.

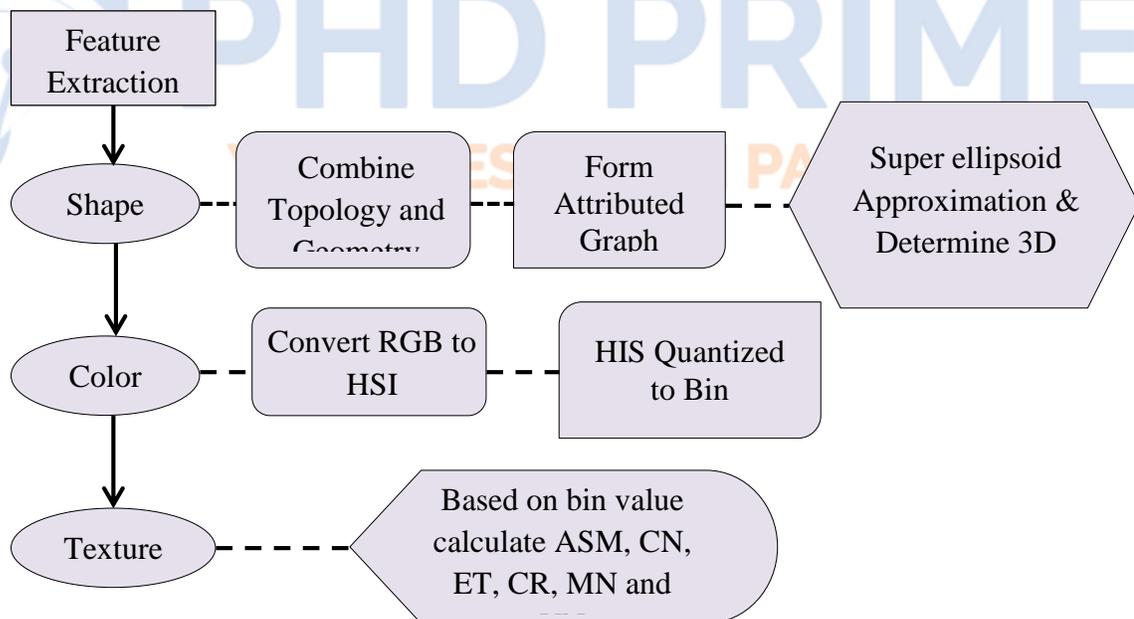


Figure 4.4 Flow of Feature Extraction Process

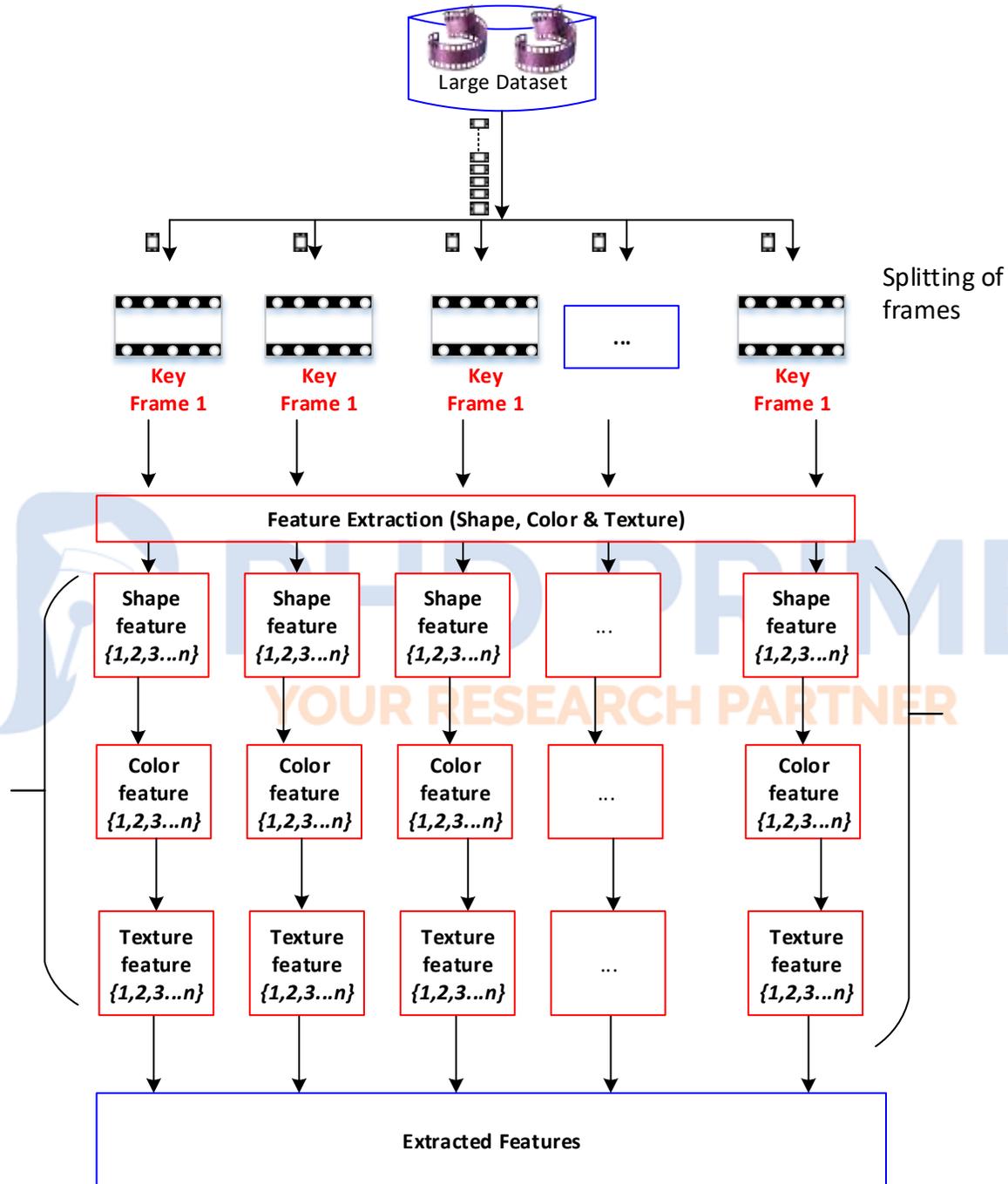


Figure 4.5 Overall Feature Extraction

In this manner, we reduce the additional execution of reducer. Hence the output from mapper is given as input to the pipeline stage. As per this work, shape feature is extracted first, followed by the other two features i.e. color and texture. The resultant feature extraction is a combination of all above mentioned features. Figure 4.5 illustrates the feature extraction process along with shape, color and texture features. Initially; the shape features are extracted with geometric and topological features. Then, both color and texture features are extracted using super ellipsoid 3D object of an image. 3D co-occurrence matrix is applied to extract the six textual features.

Shape is considered as one of the prime feature of an image. In general, 3D model consider features like geometric and topological for shape features. Previous works used anyone of the feature or combination of two for features during feature extraction [45], [68]. But previous work has many problems which were detailed in related work. In order to overcome this, we have defined some rules to combine topological and geometrical feature.

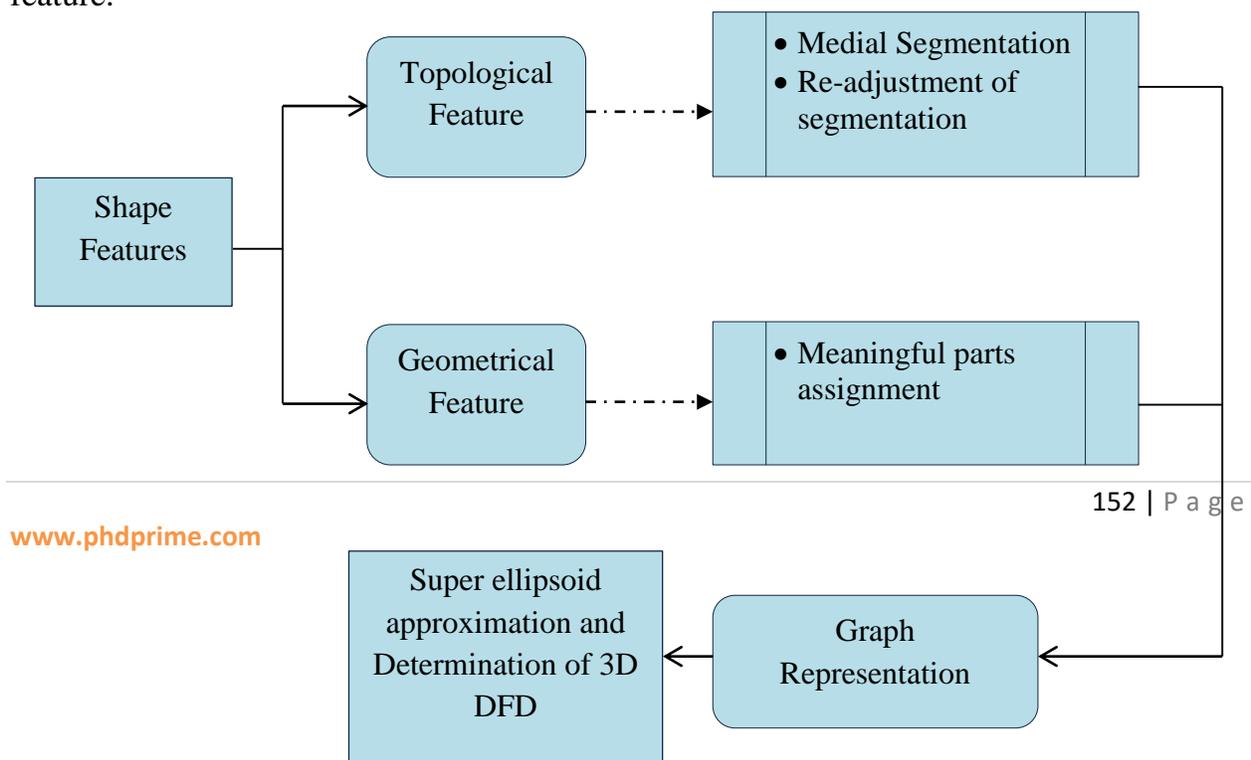


Figure 4.6 Shape Feature Extractions with Topological and Geometrical Feature

Figure 4.6 represents the shape feature extraction process along with geometric and topological features. The overall steps of shape feature extraction involve:

- Medical surface extraction and segmentation
- Re-adjustment of segmentation
- Use super ellipsoid to approximate the segmentation
- Use 3D DFD (Distance Field Descriptor)

To perform Medial Surface Extraction (M_S), we have used an appropriate Fork Strategy based Medial Surface Extraction. M_S is fast and robust to surface noise, which is applicable for objects represented in terms of topological function. We start our process with ‘ M_S ’, Medial surface extraction which is performed using Voronoi diagram method. M_S is performed by topology thinning which uses the concept of Average Outward Flux that estimates Euclidean distance as its parameter but it does not guarantee the overall geometry.



Figure 4.7 Voronoi Representation for Object

In this proposed strategy, images are partitioned into cells, edges and vertices based on the points present on it as shown in figure 4.7. Endpoints of voronoi edges are called voronoi vertices. Each point is present in equidistant from at least two edges. After forming voronoi diagram we scan 3D image in 6 directions and start to eliminate boundary region. Voxels (regular grid in 3D Shape) are eliminated one by one until the voxels have one neighboring voxel.

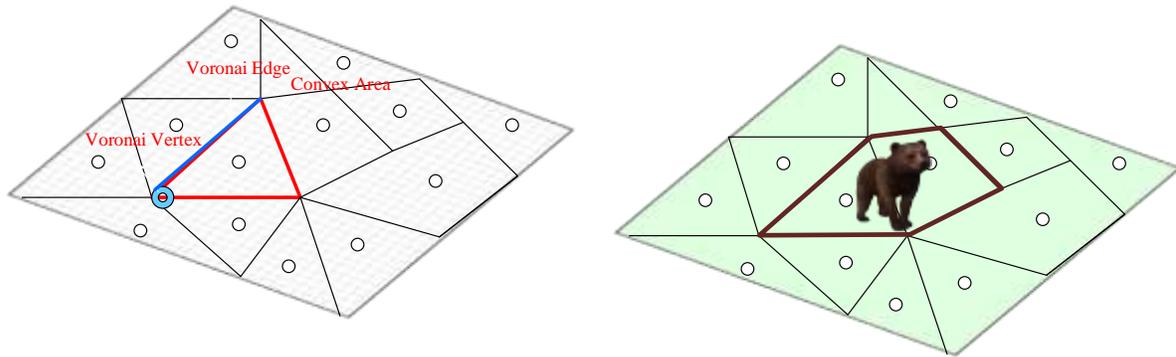


Figure 4.8 Voronai Diagram

Considering ‘n’ points, every cell is the intersection of (n-1) half-planes which are convex in shape and have at least (n-1) edges in the boundary. To split ‘n’ points in the plane, the required time is computed as,

$$T = O(n^2) \quad (4.15)$$

$$\text{Max Frame } (a, b, c) \quad (4.16)$$

Further, M_s segmentation process is involved with the use of several classification methods, which classifies every voxel on median surface. The classification methods of M_s are based on the following classification categories: Simple (voxel is an end point of line), Surface Voxel (belongs to surface segment), Line voxel (voxel belongs to line segment) and junction (voxel is a junction between two lines). The classification process is performed with respect to connectivity of neighboring medial surface and background

voxels. After the completion of classification, rules are applied to perform segmentation in M_S .

- All neighboring line voxels along with the neighboring simple voxels form a line segment
- All neighboring surface voxels along with neighboring simple voxels form a surface segment

By using few criterions, re-adjustment is performed in segmentation to overcome the problems in traditional systems. The criterions are correction criterion, elimination criterion, over segmentation and merging criterion. Now the M_S is segmented into meaningful parts, while its topology is preserved. After successful completion of re-adjustment process, following definitions are used to allocate the voxel of object's boundary surface to segmentation process.

- Set of medial surface segment is $S = \{U_{i=1}^n S_i\}$, where n is the number of medial surface segments
- Set of medial surface voxels is $S_i = \{x_i[k] \mid k = 1 \dots N_i\}$. that belong to the segment S_i , and $x_i[k]$ are coordinates of the k medial surface voxels center assigned to 'i' part. N_i is the number of the medial surface voxels assigned to the 'i' segment

- Set of boundary voxels is $P = \{p[l] \mid l = 1 \dots L\}$ where $p[l]$ are the coordinates of the 'l' boundary voxels center, $P \subset U_A$ while U_A is the set of voxels lying inside the object 'A' and 'L' is the number of objects boundary voxels.
- Euclidean distance of the 1st boundary voxel from the kth medial surface voxel of the segment S_i is $d(p[l], xi[k])$
- $Di[l] = \operatorname{argmin}_{1 \leq k \leq \{d(p[l], xi[k])\}}$ is the minimum Euclidean distance of the 1st boundary voxel from the segment S_i .
- $Pi = \{p_i[l] \mid l = 1 \dots L_i\}$ is a subset of P, which represents the set of boundary voxels assigned to segment S_i and $p_i[l]$ are the coordinates of the 'l' boundary voxels center assigned to 'i' part. L_i is the number of boundary voxels assigned to 'i' part.

A boundary voxel $p[l]$ is assigned to the segment S_i ,

$$P[l] \in p_i \Leftrightarrow W_i \cdot D_i[l] = \operatorname{arg\,min}_{1 \leq i \leq n} \{D_i(l)\} \quad (4.17)$$

Where W_i is a weight factor for given by the following equation:

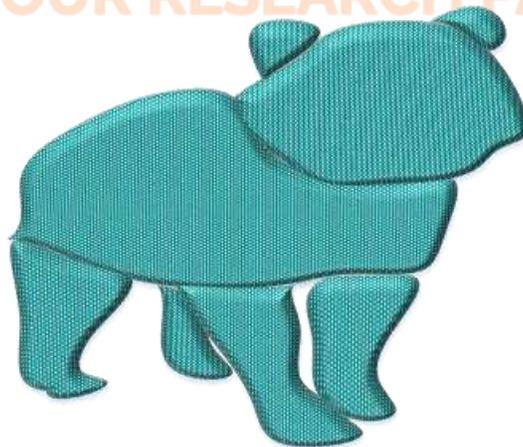
$$W_i = 1 + \frac{\sigma_i - \operatorname{arg\,min}_{1 \leq i \leq n} \sigma_i}{\operatorname{arg\,max}_{1 \leq i \leq n} \sigma_i - \operatorname{arg\,min}_{1 \leq i \leq n} \sigma_i} \quad (4.18)$$

Let σ_i be the standard deviation of $D_i[l]$ for all boundary voxels $p_i[l]$ assigned to the segments S_i . Each and every extracted segment of the 3D- objects is approximated using

super ellipsoids to extract the shape without any loss. By using the attributed graph representation, both topology and geometrical information are combined.



(a) PHD PRIME (b)
YOUR RESEARCH PARTNER



(c)

Figure 4.8 (a) Original image (b) Medial surface extraction and segmentation (c) 3D Segmentation

Figure 4.8 represents the medial surface extraction and re-adjustment segmentation of the original 3D-objects. The result obtained from medical surface extraction and readjustment procedures are considered as topological information whereas the meaningful parts are assigned to each surface segment that are well thought out to be geometrical information. The graph is represented as,

$$G = \{V, E, A\{B_i\}_{i=1}^s\} \quad (4.19)$$

Where 'V' is the non-empty set of vertices, 'E' is the set of edges and 'A' is the binary symmetric adjacency matrix. According to the mapping definition, an undirected vertex-attributed graph is constructed. From equation below, the parameters specified within the function 'F' is (x, y, z) which represents the co-ordinates of the 3D-point

$$F(x, y, z) = \left(\left(\left(\frac{x}{a_1} \right)^{\frac{2}{\varepsilon_2}} + \left(\frac{y}{a_2} \right)^{\frac{2}{\varepsilon_2}} \right)^{\frac{\varepsilon_2}{\varepsilon_1}} + \left(\frac{z}{a_3} \right)^{\frac{2}{\varepsilon_1}} \right)^{\varepsilon_1}$$

$$= \left(\frac{xa_2a_3 + ya_1a_3 + za_1a_2}{a_1a_2a_3} \right)^2 = 1 \quad (4.20)$$

Function (4.11) is commonly called inside-outside function, since it a 3-D point with coordinates (x, y, z) :

$$\begin{cases} F(x, y, z) > 1, & \text{if } (x, y, z) \text{ lies outside} \\ & \text{the surface of the object} \\ F(x, y, z) \leq 1, & \text{if } (x, y, z) \text{ lies inside or on} \\ & \text{the surface of the object} \end{cases} \quad (4.21)$$

The issue of modeling 3-D object using super quadratic, which is prevailed over by reducing it to the least squares minimization of nonlinear inside-outside function $F(x, y, z)$ with respect to several shape parameters,

$$F(x, y, z) = F(x, y, z; a_1, a_2, a_3, \varepsilon_1, \varepsilon_2, \varepsilon_3, \Phi, \theta, \chi, t_x, t_y, t_z) \quad (4.22)$$

Where t_x, t_y, t_z and Φ, θ, χ are Euler angles and translation vector coefficients, $a_1, a_2, a_3, \varepsilon_1, \varepsilon_2, \varepsilon_3$ are the super quadratic shape parameters and (x, y, z) are the coordinates of 3D objects. Using mean-square error, the above found parameters are minimized as,

$$MSE = \sum_{i=1}^N \sqrt{a_1 a_2 a_3} (F(x_i, y_i, z_i) - 1)^2 \quad (4.23)$$

Where 'N' is the number of points of the 3D object

The super ellipsoid approximation provides an acceptable representation of object's meaningful parts. Here, the shape information of each part is reduced to five attributes. These attributes are the parameters of super-ellipsoid. As mentioned in over segmentation criteria, all the parts of the image will be merged into a single part, then the number of resulted parts is comparable to the number of medial surface voxels. In case of merging, only geometry feature is considered, by ignoring the topology information. Further, 3D DFD is used to compute the differences between the surface of an ellipsoid and surface of

an object. Using the following scaling procedure, the 3D DFD is extracted from every segment of 3D object,

$$d_{DF}^s = \frac{1}{\sum_i |d_{DF}(i)|} d_{DF} \quad (4.24)$$

$$d_{DF} = [||F_{DF}(0,0)|| ||F_{DF}(0,1)|| ||F_{DF}(1,0)|| ||F_{DF}(0,2)|| ||F_{DF}(1,1)|| ||F_{DF}(0,2)|| \dots] \quad (4.25)$$

$$F_{DF} = FT\{DF\} \quad (4.26)$$

$$DF = [d_{i,j}] \quad (4.27)$$

Where $d_{i,j}$ is the signed distance between the surface of an object and surface of the ellipsoid at the same (θ_i, Φ_i) . Equation for 3D Distance field Description is acquired as a result of shape based feature extraction,

$$(d_{DF}^s) = \frac{1}{\sum_i |d_{DF}(i)|} d_{DF} \quad (4.28)$$

The above mentioned equation reports about 3D Distance Field Descriptor (DFD) which is obtained from maximum occurred frames. This value is further applied for color and feature extraction process. On completion of extracting the shape features, remaining two other features are extracted. The color and texture are also significant attributes in image analysis. Several techniques were studied for color and texture feature extraction based on color histogram. Although, such techniques have been extensively used in certain applications, they have some disadvantages since spatial information is not incorporated into histogram. To tackle this problem, color and texture features are extraction using the approximated super ellipsoid 3D object of an image. In reference of a

previous article, this work uses 3D co-occurrence matrix as feature descriptor for both color and texture.

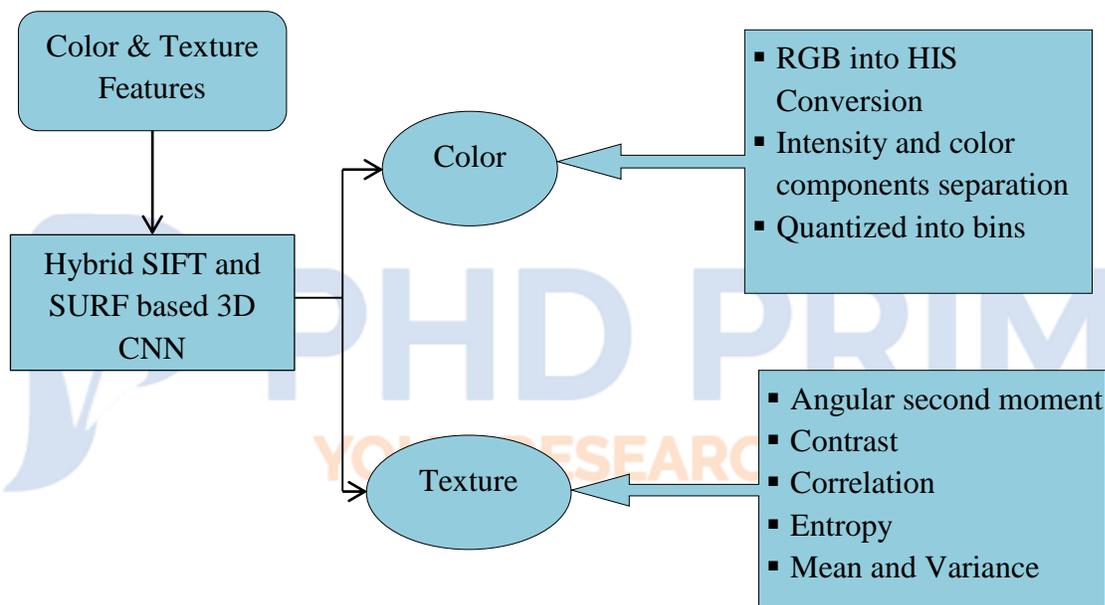


Figure 4.9 Color and Texture Features Extraction

Figure 4.9 illustrates both color and texture features based on 3D co-occurrence matrix. As an initial step, RGB color components are converted into HIS color component, since it separates intensity and color components. Further, H, S, and I components are quantized into 8, 4, and 4 bins respectively. Next, the image is split into

squared window. Our concept includes overlap regions as an added advantage. Nine orientations are employed in each window to define the neighborhood of each pixel along the HIS plane. Table 4.1 shows the list of shape features extracted in the 3D Shape Representation Model.

Table 4.1 List of Shape Features

Features	Formulas
Area - It is computed by the number of pixels in largest axial slice lar_{as} multiplied by the resolution of pixels	$I(x, y) \times \Delta A$
Aspect Ratio - It is computed by the width and length of the frame	L/W
Roundness - It is computed by the similarity of the scene region to the circular shape	$\frac{4\pi A}{L}$

<p>Perimeter - It is defined as the structural property for the list of co-ordinates and also the sum of distance from each coordinate</p>	$\sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2}$
<p>Circularity - It is computed by the lar_{as} of each scene or frame region</p>	$\frac{4\pi A}{Q^2}$ <p>ΔA is the area of one pixel in the shape of $I(x,y)$, X_i and Y_i is the ith pixel coordinates, A is the object area, and L is the object region boundary length, $MA(x,y)[L]$ is the major axis length and $MI(x,y)[L]$ is the minor axis length.</p>
<p>Uniformity - It is defined by the intensity uniformity for the Histograms</p>	$\sum_{i=0}^{l-1} H^2(R_i)$
<p>Mean - It is represented by the average intensity measure</p>	$\sum_{i=0}^{l-1} R_i \times H(R_i)$
<p>Standard Variance - It is represented by the 2nd moment of the average values</p>	$\sum_{i=0}^{l-1} (R_i - M)^3 \times H(R_i)$

4.4.4 CNN for texture feature extraction

Our SURF and SIFT based algorithm extracts texture features by implementing the CNN descriptor. The reason for selecting CNN for texture feature extraction is that it provides better performance in extracting features from each image using different layers. Using this descriptor, our proposed method extracts six texture features from the given image. They are Contrast, Dissimilarity, Entropy, Homogeneity, Correlation and Angular Second Moment. The CNN descriptor comprises of three significant layer that are convolution layer, pooling layer and fully connected layer as depicted figure 4.10. The layers in the CNN descriptor are described as follows:

1) Convolutional Layer

The convolutional layer collects input from the input layer to extract features from the input image. The convolution layer contains multiple filters to extract high level features from the given image. The convolutional layer comprises set of learnable filters to produce the feature map. In this, six filters are used to create six feature maps from the input image. The feature map acquired from the individual filter is convolved through the whole image. Each feature map obtained from the filter signifies the precise features of the image. In this layer convolution operation is performed which combines the two different functions to generate a third function. The convolution operation is illustrated as follows:

$$x_j^l = a_f(\sum_{i \in M_l} x_j^{l-1} * f_{ij}^l + b_j^l) \quad (4.29)$$

Here, a_f designates the activation function, j signifies the specific convolution feature map, l exemplifies the layer in the CNN, f_{ij} represents the filter, b_j is referred as the feature map bias and M_l defined as the selection of feature map.



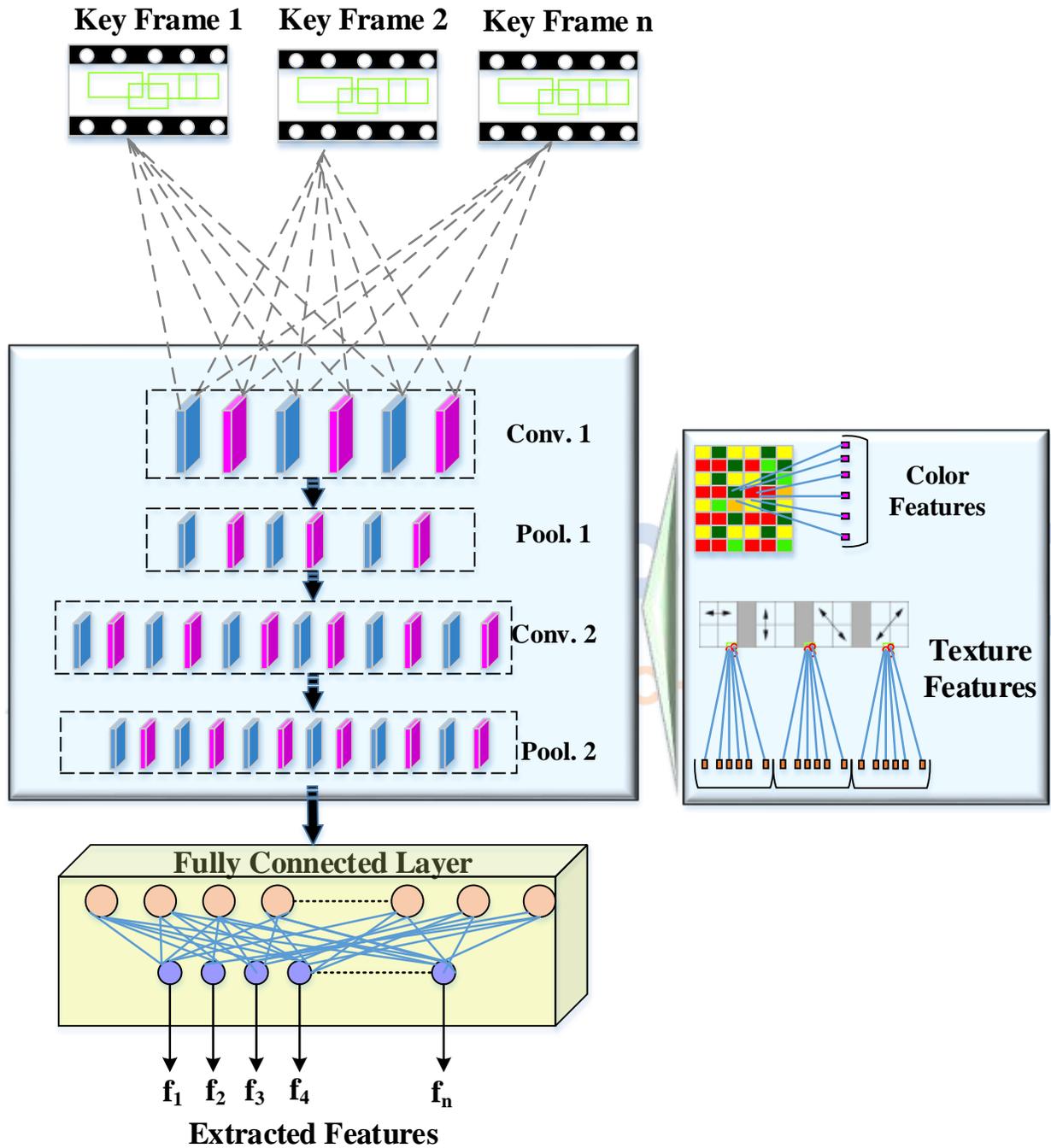


Figure 4.10 Feature Extraction in CNN

2) Pooling Layer

The pooling layer is utilized to accomplish down sampling operation in CNN algorithm. Herein, pooling operation is accomplished to diminish the spatial size of the representation in order to diminish the volume of parameters and computations in the network. It functions on each feature map individually. The pooling operation is expressed as follows:

$$p_j^l = a_f(C_j^l \text{pool}(p_j^{l-1}) + b_j^l) \quad (4.30)$$

Where p_j^l denotes the pooling region result applied on the j^{th} region in the input image, p_j^{l-1} refers the j^{th} region of interest taken by the pooling mask in preceding layer and C_j^l represents the trainable co-efficient.

3) Fully Connected Layer

The fully connected layer is used to extract features that are extracted from the previous layer. It provides the extracted features as output to the upcoming processes. This layer is final layer in the CNN based feature extraction that obtains results from the preceding layers in order to give extracted features.

3.4.2.1 SIFT descriptor

SIFT descriptor is one of the feature detection algorithm used in computer vision technology to detect and outline the local features in the image. It transforms the image information into the scale invariant coordinates. It incorporates the four successive phases

that are listed as follows:

- i. Scale Space Extrema Detection
- ii. Localization of key point
- iii. Orientation assignment
- iv. Key point descriptor

1) Scale Space Extrema Detection

The initial stage of the feature extraction is to recognize the locations and scale of the periocular region of the face image. In this, scale space is used to identify the locations that are invariant to the scale change of the image. The gaussian kernel is used to search all the scale and image locations effectually. The periocular region of the image $P(x, y, \sigma)$ is obtained by performing the convolution of the variable scale gaussian function with the input image.

$$P(x, y, \sigma) = \text{Gaussian}(P(x, y, \sigma)) * \text{Input}(P(x, y, \sigma)) \quad (4.31)$$

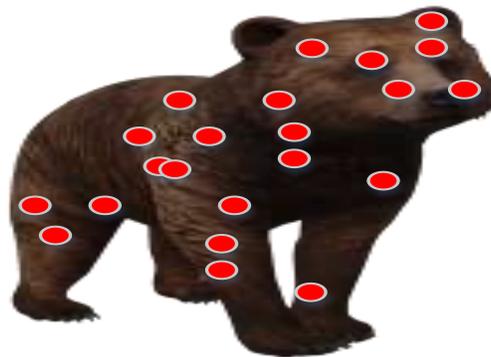


Figure 4.11 key point location using SIFT

Figure 4.11 illustrates the key point location identified using the SIFT feature descriptor in the periocular region of the face image. The gaussian function can be expressed as follows,

$$\mathbf{Gaussian}(P(x, y, \sigma)) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (4.32)$$

The difference of Gaussian (DoG) function is estimated which is used to determine the difference between the two adjacent scales of the image with unglued factor d.

$$\mathbf{DoG}(x, y, \sigma) = (\mathbf{Gaussian}(x, y, d\sigma) - \mathbf{Gaussian}(x, y, \sigma)) * \mathbf{Input}(x, y) \quad (4.33)$$

$$= P(x, y, d\sigma) - P(x, y, \sigma) \quad (4.34)$$

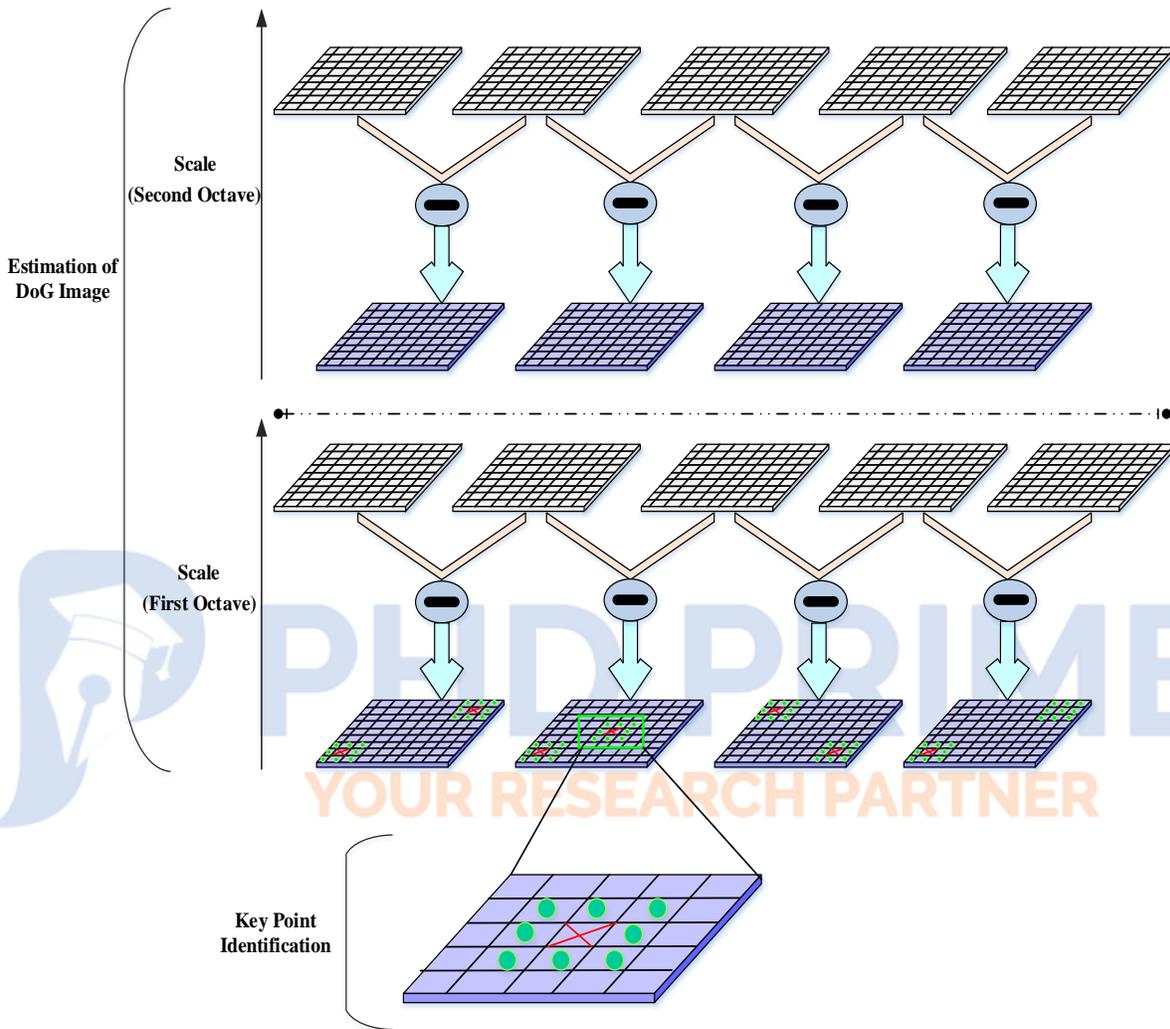


Figure 4.12 DoG Estimation and Key point identification

Figure 4.12 depicts the estimation of DoG image and key point identification in the SIFT descriptor for the given input image. Once the DoG of the image is obtained, then key points are identified as the local minima/maxima of the DoG images across different scales. This is accomplished by comparing each pixel in the DoG images to its

eight neighbor pixels at the same scale and nine neighbors in the above and nine neighbors in the below scales. If the pixel value is maximum or minimum among all compared pixels then it is selected as a candidate key point.

2) Localization of Key Point

If a candidate keypoint is found by associating a pixel to its neighbors, then it is used to calculate the location and scale data in the pericocular region of the image. The key point candidates which have low contrast or poorly constrained along edges are avoided by this localization of the key point. This process is significant in feature extraction since key points detected from the scale space extrema contain more unstable points. The key points are selected based on the measurement of their stability and interpolation among nearby data in order to find an accurate position of the key point. This is achieved using the quadratic Taylor expansion of the DoG scale space performance.

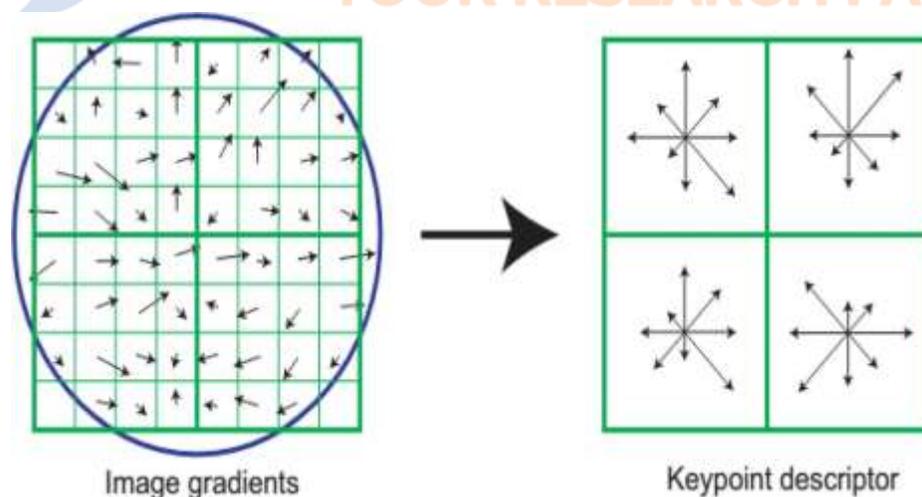


Figure 4.13 key Point descriptor

Figure 4.13 describes the performance of the key point descriptor in SIFT with image gradients. The Taylor expansion is expressed by,

$$DoG(k) = DoG + \frac{\partial(DoG)^T}{\partial k} k + \frac{1}{2} k^T \frac{\partial^2 DoG}{\partial k^2} k \quad (4.35)$$

Where, k is the offset from the sample key point that can be defined as $k = (x, y, \sigma)^T$. The location of the offset extremum \hat{k} is expressed as,

$$\hat{k} = -\frac{\partial^2(DoG)^{-1}}{\partial k^2} \frac{\partial DoG}{\partial k} \quad (4.36)$$

The function value of the extremum \hat{k} is valuable for refusing unstable extrema with low contrast. This can be acquired through substituting equation (4.35) into (4.36) that provides,

$$DoG(\hat{k}) = DoG + \frac{1}{2} \frac{\partial(DoG)^T}{\partial k} \quad (4.37)$$

If $DoG(\hat{k})$ value is less than the 0.3, then maxima or minima extrema will be prohibited.

Figure 4.14 depicts the estimation of the integral images using the SIFT feature descriptor. The second enhancement is to remove some unstable points that provide a high response as they are positioned at the boundaries. This problem is overwhelmed by the Hessian matrix (Hm) by checking the ratio whenever it is less than the precise threshold,

$$\frac{Trace(Hm)}{Deter(Hm)} < \frac{(\gamma+1)^2}{\gamma} \quad (4.38)$$

Where $Trace(Hm)$ and $Deter(Hm)$ are categorized as the sum of the Eigenvalues obtained from the trace of $Hm(2 * 2 matrix)$ and their product from the determinant. γ is referred as the ratio of the largest magnitude eigenvalue with the smaller one.

3) Orientation Assignment

In this phase, each key point is allotted with one or more orientations based on the local image gradient directions. This is a significant step to attain invariance to rotation as the key point descriptor that can be represented comparative to this orientation hence acquire invariance to the image rotations. For an image sample $I(x, y)$ at scale σ , the gradient magnitude $Ma(x, y)$ and orientation $\theta(x, y)$ is expressed as follows:

$$Ma(x, y) = \sqrt{(P(x+1, y) - P(x-1, y))^2 + (P(x, y+1) - P(x, y-1))^2} \quad (4.39)$$

$$\theta(x, y) = \tan^{-1} \frac{P(x, y+1) - P(x, y-1)}{P(x+1, y) - P(x-1, y)} \quad (4.40)$$

4) Key Point Descriptor

A key point descriptor is created by estimating the gradient magnitude and orientation at each image sample point in a region near the keypoint location. These keypoint locations are weighted by means of a gaussian window and selected using the overlaid circle. The sample acquired is then combined into the orientation histograms succinct the contents over $4*4$ sub regions. The gaussian and box filters are performed in two different directions that are described as the X direction and XY direction in order to describe the key points present in the periorcular region of the facial image.

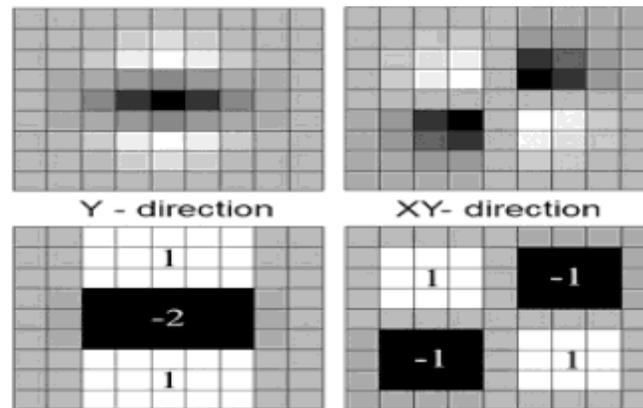


Figure 4.14 Second order Gaussian and Box Filter

Figure 4.14 demonstrates the second order gaussian and box filters used in the SIFT descriptor. The SIFT descriptor uses the image gradients instead of the intensity values owing to the directions are extremely invariant to variations with respect to the brightness and contrast.

3.4.2.2 SURF descriptor

SURF descriptor is one of the feature descriptors in computer vision technology which is partially inspired by the SIFT descriptor. It centers on scale space theory and popular for its computing speed. It is a noteworthy algorithm on the base of multi scale space theory and it is robust to the variations such as scale, illumination, rotation, etc... It encompasses four sequential steps:

1. Generation of Integral Image

2. Detection of a key point
3. Orientation assignment
4. Generation of descriptor

The principles and steps of the SURF algorithm are similar to that of the SIFT algorithm.

1) Generation of Integral Image

SURF utilizes square shaped filters as an approximation of gaussian smoothing whereas SIFTS uses cascaded filters in order to find the scale invariant points. The summation of all the pixels exist in the input image at location $\mathbf{l} = (x, y)$ in a rectangular region is known as integral images that can be expressed as,

$$In_{im} \Sigma(x, y) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} Ip_{im}(i, j) \quad (4.41)$$

Where In_{im} represents the integral image and Ip_{im} represents the input image.

2) Detection of Key Point

SURF algorithm utilizes the blob detector based on the determinants of hessian matrices to find the interest points in the given image. The determinant of the hessian matrix is used to measure the local variations around the point and the points are selected where this determinant is high. The hessian matrix is determined as follows,

$$Hm(x, \sigma) = \begin{pmatrix} P_{xx}(s, \sigma) & P_{xy}(s, \sigma) \\ P_{xy}(s, \sigma) & P_{yy}(s, \sigma) \end{pmatrix} \quad (4.42)$$

Where, $P(\mathbf{s}, \sigma)$ indicates the convolution of the gaussian second order derivative for the given input image in point \mathbf{s} . A box type filter approximation is used in the SURF algorithm instead of the gaussian filter in order to enhance calculating speed. The Hessian matrix associated with gaussian kernel in the box type filter is shown as below,

$$Hm = \begin{bmatrix} DoG_{xx} & DoG_{xy} \\ DoG_{xy} & DoG_{yy} \end{bmatrix} \quad (4.43)$$

The determinant of the hessian matrix at different scales in the image is indicated by:

$$Det(Hm)_{approx.} = DoG_{xx}DoG_{yy} - (\omega DoG_{xy})^2 \quad (4.44)$$

Where ω indicates the weight function that is used to sustain energy within the gaussian kernel. The value of $Det(Hm)_{approx.}$ is constant when the minimum or maximum value is reached. In accord to obtain the extreme point, the following equation is obligatory:

$$Hm(\mathbf{k}) = Hm + \frac{\partial Hm^T}{\partial k} \mathbf{k} + \frac{1}{2} \mathbf{k}^T \frac{\partial^2 Hm}{\partial k^2} \mathbf{k} \quad (4.45)$$

Where, $\mathbf{k} = \mathbf{x}, \mathbf{y}, \sigma$. The position of the local extreme point is computed by assigning the derivation of equation 3.5 as zero. The position of the extreme point is denoted as below,

$$\hat{\mathbf{k}} = \frac{\partial^2 Hm^{-1}}{\partial k^2} \frac{\partial Hm}{\partial k} \quad (4.46)$$

3) Orientation Assignment

The dominant direction is assigned to each interest point as rotational invariance. Here, Haar wavelet responses or filters are employed to find the gradient of the key points in the either vertical or horizontal direction. In this, three steps are exploited to generate the key point candidate as the rotation invariant.

- A Haar wavelet response is used to measure the gradient of each key point in the scale space. These points should be gratified with the arena where the circle is dignified with the radius 6σ nearby the interesting point that has been deliberated as the center point.
- The six vectors are obtained through rotating the window of 60 degrees around the center point of the circle. The reaction to the Harr wavelet is calculated in every sector and the solutions are collected. After that, six specific direction vectors are acquired.
- The direction which has maximum summation is desired as the key direction of the feature point.

4) Generation of Descriptor

In this step, the square or box filter is revolved into the main direction that is performed after selecting the neighborhood's points. The square region is divided into different sub regions. In each sub region, Harr wavelet response is used for frequently spaced key points. Therefore, each sub region contains a four dimensional vector and can be generated as follows;

$$v = (\sum d_{ho}, \sum d_{ve}, \sum |d_{ho}|, \sum |d_{ve}|) \quad (4.47)$$

Where d_{ho} represents the horizontal direction in Harr wavelet response and d_{ve} indicates the vertical direction in Harr wavelet response. The output of the SURF algorithm for the image in scale space generates a group of interest point locations instead of describing the regions around the interest points.

The key benefits of the proposed SIFT and SURF feature extractions are robust to the scale and rotation of the given input image. Therefore, our proposed method extracts better feature points from the facial image that enhances the accuracy in face recognition across aging variations. **The texture features are calculated using following features,**

- Angular Second Moment - Local intensity variation of an image and it will favor contributions from $P(i, j)$ away from the diagonal (i.e.) $i \neq j$. Contrast is calculated using the following equation

$$ASM = \sum_i^U \sum_j^V P^2[i, j] \quad (4.48)$$

- Entropy (ET) - This parameter calculates the randomness of gray-level distribution

$$Entropy = \sum_i^U \sum_j^V P[i, j] \log P[i, j] \quad (4.49)$$

- Correlation (CR) - It provides the correlation between two pixels present in the pixel pair

$$Correlation = \sum_i^U \sum_j^V \frac{(i-\mu)(j-\mu)P[i, j]}{\sigma^2} \quad (4.50)$$

- Mean (MN) – This parameter is used to calculate the mean of gray level in an image. It can be manipulated as,

$$Mean = \frac{1}{2} \sum_i^U \sum_j^V (iP[i, j] + jP[i, j]) \quad (4.51)$$

- Variance (VN) - Variance is used to explain the overall distribution of gray level. The following equation is used to calculate the variance,

$$Variance = \frac{1}{2} \sum_i^U \sum_j^V ((i - \mu)^2 p[i, j] + (j - \mu)^2 P[i, j]) \quad (4.52)$$

The above mentioned equations are applied to nine texture matrices and resulted as a single entity. The extracted features are the combination of all the features mentioned in this work, which is given as

$$Etd_f = [ASM, CN, ET, CR, MN, VN] \quad (4.53)$$

Equation (4.53) gives the feature extracted (Etd_f) from all the three features taken into account (Shape, Color, texture).

Color Feature Extraction

It refers to the intensity distribution of the scene. Color moments can be used to describe the intensity distribution. At first, color space conversion is performed

$$V = \frac{1}{3} (R + G + B) \quad (4.54)$$

$$S = V - \frac{3}{R+G+B} [\min(R, G, B)] \quad (4.55)$$

$$H = \begin{cases} \theta, & B \leq G \\ 360^\circ - \theta, & B > G \end{cases} \quad \text{Where, } \theta = \arccos\left(\frac{[(R-G)+(R-B)]/2}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right) \quad (4.56)$$

Further values of HSV are estimated,

$$h^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} h_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq h^* \leq 360) \quad (4.57)$$

$$s^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} s_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq s^* \leq 360) \quad (4.58)$$

$$v^* = \frac{\sum_{i=\tau_{leave}}^{\tau_{enter}} v_i}{\tau_{enter} - \tau_{leave}} \quad (0 \leq v^* \leq 360) \quad (4.59)$$

Where ' τ_{enter} ' - number of frames that are entered into tracked object view, ' τ_{leave} ' - number of frames that are leaved into tracked object view, (h_i, s_i, v_i) are the i^{th} frame of hue, saturation and value

Then, compute 3 color moments

- Moment-I (Mean)
- Moment-II (Variance)
- Moment-III (Skewness)

Table 4.2 Color Features

Feature	Formula
Mean - defines the contribution of individual pixel intensity for the whole scene	$\mu_i = \frac{1}{N} \sum_{j=1}^N P_{ij}$
Variance - determines how each pixel can be varied from the neighboring or center pixel.	$\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^N (P_{ij} - \mu_i)^2 \right)}$
Skewness - refers to symmetry measure in a particular scene. It means when pixel values occur at the regular interval.	$SW_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^N (P_{ij} - \mu_i)^3 \right)}$

MATHEMATICAL ANALYSIS OF FEATURE EXTRACTION

Firstly – Let shape features (Topology & Geometry) is computed as \Rightarrow [8]

Assume that voronoi segments be 250. So total shape features S is given as,

$$S \rightarrow 8 * 250 = 2000$$

Secondly - Let RGB and HSV be two color spaces C, i.e. given as,

$$R \ G \ B \Rightarrow [3] \ // \ \text{number of channels}$$

$$H \ S \ V \Rightarrow [3]$$

Color Space Conversion,

$$R \ G \ B \rightarrow H \ S \ V \rightarrow [3]$$

Color moments as, $\{I, II, III\} \rightarrow [3]$

$$\text{So, } C \rightarrow 3 * 3 = 9$$

Thirdly, let texture features T is given as [7]

Compute 7-feature vectors in 3-directions (Horizontal, Vertical, and Diagonal) and different feature maps (e.g. 5). After computing this, the total number of T is $7 \times 5 = 35$.

Lastly, S, C and T feature vectors are concatenated for each frame F as below,

$$(S, C, T) = [2000 + 9 + 35]$$

$$= 2004$$

Visual Vocabulary Generation using CT

Construction of Visual vocabulary (V_C) is significant in CBVR system. It is the next stage of BOVW. In those previous works, several clustering algorithms were used for vocabulary generation process [13], [22], [72]. K-Means clustering algorithm is frequently

used technique but it limits its merits in case of handling large amount of datasets. To overcome this problem, in this work Clustering Tree (CT) is used.



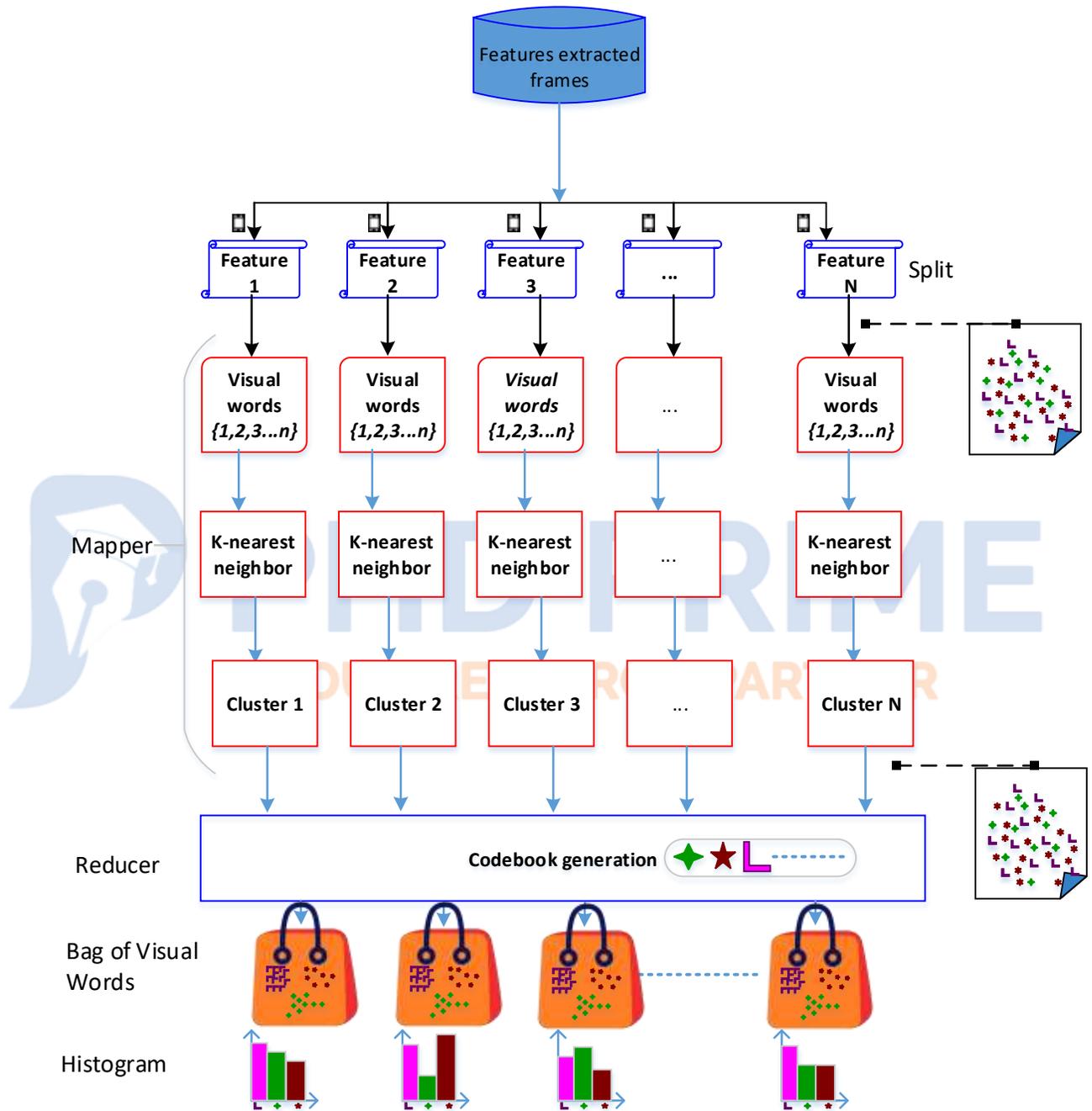


Figure 4.15 Features Clustering using MapReduce

To construct PCT, we combine the concept of predictive clustering and randomized tree construction. PCT is provoked using a decision trees algorithm. During the tree construction process, local descriptors values are used in a random manner. Algorithm 1 shows the TDTDT algorithm procedure in detail.

Algorithm 1: Decision Tree

```

procedure PCT(E) returns tree
1: (t*,h*,p*) = BestTest(E)
2: if t* ≠ none then
3:   for each Ei ∈ p* do
4:     treei = PCT(Ei)
5:   return node(t*,Ui{treei})
6: else
7:   return leaf(Prototype(E))
procedure BestTest(E)
1: (t*,h*,p*) = (none, 0, Φ)
2: for each possible test t do
3:   p= partition induced by t on E
4:   h= Var(E)- ∑Ei∈p  $\frac{|E_i|}{|E|}$  Var(Ei)
5:   if(h>h*) ^ Acceptable (t,p) then
6:     (t*,h*,p*) = (t,h,p)
7: return (t*,h*,p*).
  
```

In this proposed work, the concept of Threshold Based-Predictive Clustering Tree (CT) is involved which include threshold values for clustering. Algorithm 2 depicts the overall working procedure for the construction of visual vocabulary using threshold. In visual word construction, noise is the problem, which occurs when the repeated visual words are entered. These noisy visual words are referred as unusable visual words which create difficulty during retrieval. The problem of visual words in clusters is solved by using Numeric Semantic Analysis. To reduce noise, we assign numbering for each visual word and repeated visual words are ignored. Hereby, we solve the problem of noise into generated visual words. By this technique, we find similar visual word that occurs more than once by providing similar integer numbers for similar visual words. Hence, we reduce the problem of noise which in turn reduces the computation time for the construction of visual words.

Algorithm 2: Construction of visual vocabulary using CT

Input: set of subsets of local descriptors as

$$S_i = \{s_1, s_2 \dots s_n\}$$

Output: visual vocabulary.

1: Begin

2: **Choose** s_i based on *threshold*(t)

3: **Set** $TS = s_i$

4: **Assign** $da = ta$ & $ca = 128$ vector

5: Perform cluster c using PCT

6: **Construct** $tree(T)$ using c based on randomized tree

7: *leaf* (l) = Vw

8: $Vc = \Sigma l;$

9: End

In initial step, local descriptors are congregated and partitioned into subsets (S_i) using machine learning algorithm and predictive clustering. Resultant subsets are given as input for vocabulary construction and output from this is used to construct V_c . Subsets of local descriptors (s_i) are selected using threshold value (t).

The selected values (s_i) constitute a training set (TS). For tree construction (T), we assign descriptive attribute (d_n) as target (t_n) and also clustering attribute (c_n). The value of descriptive attribute (d_n) is 128 dimensional vectors. This is one of the unique characteristics of the PCT. Root node maintains the overall information of the image and it recursively partition to construct tree. Apply pre-pruning techniques to control the size of visual vocabulary. Pre-pruning requires information regarding the minimum number of descriptor in each tree leaf (l). We can easily determine the number of instance required for a leaf to obtain the desired number of leaves for the tree, for the given dataset. Each leaf 'l' of the tree has a separate visual word (V_w). All the leaves are combined to form (V_c) (I.e.) visual codebook (V_{cb}) for frame.

These processes are managed by Map and Reduce functions in parallel manner. The Map and Reduce functions is given as,

$$\text{Map}(fl^{(n)}, ft^{1\dots F(l)^n}) \rightarrow [\langle l(1), ft(1) \rangle, \dots, \langle l^{(F(l)^n)}, ft^{F(l)^n} \rangle],$$

$$\text{Reduce}(l, ft^{1\dots F(k)}) \rightarrow [\langle l, ft^{avg} \rangle] \quad (4.60)$$

The map task holds the information of leaf node. CTs are efficient in enumeration. Here we resolve the problems of handling small amount of dataset by using random forest tree. By concentrating the individual V_{cb} from each of the CT, we can obtain final codebook. The formula for evaluating CT is as given below,

$$V_{Cb} = \sum_{i=1}^n [V_{cb(i)}] \quad (4.61)$$

After the completion of V_{cb} computation process, the next step is to constitute each 3D model (frames in dataset or query) as a histogram which is nothing but occurrences of the codebook element. For this purpose, we sort all the descriptors along the tree. As a further step, the number of descriptors are counted that is present in given 'l' (i.e.) leaf, which accounts for the number of visual word. We have presented the frame as a histogram of descriptors per visual word.

3) MapReduce

Matching is the final stage of CBVR where we have used the distance function and similarity schemes to retrieve proficient results. To measure the similarity between two images, a distance function is required. In order to accomplish this task, we have used L_2 (Euclidean) distance function and soft weighting scheme. Map Reduce task is partite as map and reduce function.

- Map Stage
- Reduce Stage

Mapper function produces the key value pair for both testing and training samples and it calculates the distance between each testing sample and training sample. In map stage, measurement between one key point and its neighboring visual word is performed for computing similarity function using k-nearest neighbors on Hadoop platform. Then, the mapping process is considered by mapping the key point to its top-k nearest neighbor. K-nearest neighbor is widely used for classification decision on the basis of the closest k-nearest neighbors in the feature space.

From each mapper, we get result as the number of output pairs $(Ix^{(n)}, vt_k^{(n)})$. Here, $Ix^{(n)}$ indicates the key $(Ix^{(n)})$ which is the index of the i^{th} ranking visual word in the top-k results and $vt_k^{(n)}$ indicates value $(vt_k^{(n)})$ which represent the similarity score with partial weight according to the rule that forms proximity with visual word. Priority is given to the highest similarity score during the process.

In the reduce phase, the reducer function is applied to determine the global nearest neighbors. Here, the histogram is computed by arranging the (key, value) $(Ix^{(n)}, vt_k^{(n)})$ pair for video representation. In particular, reduce stage gathers the partial weight value of each key pair-value $(Ix^{(n)}, vt_k^{(n)})$ to its corresponding key index.

Further, we propose a soft-weighting scheme for determining the weight of each visual code word. Our proposed soft-weighting scheme is constantly better than traditional weighting schemes used in other previous works. For each key point in the image (query image and dataset frame), we select top-k nearest visual words instead of searching only for the nearest visual words. If suppose we have a visual vocabulary of k

visual words then, k-dimensional vector $VT = [vt_1, vt_2 \dots vt_k]$ will be represented with each component vt_k . Here, vt_k represents the weight of a visual word 'k' in an image. It can be calculated as follows,

$$vt_k = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j, k) \quad (4.62)$$

$$sim(j, k) = dis_{j,k} \quad (4.63)$$

Where M_i represents the number of key points whose i^{th} nearest neighbor is visual word 'k'. Estimating $sim(j, k)$ constitutes the similarity between key point 'j' and visual word 'k'. Key point is found based on its similarity to work 'k', weighted by $\frac{1}{2^{i-1}}$ where the constituted word is its i^{th} nearest neighbor. Usually, $N=4$ is reasonable to be set. The component $dis_{j,k}$ represents the distance between two images. To predict the distance, we have used L_2 distance functions. Occurrence of visual word is constructed as a tree like index structure, which proves to be computationally efficient. We fetch weights from both (i.e.) visual vector of query image and the image retrieval from nearest neighbors. Equation used to determine L_2 distance is given as,

$$dis = \sqrt{\sum_{x=1}^k (Vt_{xi} - Vt_{xj})^2} \quad (4.64)$$

Where, $x = \{1,2,3 \dots k\}$, and V_t - visual vector of image. Based on the calculated visual vector similarity, ranking is performed using k-nearest neighboring visual vector images. It makes our proposed system efficient and effective. Finally, the matched results

are ranked according to the index value of images and it is returned back to the queried users as feedback

Pseudocode: 3D CBVR

- 1: Input - $F_1, F_2, F_3, \dots \dots \dots F_n$
- 2: $F_1, F_2, F_3, \dots \dots \dots F_n \leftarrow$ STPGF
- 3: $KF_{E1}, KF_{E2}, KF_{E3} \dots \dots KF_{En}$
- 4: STPGF \leftarrow Max (a,b,c)
- 5: Max (a,b,c) \leftarrow M_{SE}
- 6: $M_{SE} \leftarrow$ Computer(W_i)
- 7: If $F(x,y,z) = 1$
- 8: Then $SQSP \leftarrow \{x, y, z; a_1, a_2, a_3, \varepsilon_1, \varepsilon_2, \varepsilon_3, \Phi, \theta, \chi, t_x, t_y, t_z\}$
- 9: Compute MSE
- 10: $SE_d \leftarrow dS DF$
- 11: Construction of Vcb
- 12: $d^S_{DF} \leftarrow CT(Vcb)$
- 13: $V_{cb} \leftarrow Vtk$

14: Retrieves relevant results $\leftarrow V_1, V_2, V_3 \dots \dots \dots V_n$

The overall processing pseudo code of 3D-CBVR framework is illustrated below. The above mentioned procedural steps are proceeded one after another for the achievement of accurate results. From ‘n’ number of frames, key frames are selected as well as further feature extraction, visual codebook generation and similarity matching is sequentially performed. The notations used in pseudo code are as follows: $F_1, F_2, F_3, \dots \dots \dots F_n$ – video frames, KF_{En} - Extracted key frames, SQ_{SP} - Super Quadratic Shape Parameters, MSE-Mean Squared Error, and SE_d - Super Ellipsoid.

4.5. Results Discussion

To evaluate the retrieval performance of 3D-CBVR, this section involves with three subsections such as 3D model dataset and tools, performance evaluation and experimental results. In dataset and tools, we describe about proposed computer system and its specifications. Next sub-section describes the significant parameters for evaluating 3D video retrieval of this proposed system. Final sub-section provides the results of our proposed work, which is better than previous work. We show plots in graphical representations that give completeness for showing improvements of our research work.

In our proposed method, Multi-modal graph learning (MMGL), Predictive clustering tree (PCT), Sequential search algorithm (SSA), and Segmentation based query (SBQ) are used for CBVR across different characteristics.

4.5.1 Testing Scenarios

- Scenario 1: N_1 & N_5 are used
- Scenario 2: N_1 , N_2 & N_5 are used
- Scenario 3: All the five nodes ($N_1 - N_5$)

Where $N_1 - N_4$ are slave nodes and N_5 is the master node. Table 4.3 represents the system configurations for five different nodes.

Table 4.3. Hardware Configuration of our Implementation

Specifications	N_1	N_2	N_3	N_4	N_5
Processor	Intel Core i7 – 7700	Intel Core i5 - 10500	Intel Core i5	AND Fx-6100	Intel Core i7-10700F
CPU Core	Quad Core	6	4	6	8
CPU Speed	3.6GHz	3.1 GHz to 4.5GHz	3.7 GHz	3.9 GHz	3.1 GHz to 4.8 GHz
RAM	8GB	8GB	8 GB	8 GB	8GB

In our experiment, we use video files for retrieval process in which videos are converted into frames using frame conversion tool. Then, converted frames are considered as dataset.

4.5.2. Dataset Description

Our method utilizes three public 3D datasets such as YouTube (I), Atomic Visual Action (AVA) dataset, KTH Dataset (III). These databases are employed to estimate the performance of the descriptor. All these datasets consists of set of videos for different human actions and scene description.

- Encoding format: flv, wmv, avi, mpg, mp4, ram ...
- Frame rate: 15fps, 25fps, 29.97fps ...
- Frame resolution: 174x144, 320x240, 240x320 ...

Table 4.4 exemplifies the datasets used in deep learning based 3D videos retrieval for a given query image. The main of attributes of each dataset is about its description, number of videos, number of classes, video features and video type.

Table 4.4 Dataset Description

Dataset	Description	# of videos	# of classes	Video characteristics	Video type
YouTube	Entertainment (games, news, shopping)	1000	> 20	Colorfulness, object & camera motion and texture	HD, full HD, ultra HD

Atomic Visual Action (AVA) dataset	Human activity movie clips	430	80	Spatial, temporal and color	HD
KTH Dataset	Human different activities	2931	6	3D Motion , dynamic environment, objects from action sequences	Full HD

4.5.3. Comparative Analysis

This section is used to evaluate the performance of the proposed Hybrid 3D CNN with the use of performance indicators. The performance indicators are listed as follows: Accuracy, Recall, Precision and F-Score. These metrics are designated as follows:

4.5.3.1 Precision

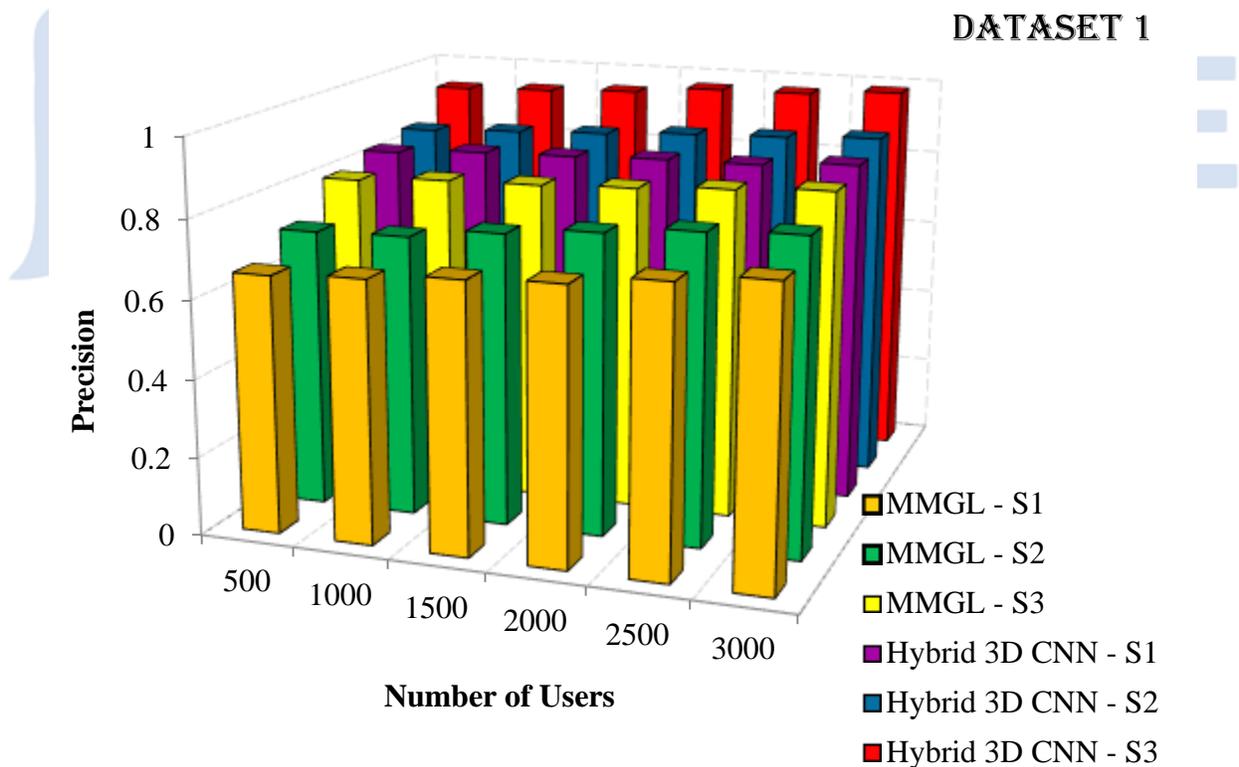
Precision evaluates the relevance (positive output) of the obtained retrieval result. For example, if 98 videos are retrieved correctly from the sum of videos retrieved (100 videos) means, then the precision value is **0.98**.

$$\text{Precision} = \frac{\text{\#relevant videos retrieved}}{\text{sum of videos retrieved}} \quad (4.65)$$

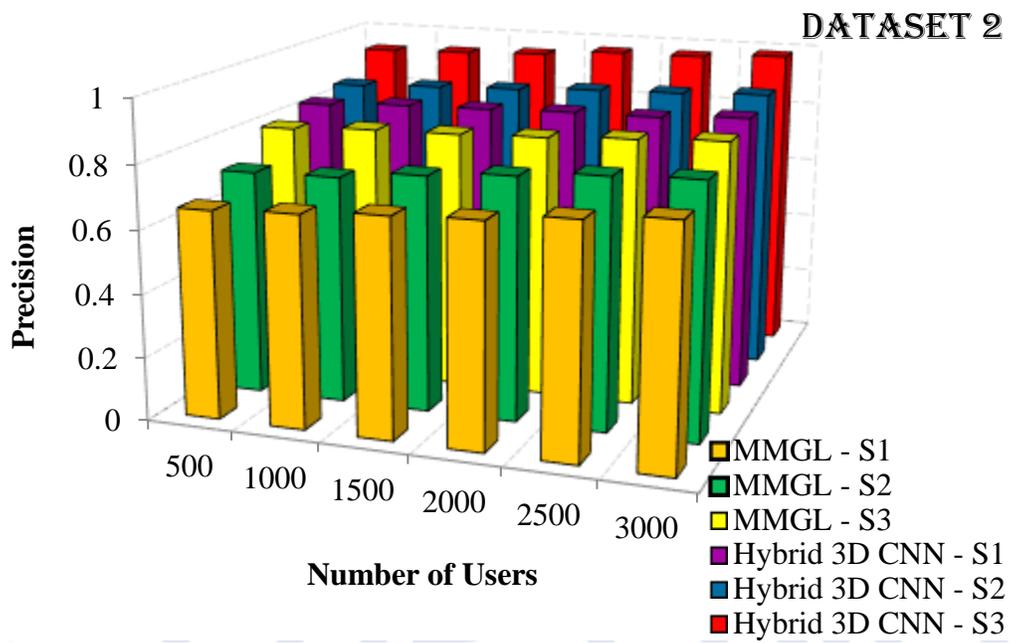
In other words, the parameter precision is computed as the ratio of retrieved relevant (Q) 3D models in top returned result (R) which are related to the specified query.

$$Precision = \frac{Q \cap R}{R} \quad (4.66)$$

Figure 4.16 shows graphical plots for three scenarios, considering the usage of Hadoop for precision. It implies that the presentation of scenario 3 is better in case of precision when compared to other scenarios. Precision is analyzed for the proposed and existing work as MMGL. In the proposed hybrid 3D CNN model is used to provide higher precision for query processing of each individual user. However, precision must be higher to show the great performance and also among 3000 of users, precision is higher especially for third scenario which provides the higher degree of precision.

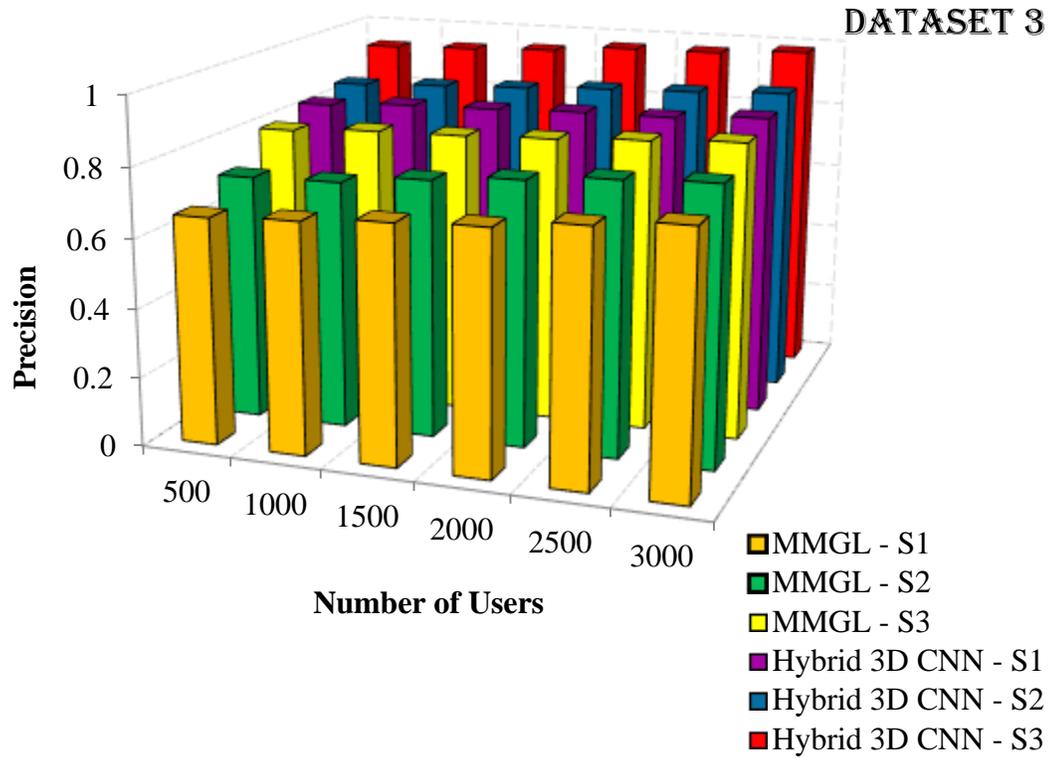


(a)



(b)

YOUR RESEARCH PARTNER



YOUR RESEARCH PARTNER

(c)

Figure 4.16 (a) (b) (c) Precision vs. Number of Users

Table 4.5 Statistical Analysis for Precision

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	0.84	0.77	0.88	0.8	0.985	0.867

YouTube 8M segments	0.85	0.775	0.89	0.82	0.981	0.875
KTH dataset	0.857	0.789	0.91	0.835	0.982	0.885

4.5.3.2 Recall

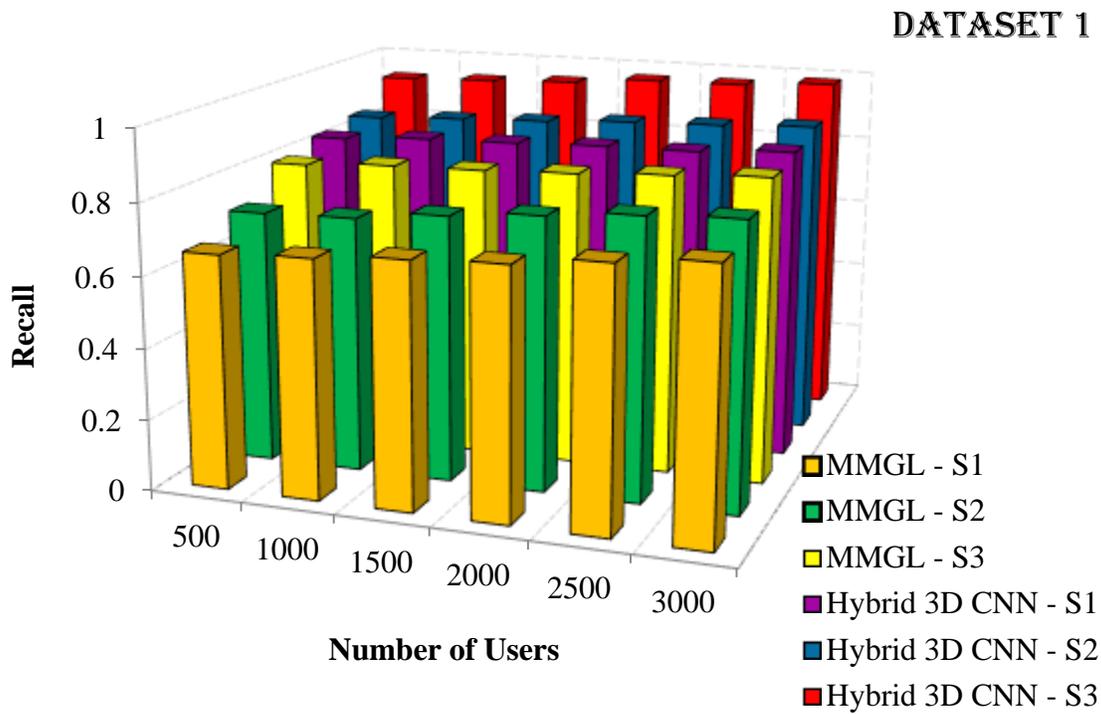
Then, the parameter Recall is defined as the ratio of successfully retrieved (Q) 3D models which are related to the specified input query.

$$Recall = \frac{Q \cap R}{Q} \quad (4.67)$$

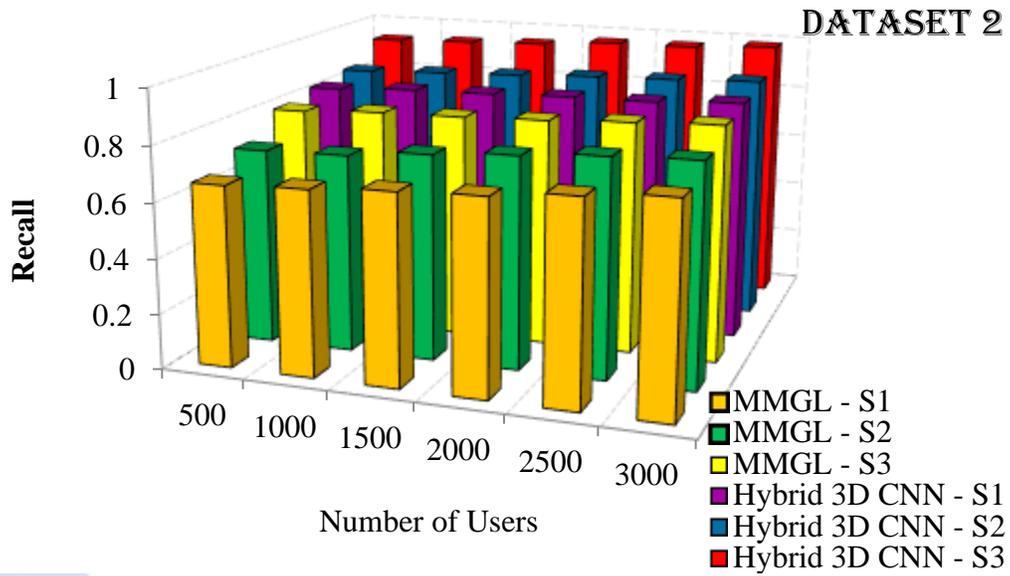
Recall defines the positive output from the total available positive results that is present in the retrieval system.

$$Recall = \frac{\text{\#. relevant videos retrieved}}{\text{sum of relevant videos in the database}} \quad (4.68)$$

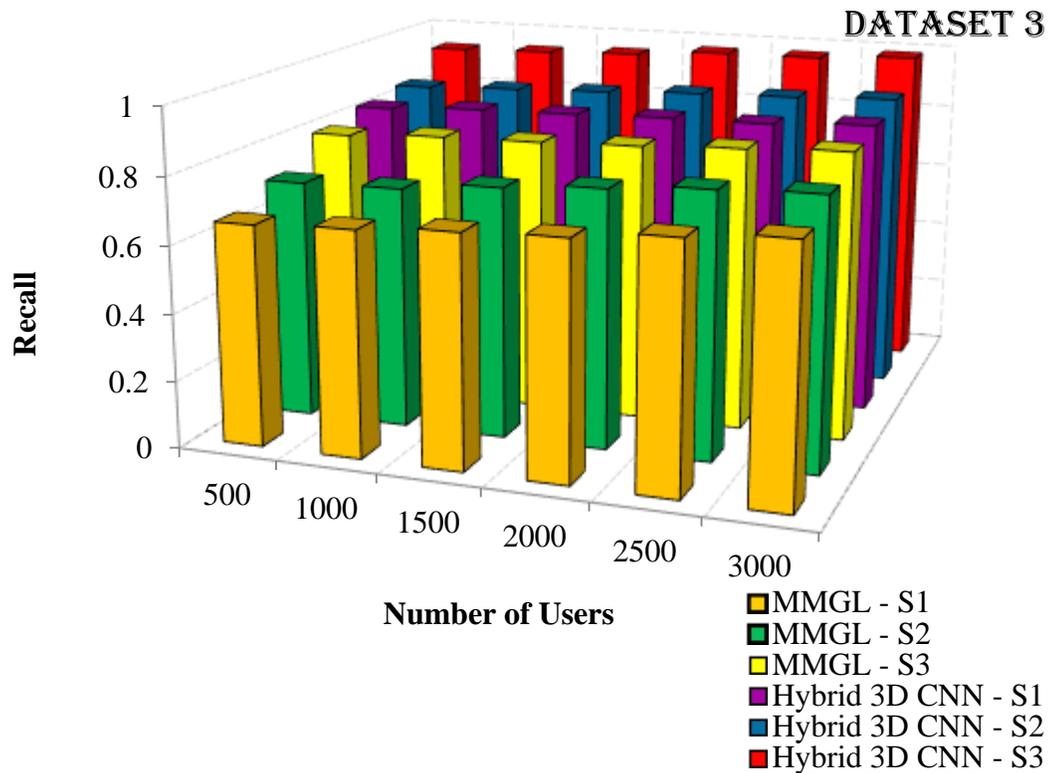
For example, if 190 videos are retrieved correctly from the sum of relevant videos in the database (200 videos) means, then the recall value is 0.95.



YOUR RESEARCH PARTNER (a)



(b) **PHD PRIME**
YOUR RESEARCH PARTNER



YOUR RESEARCH PARTNER

(c)

Figure 4.17 (a) (b) (c) Recall vs. Number of Users

Figure 4.17 shows the performance recall for number of users. 4.6 represents the resulted recall value taken for the three scenarios. In this result, training and testing results (validation loss and training loss) varies for each scenario. This analysis shows the difference in the recall based on our scenario. On comparison of recall, approximated 15 – 22% of time is higher for testing than training the dataset. We have compared the

scenarios and finally concluded that scenario 3 shows the better performance than other two scenarios.

Table 4.6 Statistical Analysis for Recall

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	0.81	0.74	0.87	0.80	0.952	0.84
YouTube 8M segments	0.83	0.745	0.89	0.82	0.957	0.85
KTH dataset	0.845	0.75	0.91	0.83	0.96	0.865

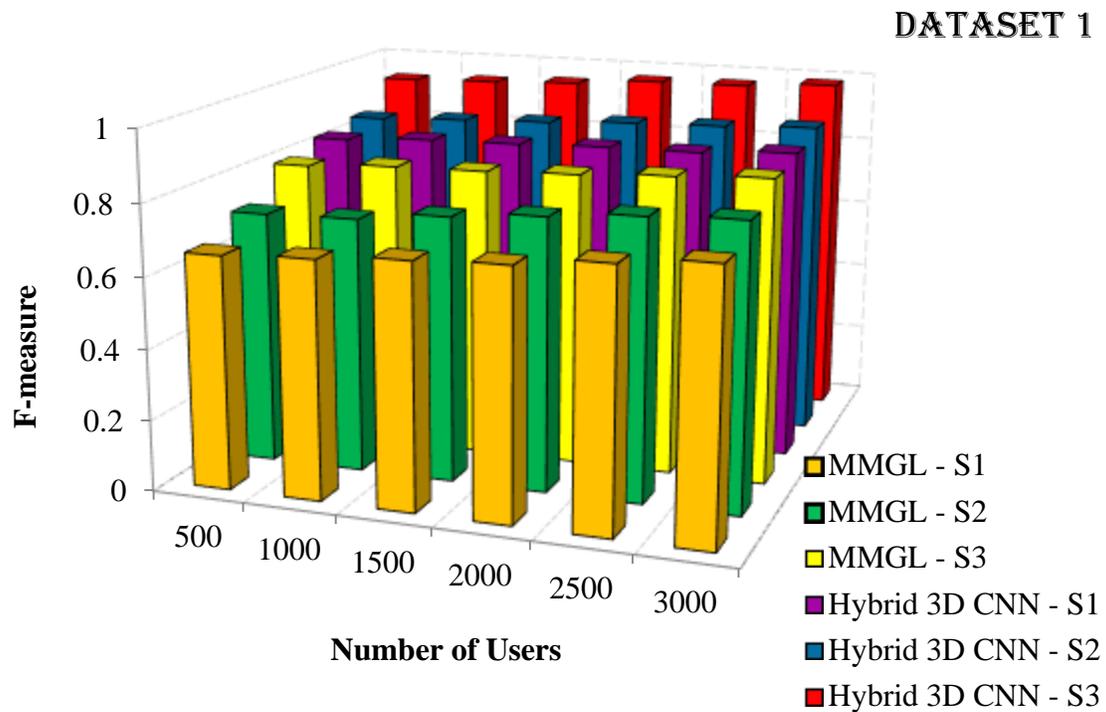
4.5.3.3 F-measure

It defined as the harmonic mean of precision and recall value. It is computed by,

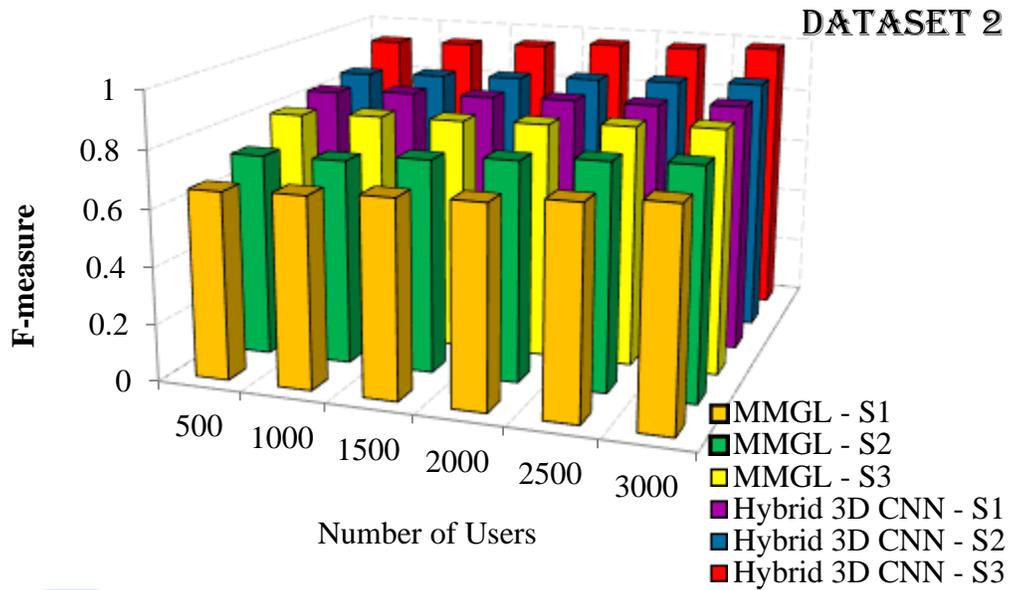
$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.67)$$

For example, if Precision value is 0.97 and Recall value is 0.95, then the F-measure is 0.959. Figure 4.18 shows the plot between number of users given queries and its response. It also shows that multiple queries from multiple users at the same time will

provide positive result only and thus the performance of F-measure increases with respect to the number of users increases.



(a)



(b) **PHD PRIME**
YOUR RESEARCH PARTNER

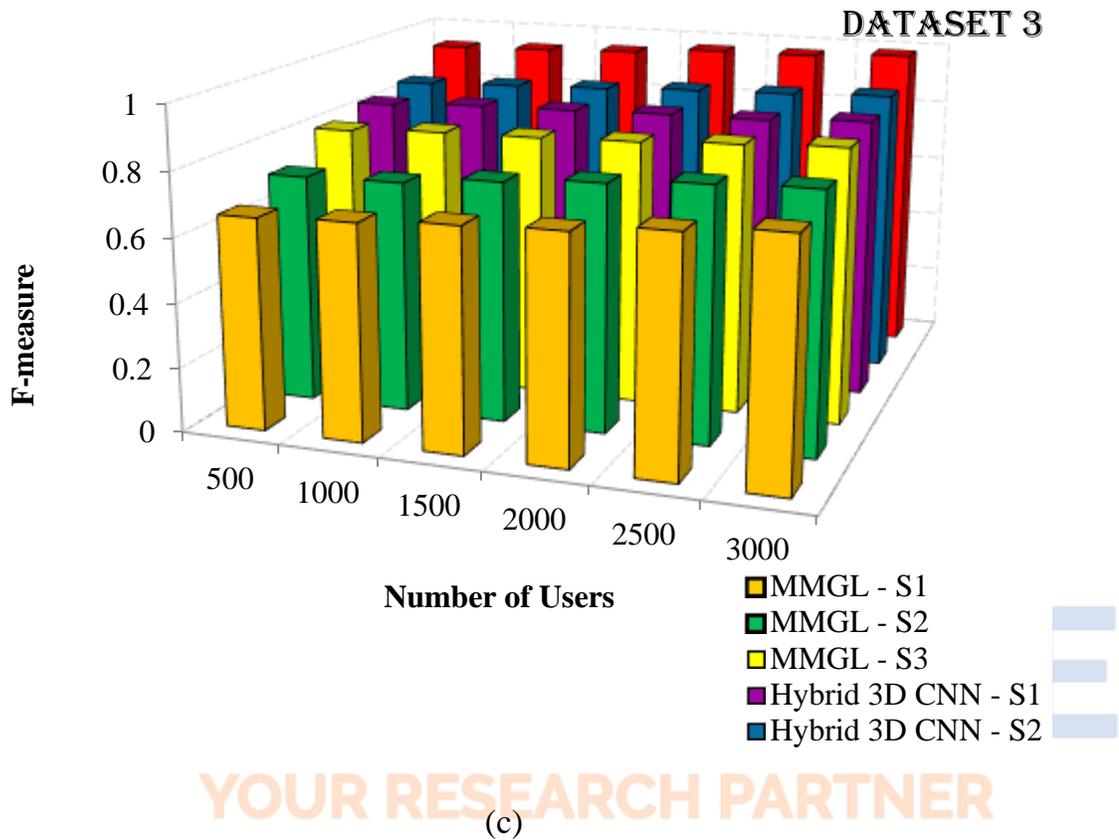


Figure 4.18 (a) (b) (c) F-measure vs. Number of Users

From this, the positive result intimates that the number of users served properly for the given query within a short period of time. It implies that the comparison between the system on using Hadoop and without Hadoop for video retrieval.

Table 4.7 Statistical Analysis for F-measure

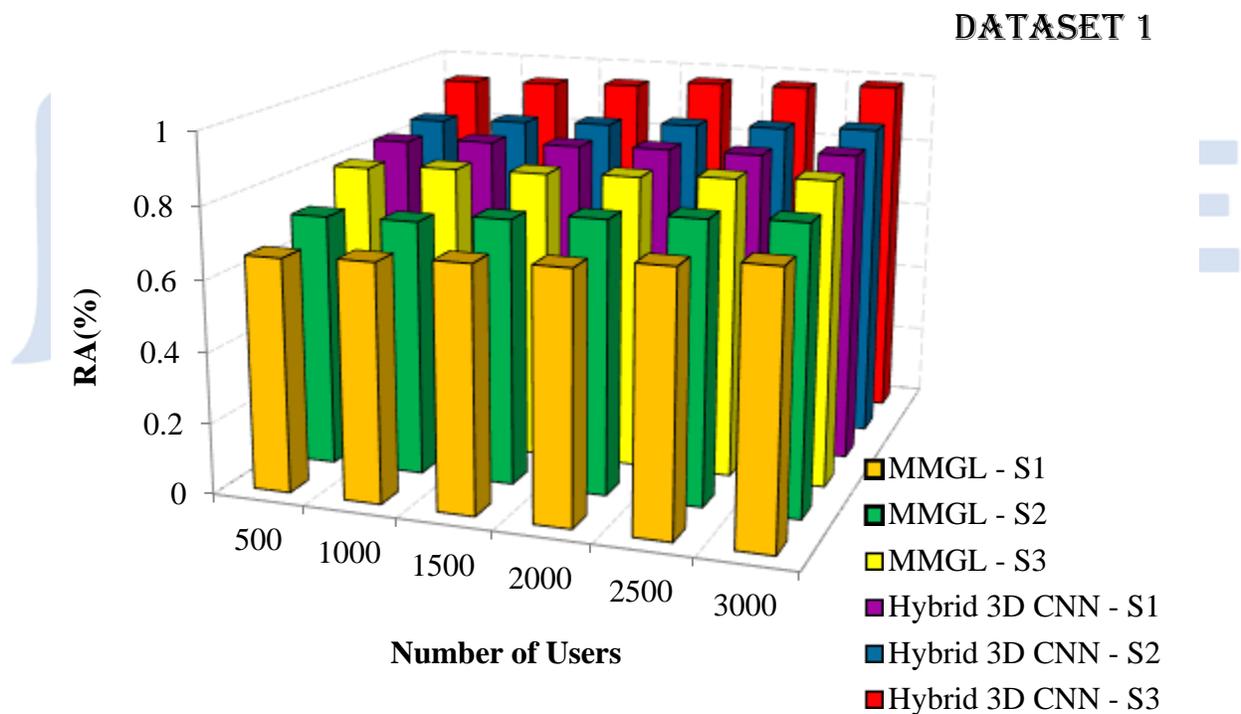
Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	0.81	0.74	0.87	0.80	0.952	0.84
YouTube 8M segments	0.83	0.745	0.89	0.82	0.957	0.85
KTH dataset	0.845	0.75	0.91	0.83	0.96	0.865

4.5.3.4 Retrieval Accuracy

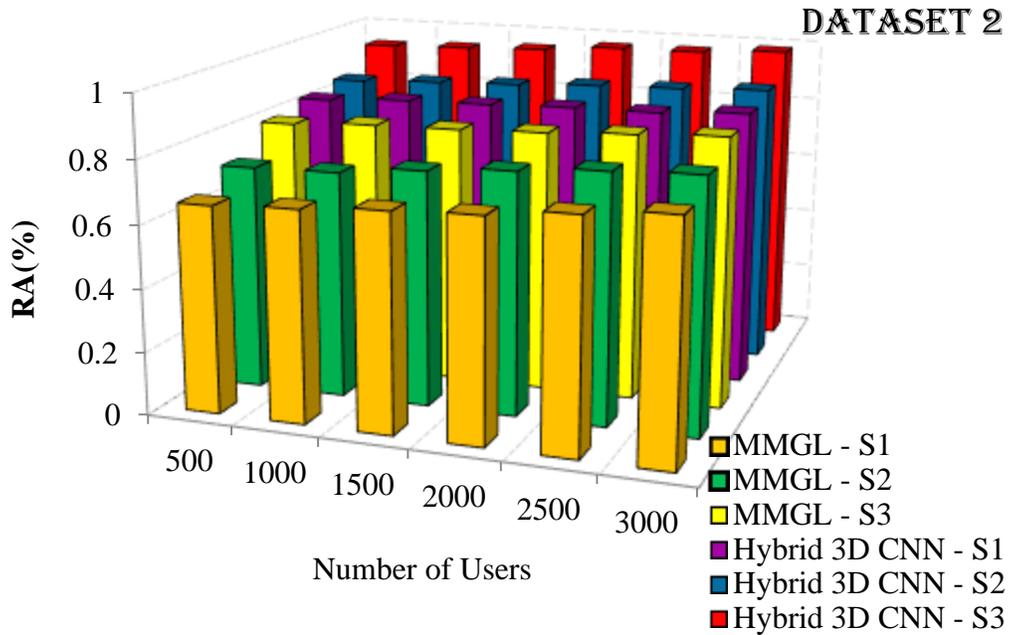
It is computed by sum of all positive and negative measures. The RA value is computed by,

$$RA = \frac{\text{All positives} + \text{All negatives}}{\text{All relevant videos in the database}} * 100\% \quad (4.68)$$

For example, if 194 videos are retrieved correctly from the sum of relevant videos in the database (200 videos) and also from 6 non-relevant videos, 3 are correctly find as non-relevant means, then the RA is 98.5%. Figure 4.19 shows the performance of the RA with respect to the number of users. The retrieval accuracy is improved in the proposed Hybrid 3D CNN using segmentation and feature extraction. Similarly, the retrieval accuracy is decreased in terms of poor segmentation and feature extraction.

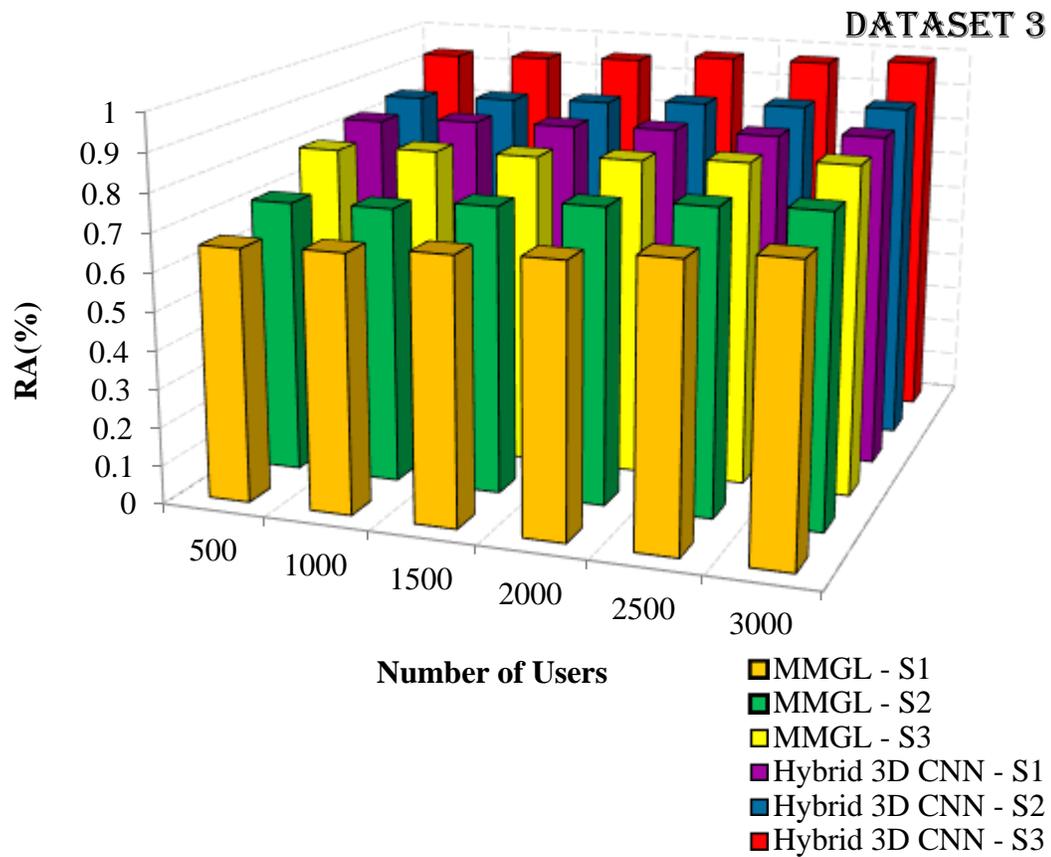


(a)



PHD PRIME
YOUR RESEARCH PARTNER

(b)



(c)

Figure 4.19 (a) (b) (c) RA vs. Number of Users

Table 4.8 Statistical Analysis for RA

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL

AVA dataset	93.91	80.2	96.2	83.4	98.9	85.2
YouTube 8M segments	93.6	80.5	96.5	83.6	99	86.5
KTH dataset	93.5	80.6	96.4	83.9	98.5	88.8

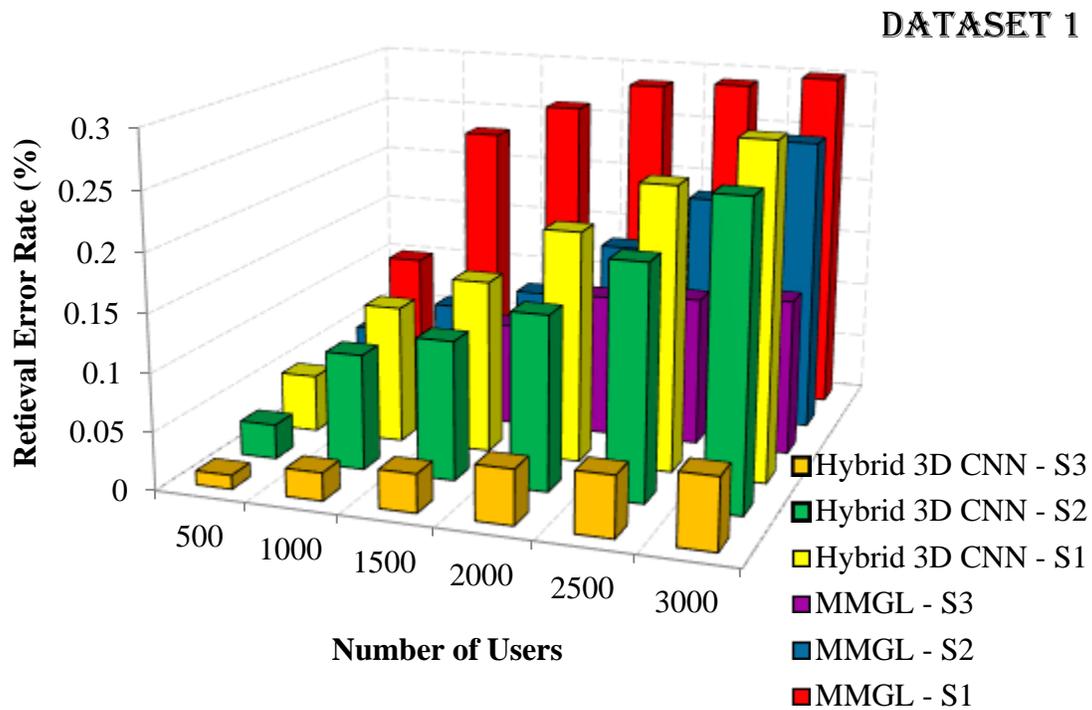
4.5.3.5 Retrieval Error Rate

It is the negative metric that is based on the specific error rates. It is also called as image false match rate. For the same features set using pre-defined threshold values for video retrieval, the false matches are occurred. It is computed as follows:



$$\text{Error_Rate} = \frac{\text{Non-Relevant videos Retrieved}}{\text{Sum of videos retrieved}} \quad (4.69)$$

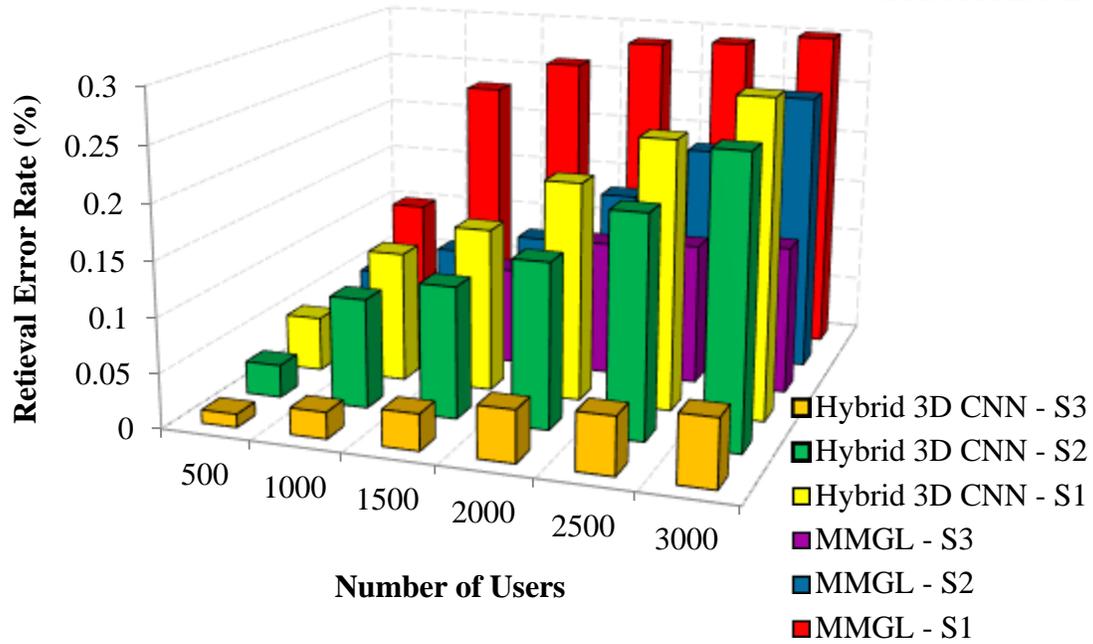
PHD PRIME
YOUR RESEARCH PARTNER



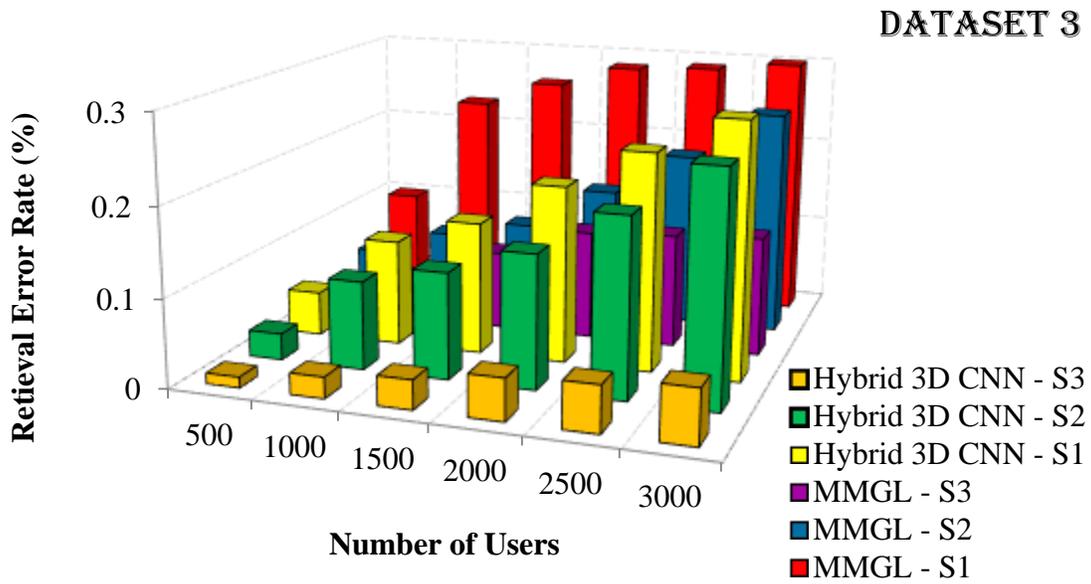
(a) YOUR RESEARCH PARTNER

For example, if 10 videos are retrieved incorrectly (non-relevant) from the sum of videos retrieved (200 videos) means, then the retrieval error rate is 0.05. Figure 4.20 shows the performance of the retrieval error rate with respect to the number of users. The performance of the MMGL and hybrid 3D CNN is compared for three different scenarios. Among all the cases, scenario 3 is suited to reduce the error rate. The proportion of errors made over the whole set of results for a given query. The error rate obtained from training data and reconstructs the training results for reducing the errors.

DATASET 2



(b)



(c)

Figure 4.20 (a) (b) (c) Retrieval Error Rate vs. Number of Users

Table 4.9 Statistical Analysis for Retrieval Error Rate

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	0.287	0.358	0.258	0.312	0.05	0.58
YouTube 8M segments	0.281	0.35	0.194	0.250	0.0514	0.56

KTH dataset	0.296	0.42	0.254	0.309	0.052	0.59
-------------	-------	------	-------	-------	-------	------

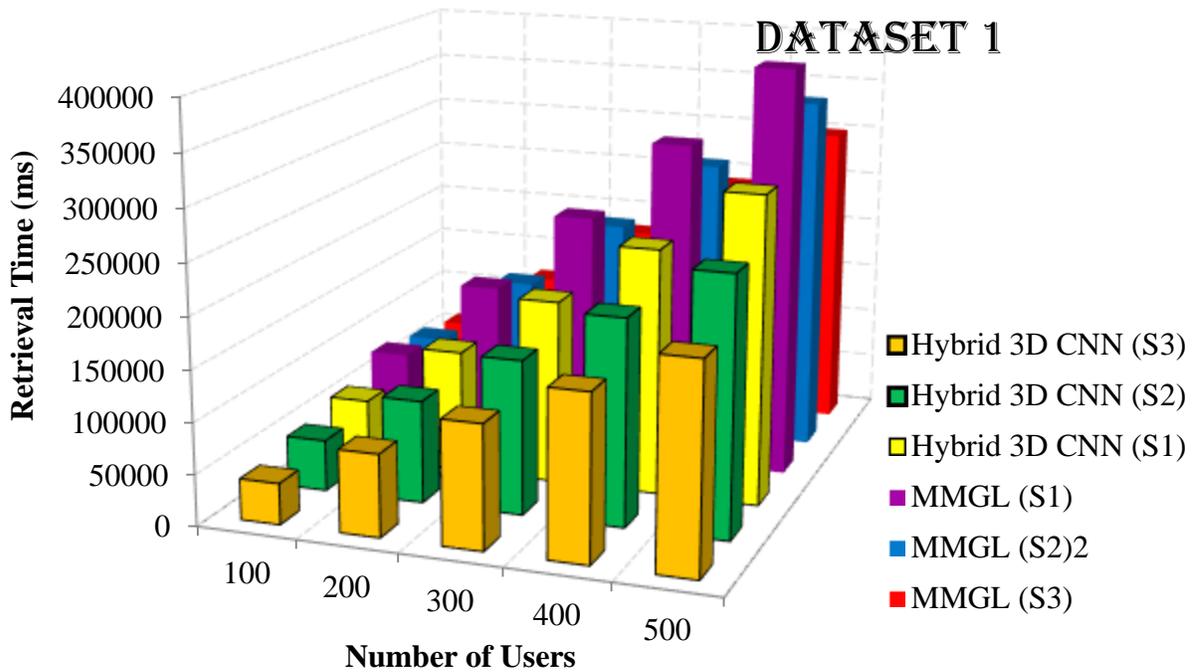
4.5.3.6 Retrieval Time

It is computed by the sum of time that is taken for each process during video retrieval and the retrieval time $R(t)$ is computed by,

$$R(t) = KF_{e(t)} + F_{e(t)} + BoVW_{c(t)} + S_{m(t)} \quad (4.70)$$

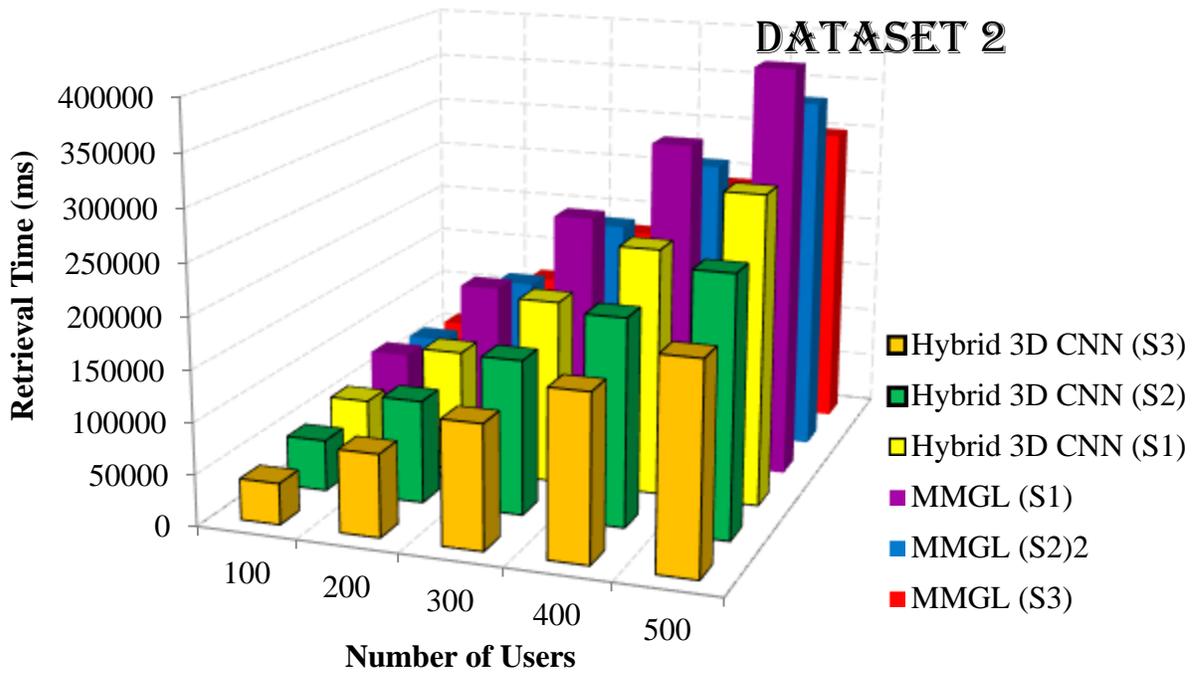
Where $KF_{e(t)}$ is the key frame extraction time, $F_{e(t)}$ is the feature extraction time, $BoVW_{c(t)}$ is the bag of visual words construction time and $S_{m(t)}$ is the similarity matching time

For example, if $KF_{e(t)}$, $F_{e(t)}$, $BoVW_{c(t)}$, $S_{m(t)}$ takes 50ms, 135ms, 150ms, 65ms respectively, then the $R(t)$ is 400ms. Retrieval time is computed from the query arrival to the HDFS system and it must be lesser to prove the system provides the better results. Figure 4.21 shows the performance of the retrieval time with respect to the number of users. When compared to MMGL, the performance of hybrid 3D CNN is higher for all three scenarios. The retrieval time shows that the performance of hybrid 3D CNN brings the effective solution i.e. 3D CNN for texture and color features extraction and also MapReduce for features clustering which constructs the tree for easy mining of videos for a given query.

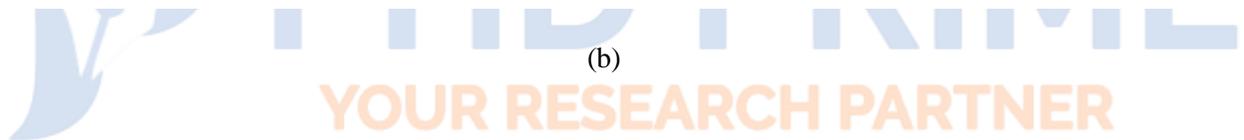


(a)





(b)



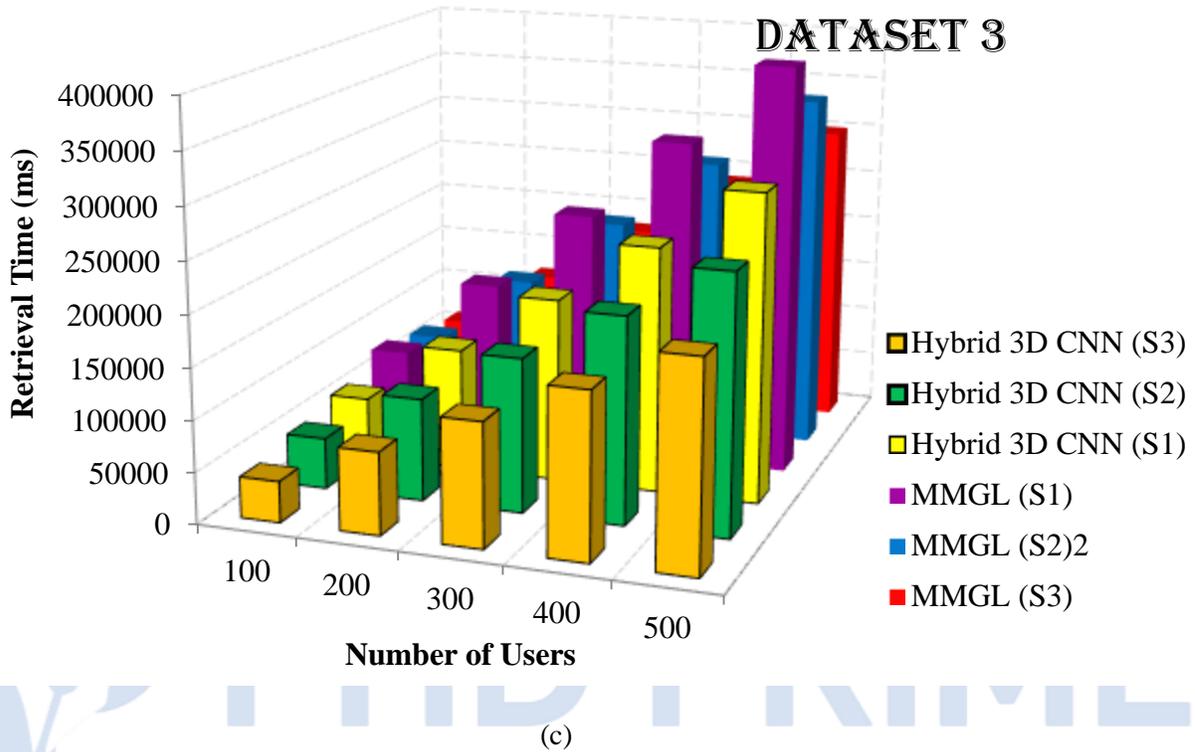


Figure 4.21 Retrieval Time vs. Number of Users

Table 4.10 Statistical Analysis for Retrieval Time

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	602000	1201000	301000	801020	252010	601050
YouTube 8M segments	652024	1201620	302014	801052	252084	601062

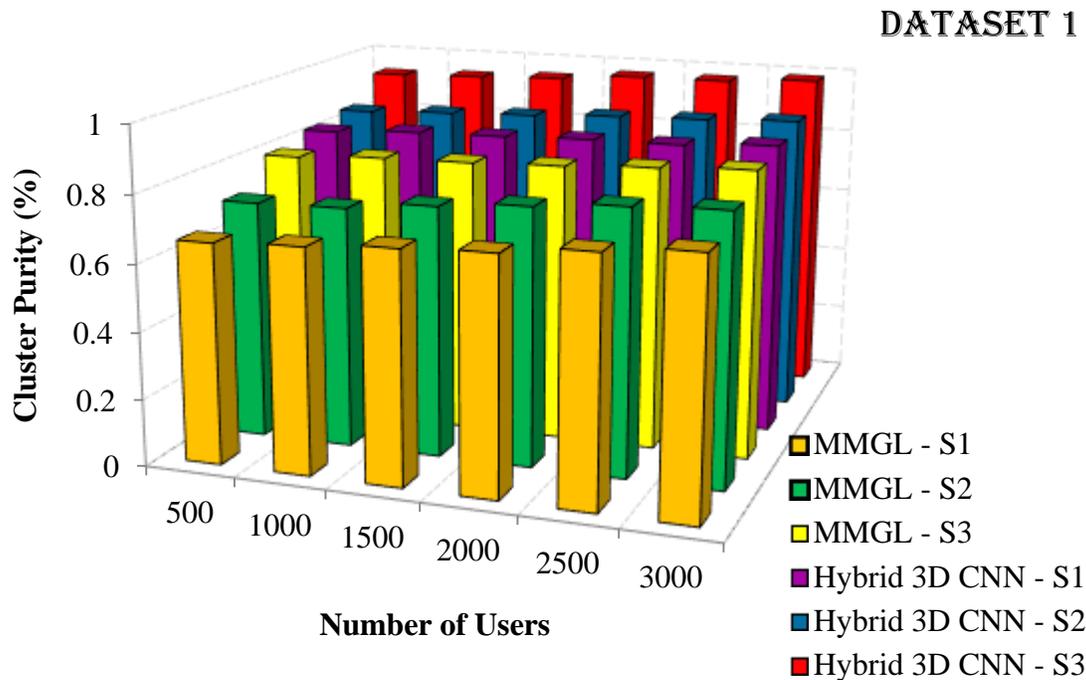
KTH dataset	602020	1201750	302010	801075	254000	601075
-------------	--------	---------	--------	--------	--------	--------

4.5.3.6 Cluster Purity

It is the measure of cluster construction in terms of grouping most similar visual words into a cluster. It is defined by

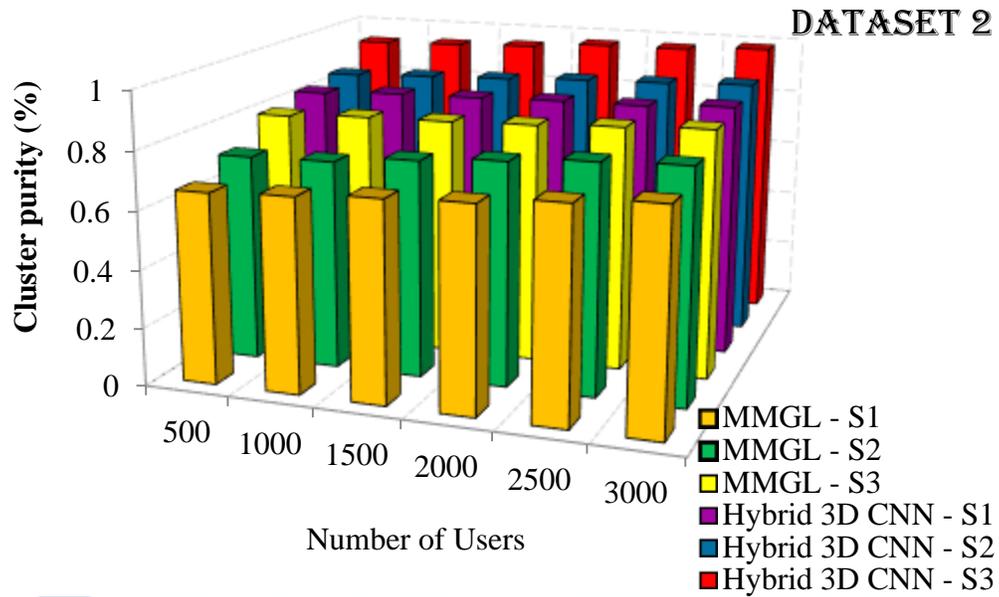
$$\text{Cluster Purity} = \frac{\# \text{ of clusters formed accurately}}{\text{Total number of clusters formed}} \quad (4.71)$$

For example, if we have 1200 clusters and 1188 clusters are formed correctly, then the cluster purity value is 0.999.

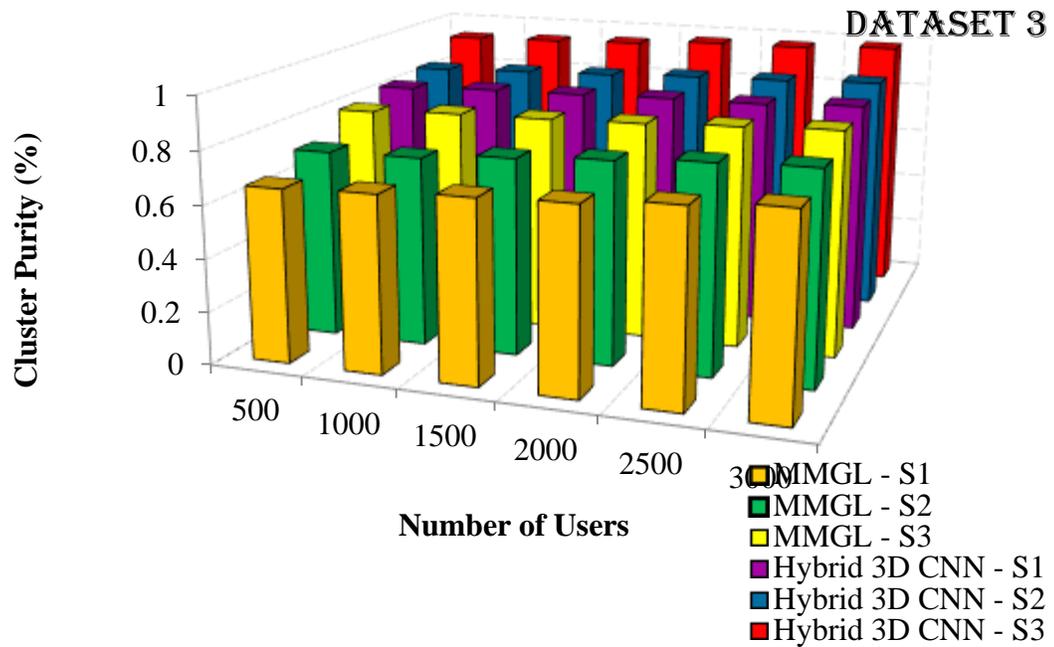


(a)

The performance of cluster purity is illustrated in figure 4.22. However, cluster purity is the measure of which clusters contain the single class. Its computation can be as follows: For each generated cluster, the number of data points from the most common class in the said cluster. The sum of overall clusters and divides by the total number of data points. Unique data points must be formed to the each individual cluster and based on the data similarity query results are validated. Table 4.11 represents the performance analysis of the proposed work vs. previous work, respectively.



(b) **PHD PRIME**
YOUR RESEARCH PARTNER



(c)

Figure 4.22 (a) (b) (c) Cluster Purity vs. Number of Users

Table 4.11 Statistical Analysis for Cluster Purity

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	85	78.9	88.9	81.5	99	83.5
YouTube 8M	86.9	79	91	82.5	99.2	85.7

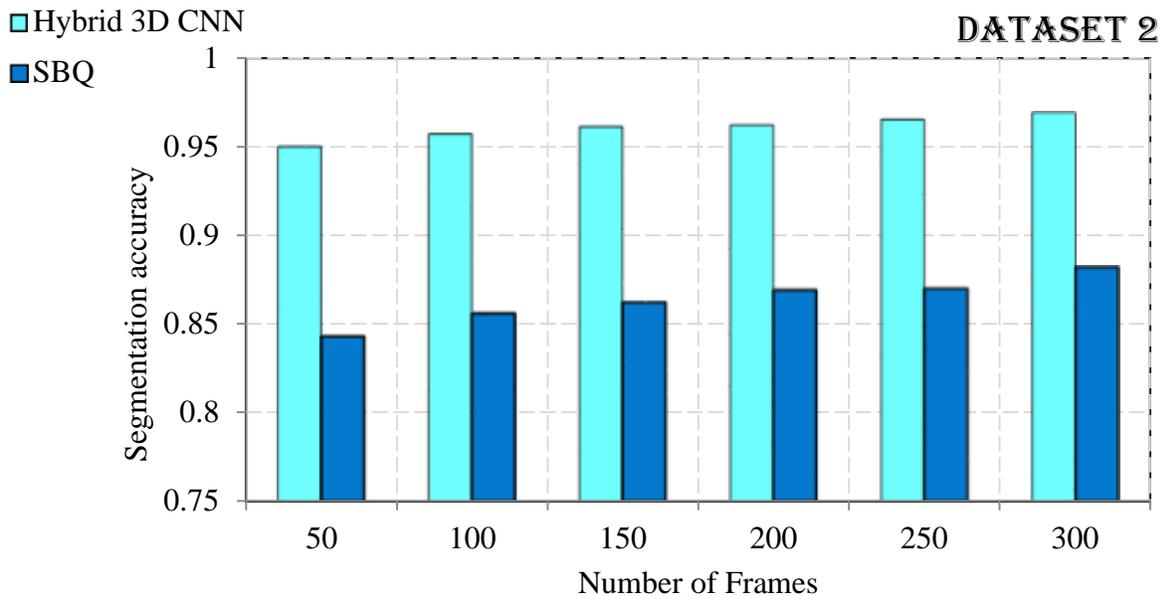
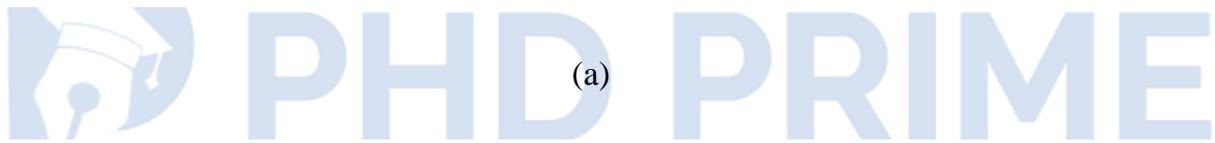
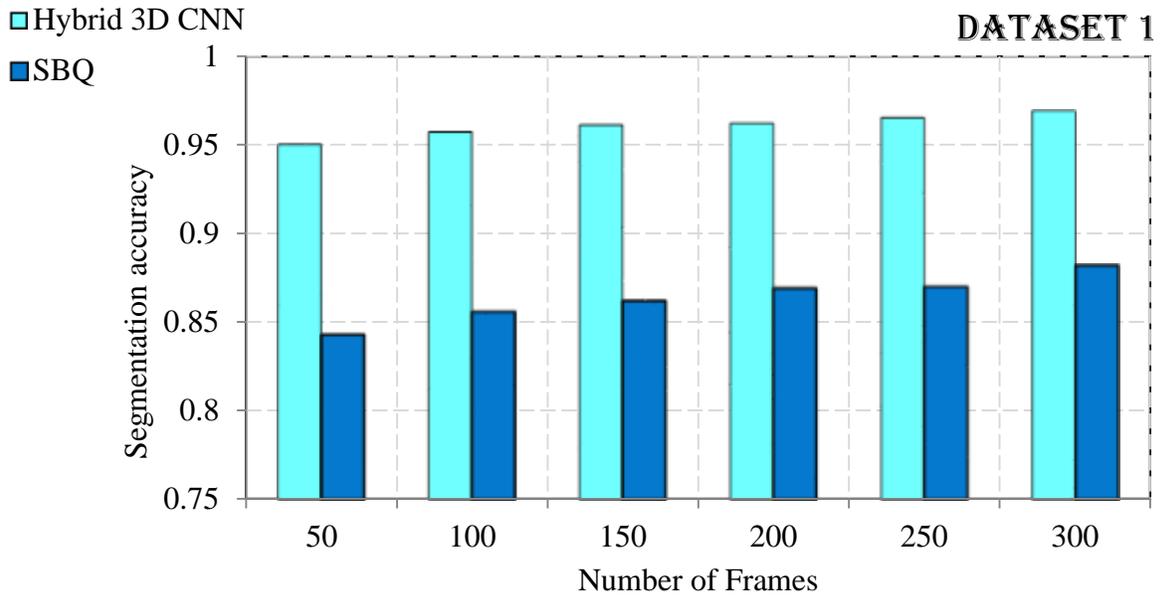
segments						
KTH dataset	87	79.8	92.4	89.8	99.5	86.9

4.5.3.8 Segmentation Accuracy

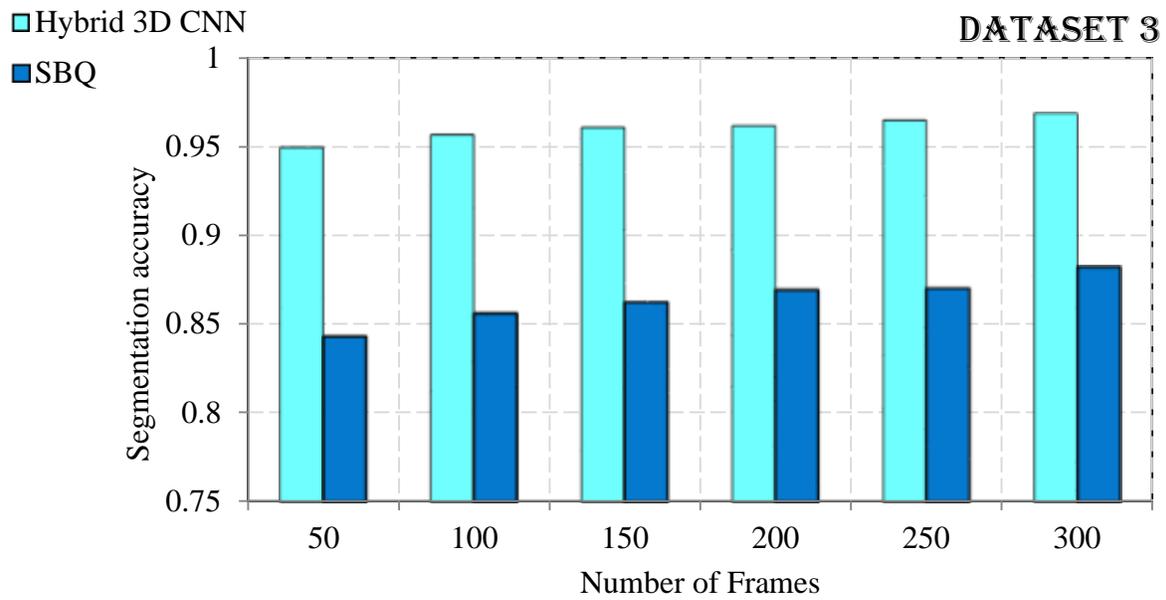
It is the measure of accuracy for the number of segments for each object in a query image and frames. However, segmentation accuracy must be higher to show the best performance.

$$SA = \frac{\text{Number of segments}}{\text{Total number of segments in a frame}} \quad (4.72)$$

For example, if we have 79 segments from the total number of segments (82) in a frame, then the SA value is 0.963. The segmentation accuracy is computed by the total number of segments classified into correct class. If the segments are classified into correctly, then the accuracy is higher in which the number of segments in a frame is useful for extracting the similar video contents for a query. Figure 4.23 shows the results of segmentation accuracy for the number of frames. For each frame, the segmentation accuracy is higher than previous work, which illustrates that the hybrid 3D CNN can provide the higher segmentation accuracy for any number of frames in a stored video.



(b)



(c)

Figure 4.23 (a) (b) (c) Segmentation Accuracy vs. Number of Users

Table 4.12 Statistical Analysis for Segmentation Accuracy

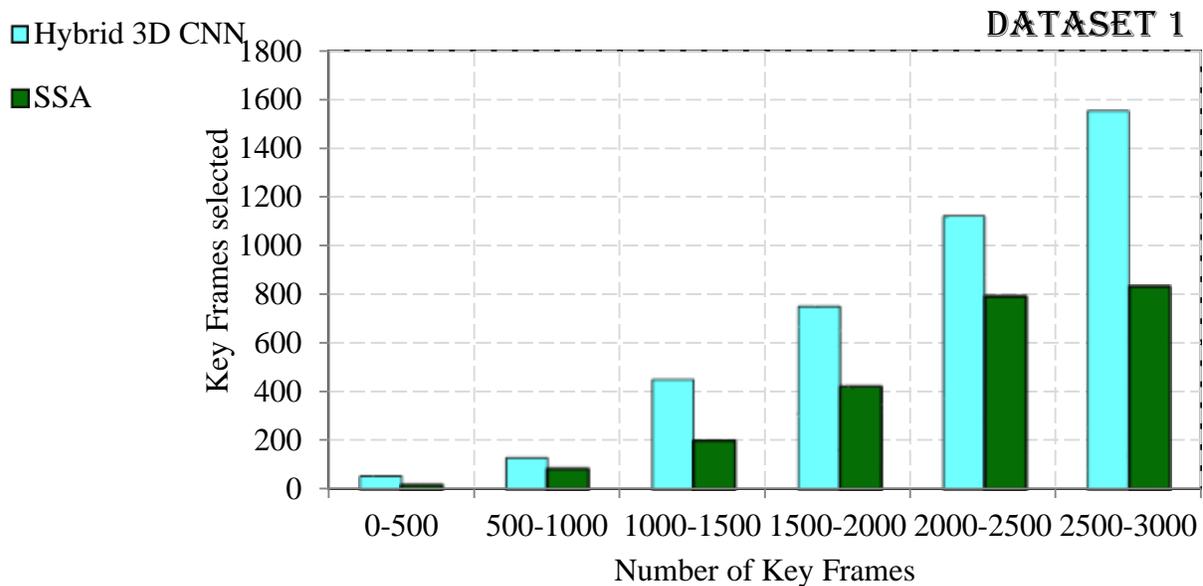
Datasets	Hybrid 3D CNN	SBQ
AVA dataset	0.959	0.8
YouTube 8M segments	0.961	0.85
KTH dataset	0.963	0.865

4.5.3.9 Key Frame Selection

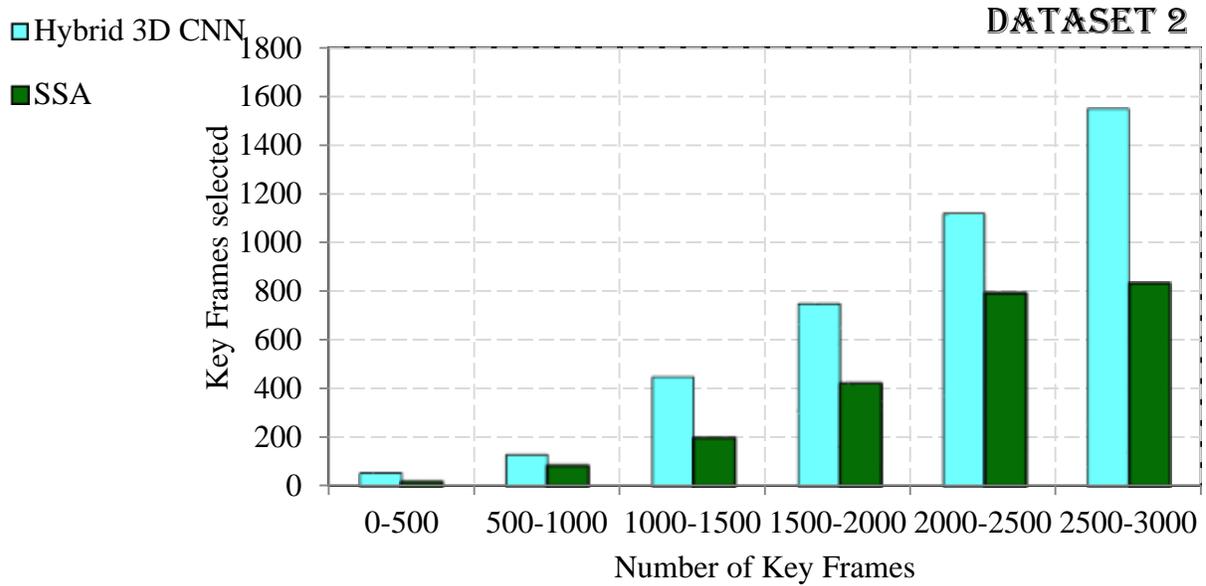
It is the selection process i.e. detects the key frames from video shots. In MF3D-CBVR, RE and SRE are used and tested for three datasets.

$$\text{Key frame selection} = \frac{\text{Selected key frames}}{\text{Total number of frames}} \quad (4.73)$$

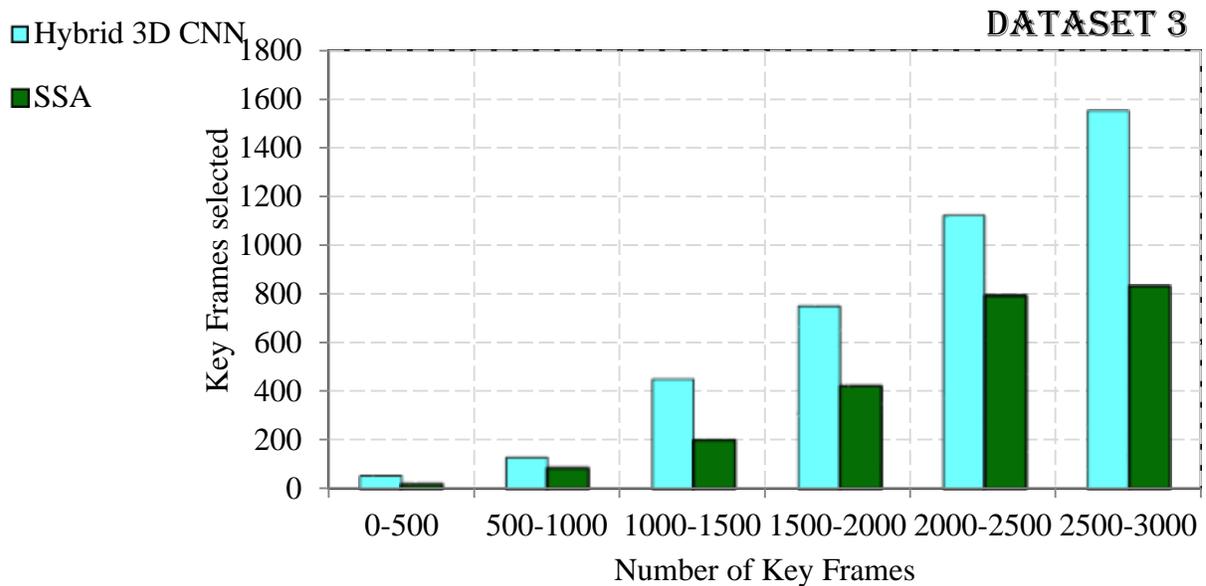
For example, if total number frames are 1500 then the selected key frames are 374.



(a)



 **PHD PRIME** (b)



(c)

Figure 4.24 (a) (b) (c) Key Frames Selected vs. Number of Key Frames

The key frame selection is the important method for CBVR. This considers the motion, time and temporally similar features. Key frames selection process reduces the overhead of relevant videos retrieval. In general, the key frame is the unique point for controlling or changing any object over the time. To extract such key frames, we must need of dynamic and specific methods for it. Table represents the statistical analysis of the segmentation accuracy for different number of key frames.

Table 4.13 Statistical Analysis for Segmentation Accuracy

Datasets	500-1000 frames		1500-2000 frames		2500-3000 frames	
	Hybrid 3D CNN	SKFS	Hybrid 3D CNN	SKFS	Hybrid 3D CNN	SKFS
AVA dataset	150	85	780	468	1500	900
YouTube 8M segments	158	92	785	478	1540	910
KTH dataset	155	99	795	482	1580	915

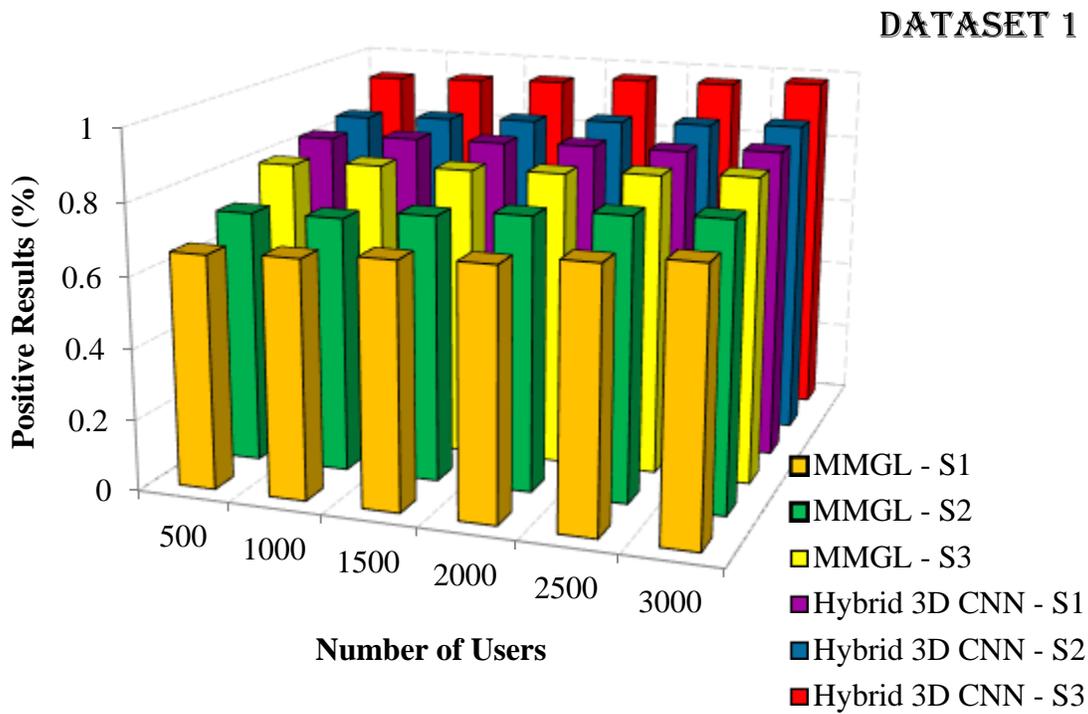
4.5.3.10 Positive Results

It is the measure of number of responses obtained for the submitted queries are most relevant to the query image. It is defined by

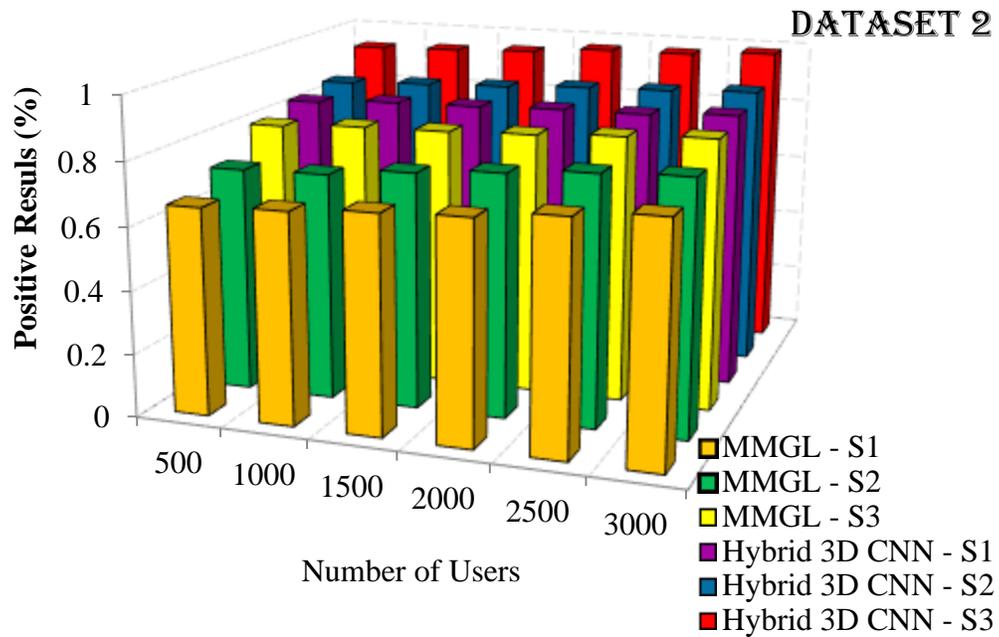
$$\text{Positive Results} = \frac{N(Q_i)Rr_v}{TN(Q_i)P} \quad (4.74)$$

Where $N(Q_i)Rr_v$ is the number of queries received relevant videos and $TN(Q_i)P$ is the total number of queries processed. For example, if 500 queries are requested to retrieve videos and 499 queries of users have obtained relevant results, then the positive results value of K is 99.8%.

The positive results show that the performance of relevant videos processing for the given number of queries. In the proposed work, the positive results are higher due to higher retrieval accuracy, fast content search, reduced memory occupation and work well with the huge amounts of data (more than 100 positive results) for a given query. An improvement of the positive results show the effectiveness of the multi-features extraction method gives the better result with respect to the number of user's queries. Figure 4.25 depicts the performance of the positive results with respect to the number of users for different scenarios.

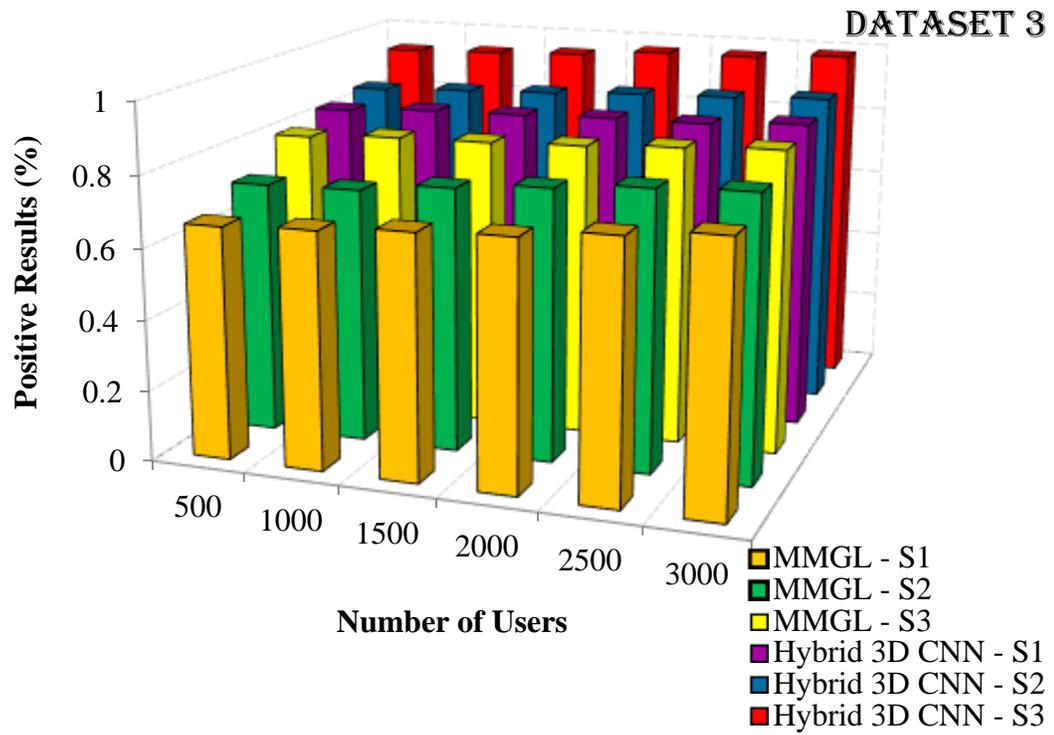


(a) YOUR RESEARCH PARTNER



PHD PRIME
YOUR RESEARCH PARTNER

(b)



YOUR RESEARCH PARTNER

(c)

Figure 4.25 (a) (b) (c) Positive Results vs. Number of Users

Table 4.14 Statistical Analysis for Positive Results

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL	Hybrid 3D CNN	MMGL
AVA dataset	93.45	88.12	94.65	89.15	98.89	90.65

YouTube 8M segments	93.75	88.45	94.68	89.4	98.9	90.7
KTH dataset	93.78	88.56	94.7	89.7	99	90.68

4.5.4 Obtained Output

The performance of this proposed 3D-CBVR is evaluated by measuring different metrics in three different scenarios. In simple, this process can be illustrated as user query based result retrieval system. According to the given query, relevant video are obtained and ranked in terms of most accurately matched videos. User given query is subjected to sequential processing and it matches with the training data for obtaining related 3D videos. The relevant videos are listed along with the length of each video i.e. their duration. Most relevant video is represented by providing higher number of stars and gradually the numbers of stars are minimized with respect to the relevancy of given query to that of the particular video.

Confusion matrix is computed for the proposed and the previous work for a scenario 3. It makes the prediction with respect to the outcome values i.e. the number of correct predictions for each category. Table represents the performance analysis of confusion matrix for the Hybrid 3D CNN and MMGL, respectively.



PHD PRIME
YOUR RESEARCH PARTNER

Table 4.15 Confusion Matrix for Hybrid 3D CNN (Scenario 3)

1	94.00	0.00	0.00	0.00	0.00	0.00	5.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	1.24	0.00
2	0.00	95.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00	0.00	0.00
3	1.45	3.12	97.45	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00
4	0.00	0.00	4.15	97.48	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	3.14	
5	2.35	1.35	0.00	6.42	97.55	3.14	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	
6	4.15	0.00	6.42	0.00	4.15	97.56	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	5.42	
7	6.42	0.00	0.00	6.42	0.00	0.00	97.84	0.00	4.15	0.00	0.00	0.00	0.15	2.35	1.35	0.15	2.35	1.35	1.45	
8	0.15	2.35	1.35	2.35	1.35	0.00	0.00	98	0.00	4.15	1.35	1.35	4.15	0.15	2.35	1.35	4.15	0.15	2.35	1.35
9	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.36	98	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.15	2.35	1.56	
10	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	98.2	1.35	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	
11	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.78	4.15	0.00	98.2	0.00	0.00	1.35	2.35	4.15	1.35	2.35	4.15	1.06

12	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	2.31	98.5	0.00	4.15	2.35	4.15	0.15	2.35	1.35	2.35
13	4.15	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	98.7	5.47	0.15	0.15	2.35	0.00	4.15	2.35
14	0.00	6.42	0.00	2.35	1.35	0.00	2.35	1.35	0.00	1.47	3.47	0.00	7.98	98.8	0.00	0.00	0.15	2.35	0.00	0.00
15	3.14	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.56	6.47	7.98	99	1.35	4.15	0.15	2.35	2.47
16	2.35	1.35	0.00	1.02	4.15	0.00	0.15	2.35	0.00	0.00	0.15	2.35	0.00	0.00	5.78	99.1	1.35	4.15	2.35	2.35
17	4.15	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	99.2	0.00	0.15	2.35
18	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	0.00	1.03	0.00	99	0.00	0.15
19	0.00	0.15	2.35	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	1.35	4.15	0.15	99	0.00
20	4.15	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	2.00	3.45	1.36	0.00	99.9	0.15	2.35	1.35	2.35	99
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20



Table 4.16 Confusion Matrix for MMGL (Scenario 3)

1	54.00	0.00	0.00	0.00	0.00	0.00	5.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	1.24	0.00
2	0.00	45.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00	0.00	0.00
3	1.45	3.12	58.75	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00
4	0.00	0.00	4.15	57	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	3.14

5	2.35	1.35	0.00	6.42	58.55	3.14	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	
6	4.15	0.00	6.42	0.00	4.15	57.89	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	5.42
7	6.42	0.00	0.00	6.42	0.00	0.00	62.53	0.00	4.15	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.15	2.35	1.35	1.45
8	0.15	2.35	1.35	2.35	1.35	0.00	0.00	78.58	0.00	4.15	1.35	1.35	4.15	0.15	2.35	1.35	4.15	0.15	2.35	1.35
9	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.36	79	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.15	2.35	1.56	
10	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	82	1.35	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	
11	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.78	4.15	0.00	82.2	0.00	0.00	1.35	2.35	4.15	1.35	2.35	4.15	1.06
12	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	2.31	78.5	0.00	4.15	2.35	4.15	0.15	2.35	1.35	2.35
13	4.15	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	88.7	5.47	0.15	0.15	2.35	0.00	4.15	2.35
14	0.00	6.42	0.00	2.35	1.35	0.00	2.35	1.35	0.00	1.47	3.47	0.00	7.98	88.8	0.00	0.00	0.15	2.35	0.00	0.00
15	3.14	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.56	6.47	7.98	89.7	1.35	4.15	0.15	2.35	2.47
16	2.35	1.35	0.00	1.02	4.15	0.00	0.15	2.35	0.00	0.00	0.15	2.35	0.00	0.00	5.78	88.7	1.35	4.15	2.35	2.35
17	4.15	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	89	0.00	0.15	2.35
18	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	1.03	0.00	86.7	0.00	0.15	
19	0.00	0.15	2.35	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	1.35	4.15	0.15	89.5	0.00
20	4.15	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	2.00	3.45	1.36	0.00	99.9	0.15	2.35	1.35	2.35	96.7
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

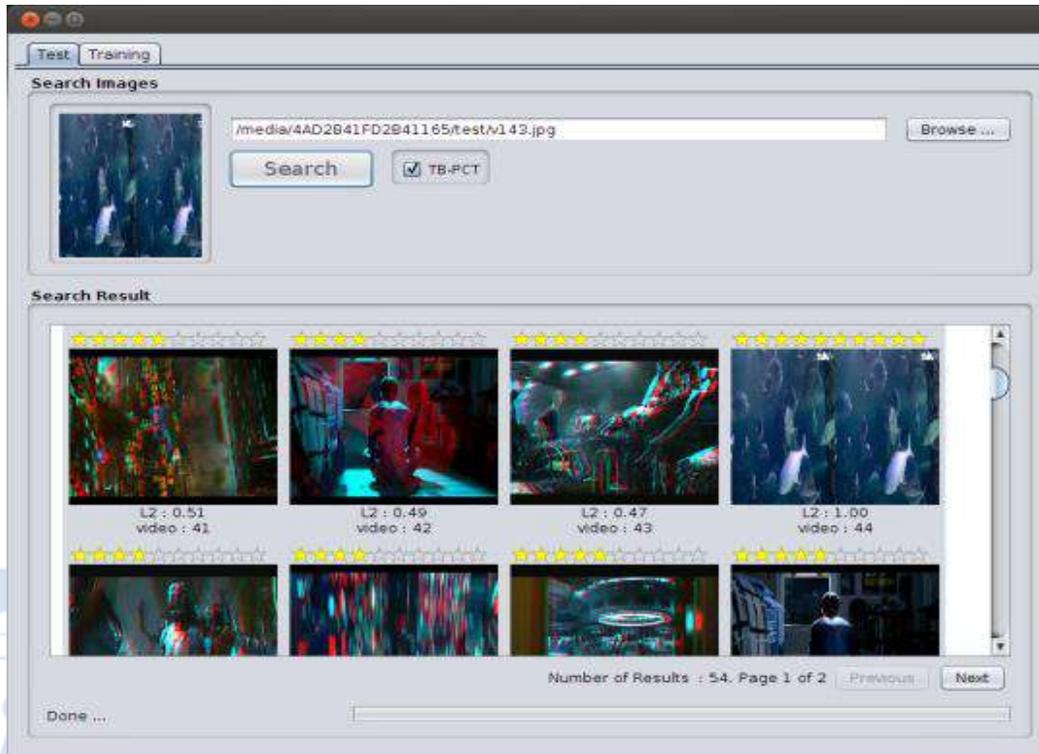


Figure 4.26 Obtained Experimental Result

Figure 4.26 represents the experimental result obtained on the user input query; it provides number of relevant and matched 3D videos. In this retrieval process, the highest ranked 3D videos are retrieved first. Map Reduce function allows the parallel processing, which splits the input data parallel so the processing time of retrieval is reduced. The designed search engine is a capable framework that withstands same level of accuracy and reduced computation time even if multiple user queries are arrived.

4.6 CONCLUSION

We have proposed a novel mechanism for 3D video retrieval process, which is processed on Map Reduce framework in Hadoop environment. It effectively supports parallel processing and improves the processing time of retrieval process. Our proposed 3D CBVR mainly concentrated on faster retrieval of videos with higher positive results for all user submitted queries. The novel 3D CBVR in map reduce includes key frame extraction, feature extraction, codebook generation and similarity matching. Initially, key frame extraction is processed by Gaussian Filtering Technique, which effectively extracts the key frames based on probability of occurrence of pixel values present in frames. Further, feature extraction involves with shape, color and texture features which are extracted for producing accurate result in forthcoming process. Voronoi and the 3D CNN feature extraction methods where SURF and SIFT descriptors are applied. Then, CT is applied for codebook generation which solves the problem of noise in the generated visual words. Finally, similarity matching is performed using soft-weighting scheme and L_2 distance which is responsible for producing relevant results with respect to user query. On completion of these processes we achieve accurate results, which overwhelm the problem of time complexity and storage complexity. Storage and time complexity

problems are significantly preserved by the use of Hadoop Map Reduce framework, which perfectly increase the accuracy rate even if multiple user queries are given. The implementation result is evaluated by considering three scenarios based on the master-slave nodes involvement in Hadoop processing. Our comparative results on accuracy and positive response for given number of queries are shown to prove betterment of this proposed research work.



CHAPTER 5

3D HOLISTIC CNN WITH MAP SHUFFLE REDUCE PARADIGM FOR CBVR



5.1 INTRODUCTION

Multimedia searching and browsing is one of the popular area in internet that is extensively played by people at all ages. Due to the use of people all over the world, the demand for video retrieval has been increased. The main goal of retrieval concept is to achieve relevant result for the given query, so visual perceptions are taken into considerations. A video file involves representations of audio, text, objects and motions. Each video is comprised of ^[105] (i) audio, text and video features, (ii) motion and color histograms and (iii) other features like color, texture, edges, shape, etc.,. The

performances of traditional video retrieval search engines were introduced with novel techniques, methods and mechanisms to improve the quality of video retrieval.

3D images and videos play a major role at present which gives more reality in visualization for viewers. 3D images may involve noise, hence it need to be preserved from corners, edges and also significant features present in it. For extracting a 3D video, temporal analysis is considered to be more significant, since the motions and objects in each frame shows minor or major differences. A video is mainly composed of spatial-temporal composition of visual features that are essentially required for analyzing. In content based video retrieval, to obtain desired videos, image processing involves several processes like indexing, feature extraction, key frame extraction, similarity matching, noise minimization and preprocessing. Video indexing includes video parsing, abstraction and content analysis. Shot boundary detection is performed in video parsing for detecting the boundaries present in consecutive shots. This boundary detection process is mainly based on six groups of techniques such as Pixel based, statistics based, histogram based, transform based, edge based and motion based. Video abstraction is a significant process which involves key frame extraction ^[110]. Feature extraction process also deals with integration of different feature for obtaining efficient results. Feature extractions are mainly classified into four categories such as metadata based, text based, audio-based and content based ^[112]. Using combination of more than one feature is to obtain accurate results while performing similarity matching process.

Most of the video retrieval systems are composed of two modes of working such as online mode and offline mode. Online mode is for customers who gives query and

retrieve relevant images / videos. Offline mode is for administrator who uploads videos / images in the database and manages the overall retrieval system. The performances evaluated over online mode are runtime and retrieval accuracy to validate the system's achievement with previous research works. Content based video retrieval process is also based of identification of human face present in videos. A video retrieval system is designed to be scalable only if it is enabled to handle larger amount of data storages ^[111]. Scalability improvising also needs minimization of cost and resource consumption.

5.2 RESEARCH METHODS

5.2.1 Denoising Methods

For image denoising, there are two different techniques are proposed, including Spatial Domain Filtering and Transform Domain Filtering. The spatial domain filtering can be implemented in image at pixel level and processing the filter on each pixel.

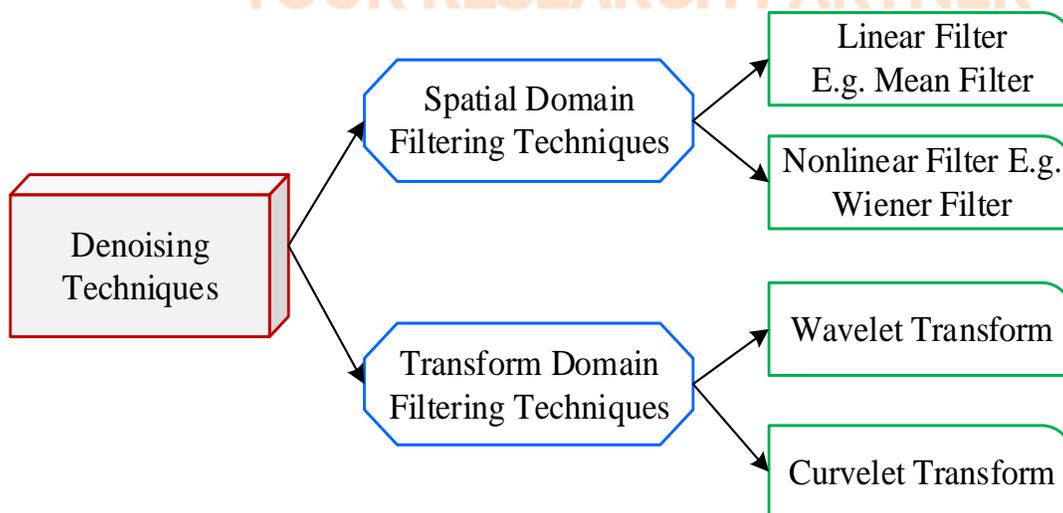


Figure. 5.1 Denoising Techniques

. The transform domain filtering can be used to convert the image into frequency domain and then process the image. The spatial domain filtering is further classified into two techniques: Linear Filter and Mean Filter, or Nonlinear Filter and Wiener Filter. The transform domain filtering is further classified into two classes: Wavelet Transform and Curvelet Transform. The overall denoising techniques can be seen in Figure.5.1 [Biswas et al., 2017, Dhivyaprabha et al., 2018, Singh et al., 2019, Iqbal et al., 2019]

A. Spatial Domain Filtering

In this filtering, the spatial domain process is applied on the image pixels, and implements the manipulation in pixel-by-pixel. It is a mask for filtering that shifts from one pixel to another by doing the several operations. It will remove the noise by smoothing the image. As mentioned in the earlier, the spatial domain filtering is classified into two classes of filters such as linear filter and non-linear filter. The mean filter and wiener filters are most widely used linear filtering techniques, and the median filter is a class of non-linear filtering. In the mean filter, the mean (average) value has taken from the computation of neighbors and the centre pixel values, which is computed in the $N \times N$ size. After the calculation of the centre pixel value, it is computed by follows:

$$Y(M, N) = \text{Mean} \{X[i, j], [i, j] \in w\} \quad (5.1)$$

Where w are the pixel positions in the neighborhood

Then we describe the Wiener filter, which is used for filtering the consistent pixel values that define the constant power additive noise. This filtering technique is used for adaptive filtering for the image in pixel-wise. By computing the pixels in neighborhood, two variations are computed including Mean and Standard Deviation. The mean is computed by follows:

$$\mu = \frac{1}{N \times M} \sum_{(n_1, n_2) \in \eta} a(n_1, n_2) \quad (5.2)$$

The standard deviation is given as

$$\alpha^2 = \frac{1}{N \times M} \sum_{(n_1, n_2) \in \eta} a^2(n_1, n_2) - \mu^2 \quad (5.3)$$

Where η represents the $N \times M$ of the current pixel. With this estimation, the pixel-wise Wiener filter was applied over the denoised image, which is computed by follows:

$$b(n_1, n_2) = \mu + \frac{\alpha^2 - v^2}{\alpha^2} (a(n_1, n_2) - \mu) \quad (5.4)$$

Where v^2 represents the noise variance.

Median filter is working by the pixels in the window that are sorted in the Ascending Order. The median value of the $N \times M$ image is changed based on the central pixel values. Then it is defined by follows:

$$Y(M, N) = \text{Median} \{X[i, j], [i, j] \in w\} \quad (5.5)$$

Where w is the pixels of neighborhood.

B. Transform Domain Filtering

In this section, we give details of three transform techniques, including wavelet transform, curvelet transform and thresholding.

(i). *Wavelet Transform*

In this transform, wavelets are used for noise removal. It is a very essential in denoting the non-linear signals. The wavelet analysis of the un-decimated wavelet transforms has been considered over the un-dimensional signal, which aims to eliminate the noise. It is a traditional method for denoising. The wavelet function is represented by following:

$$\psi(a, b)(x) = \left(\frac{a}{\sqrt{a}}\right) \times \psi\left(\frac{x-b}{a}\right) \quad (5.6)$$

Where ψ is the Mother Wavelet, and a represented Dialation Parameter and b represents Translation Parameter

The continuous wavelet transform is represented by the following:

$$\psi(a, b) = \int_{R^2} f(x)\psi_{a,b}(x). d(x) \quad (5.7)$$

Then reconstruct $f(x)$, and hence the Inverse of Transform is completed by following.

$$f(x) = \int \int_{-\infty}^{\infty} w(a, b). \psi(a, b)(x) \frac{da.db}{a^2} \quad (5.8)$$

Wavelet transformation is the significant technique that can be used for the non-linear signals representation. It is operated by the original image, which is noisy in nature so it decomposes into two elements: Time domain and Frequency domain. Based on the original noisy image, its decomposition is follows: Low (L), High (H), and frequency

bands are known as LL, LH, HL, and HH. Furthermore, the LL subsample is decomposed into four subsamples at level two and so on.

(ii). Curvelet Transform

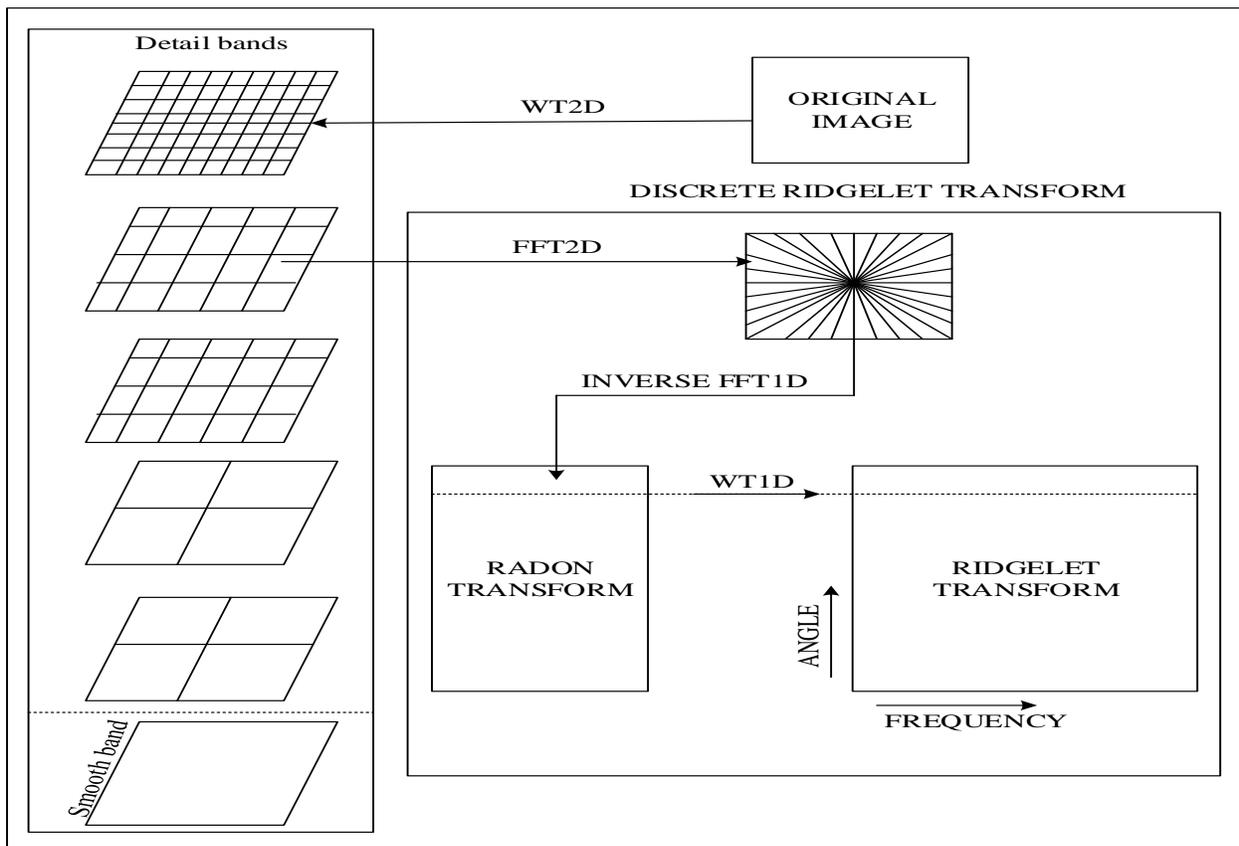


Figure 5.1 (b) Block Diagram for Curvelet Transform

It mitigates the disadvantage of the wavelet transform i.e. it does not remove the noise over the image curves and it results the loss details. Ridgelet transform is a solution to solve these problems in the curvelet transform, which is finished by conversion it into Radon Transform. In curvelet transform, the original image is decomposed into number

of sub-bands, which is followed by each sub-band spatial partitioning. One of property in the ridgelet transform is called as the scaling or support interval, which is based on the Anisotropy Scaling Relationship [Wolterink et al., 2017, Thakur et al., 2015, Schaap et al., 2008, Lin et al., 2013., Candes et al., 2002].. Figure 5.1 (b) shows the Curvelet Transform.

Mathematical calculations for Curvelet Transform are follows:

$$width = length^2 \quad (5.9)$$

In the 1st generation of the curvelet transform by multiscaling ridgelet, the curve is splits into number of blocks and the sub-blocks, which are approximated into the straight line and implement the ridgelet analysis. The process involving in the curvelet decomposition are follows:

$$f \mapsto (P_0f, \Delta_1f, \Delta_2f, \dots) \quad (5.10)$$

f is the sub-band decomposition, P_0 is the sub-band filters, Δ_s is sub-bands, $s \geq 0$, $\Delta_s f$ consists of the details of 2^{-2s} wide. The smooth window is the $W_Q(x_1, x_2)$, which is localized in dyadic squares, and it is given by follows:

$$Q = [k_1/2^s, (k_1 + 1)/2^s] \times [k_2/2^s, (k_2 + 1)/2^s] \quad (5.11)$$

This square values are re-normalized into unit scale values, which is given by follows:

$$g_Q = T_Q^{-1}(w_Q \Delta_s f), \quad Q \in Q_s \quad (5.12)$$

$$(T_Q f)(x_1, x_2) = 2^s f(2^s x_1 - k_1, 2^s x_2 - k_2) \quad (5.13)$$

In above equation, T_{Qf} is the renormalization operator, after the re-normalization the ridgelet transform is computed by follows:

$$\alpha_{\mu} = \langle g_Q, p_{\lambda} \rangle \quad (5.14)$$

(iii). Thresholding Technique

It is a type of transform domain based filtering technique that further classified into two techniques including Hard Thresholding and Soft Thresholding. The main of the thresholding technique is to eliminate the unwanted noise signals. The hard thresholding helps to eliminate all the pixels values greater than the threshold value, whereas the soft thresholding reduces the range of noise intensity into zero, which is defined by following:

$$y(t)_{Hard} = \begin{cases} x(t) & |x(t)| \geq T \\ 0 & |x(t)| < T \end{cases} \quad (5.15)$$

$$y(t)_{soft} = \begin{cases} sign(x(t)).(|x(t)| - T) & |x(t)| \geq T \\ 0 & |x(t)| < T \end{cases} \quad (5.16)$$

Where T denotes the threshold value, x and y are the input and output coefficients in the transform domain

In the wavelet domain, threshold values are computed by VisuShrink method, which is based on the universal threshold value, which is given in follows:

$$T_w = \sigma \sqrt{\log(N)} \quad (5.17)$$

Where σ presents the noise variance, and N is the image size

In the curvelet domain, threshold value is computed by value of $3 \cdot \sigma$ and $4 \cdot \sigma$ for coarse scale and fine scale elements

$$T_c = 3 * \sigma + \sigma * (s == \text{length}(C)) \quad (5.18)$$

Where C represents the decomposed images size, and $s = 2$ to length of C [Mustafa et al., 2014, Chithra et al., 2017, Kamezawa et al., 2014, Madhura et al., 2017]

5.2.2 Feature Clustering Methods

Clustering is a process of categorizing a set of data into homogeneous clusters, which is also called unsupervised classification. The data elements in each cluster should be similar. Using centroid, clustering can be very useful between cluster homogeneity, and cluster separation. Thus, the similarity between individuals in the same cluster must be small and should be high between the different clusters. Distance measure is used to measure the similarity between two points. Formally, the ultimate goal of data clustering is to partition a set of unlabeled data samples $D = \{d_1, d_2, \dots, d_n\}$ into k clusters. Each data sample is characterized by a feature vector $f = \{f_1, f_2, \dots, f_x\}$, where x denotes its dimension. A large number of clustering methods are developed, which can be illustrated in figure.

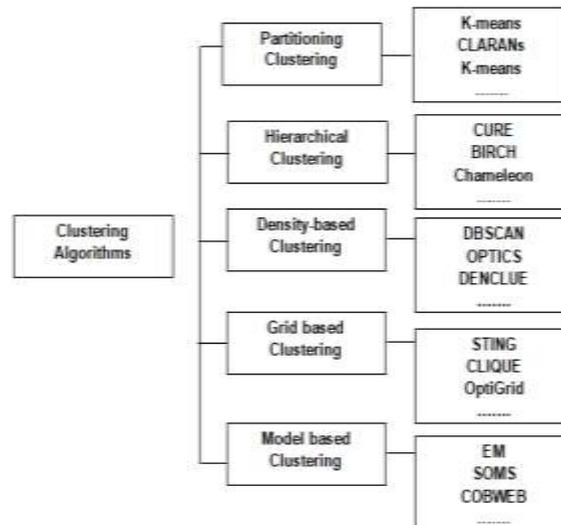


Figure 5.2 Clustering Algorithms

- Partitioning Clustering:** In this type, data divided into many clusters. To reach this objective, the clustering function is used. Initially, groups (data clusters) are formed and organized in order to have the final clusters. A simple groups for a set of data items or objects into sub-groups by moving the data objects from one cluster and another cluster group. K-means, CLARA, PAM clustering algorithms are some of the partitioning algorithm. Construct a partition of a database D of objects n into a set of k clusters

Given a k clusters, here find a partition of k -clusters that optimizes the chosen partitioning criterions: global optimal and heuristic methods

Example: k-means clustering

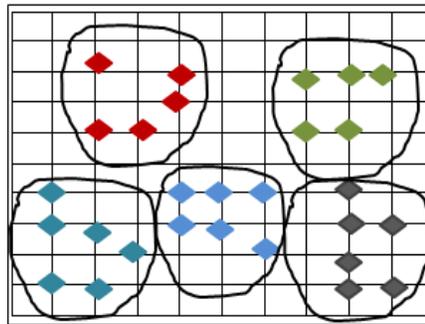


Figure 5.3 K-Means Clustering

- **Hierarchical Clustering:** This type of clustering is simple one and group objects into a tree of clusters. The algorithm is divided into two parts: Divisive (top to bottom) and agglomerative (bottom to top). The first part of the algorithm is puts all the data into a single cluster, then it divided hierarchically until find the final clusters. The second part of the algorithm is puts each object (data sample) of the database in one cluster, then merges them until forms the last clusters. Some of the major hierarchical clustering techniques are CHAMELEON, BIRCH, CURE, ROCH, etc.

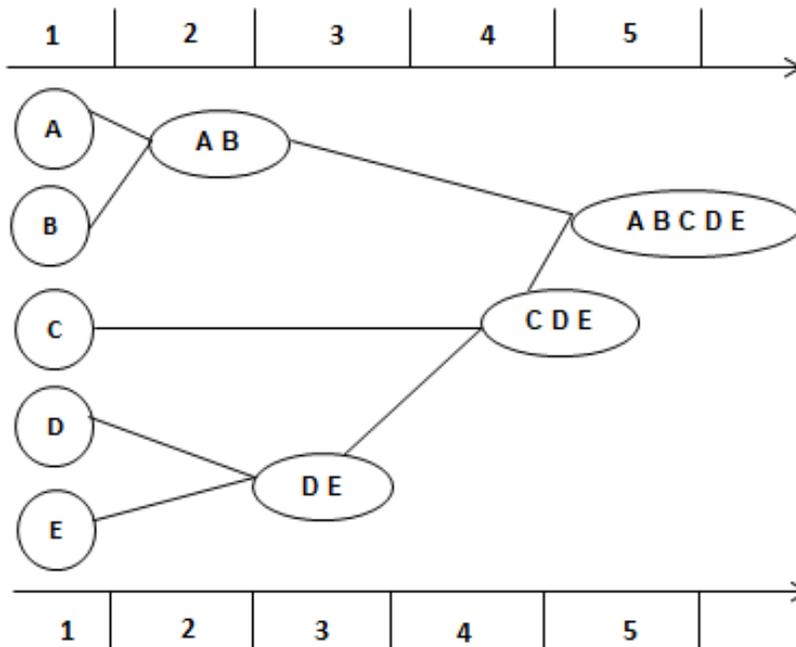


Figure 5.4 Hierarchical Clustering

- Moreover, these methods are not suitable for large scale processing. The disadvantages of the clustering algorithms are argued in below:
 - ✚ Time Complexity: The clustering processing time consuming and the time required for clustering is $O(n^2 \cdot \log n)$, where n denotes the total number of records in dataset
 - ✚ Space Complexity: Most of the clustering algorithms need huge size for storing a similarity matrix of size $O(n^2)$
 - ✚ Syndrome Irreversible: Original datasets can't be tracked since all the actions of step-up or a split down cannot be reversed

- **Density based clustering:** This type of clustering is effective and simple one. The algorithm in this type are defined by density. To form a cluster, the objects are classified based on the regions density. This type of algorithms able to discover classes of arbitrary shapes and remove noisy objects. The major features of density based clustering method are: handle noise, one scan, need density parameters as termination criterion, and discover clusters of arbitrary shape. Several clustering algorithms are DBSCAN, DENCLUE, CLIQUE, OPTICS, etc.

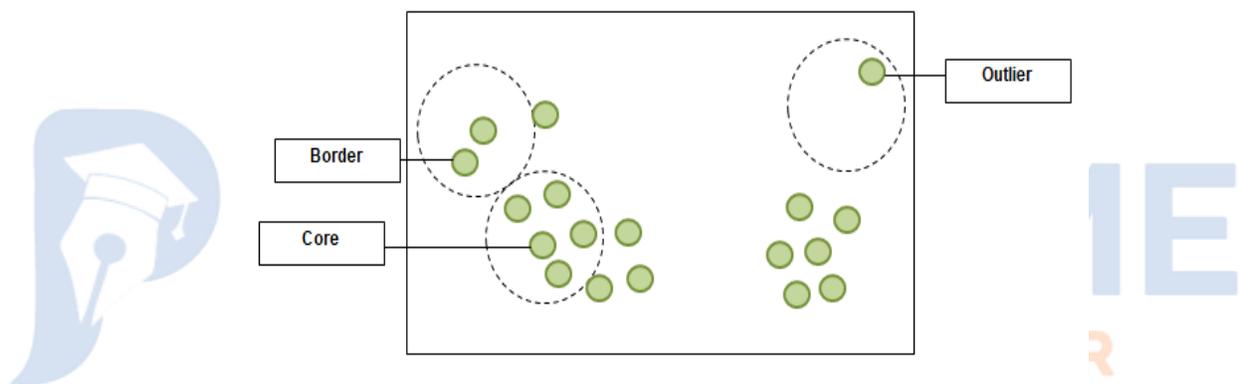


Figure 5.5 Density based Clustering

Figure 5.5 shows the density based notion of cluster. Here discovers clusters of arbitrary shape in spatial databases with noise

- **Grid based Clustering:** This type of clustering is applied on the grid where data are divided into grid of objects. Grid based clustering uses multi-resolution grid structure. Several interesting clustering methods are illustrated in follows: STING (a Statistical Information Grid Approach), Wave cluster (wavelet method), CLIQUE, etc.

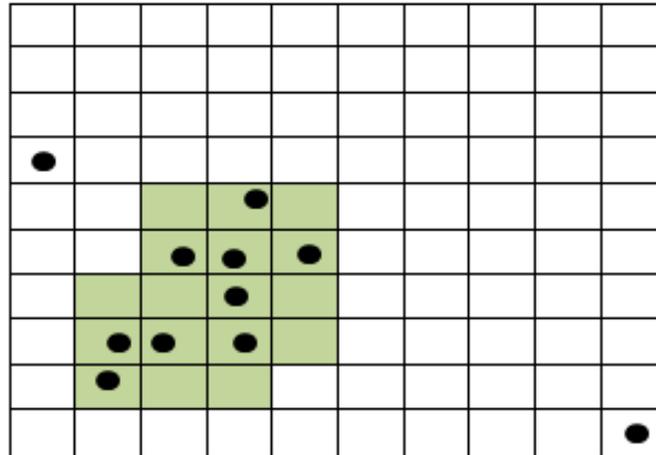


Figure 5.6 Grid based Clustering

- Model based Clustering:** In this type of clustering, hypothesis is used which is generated by a probability distributions. These methods aimed to remove a model assumption for each cluster. Finally, determine the best fit of the data to the model. AI and statistical approach is attempt to optimize the fit between the data and some methemathical models are AutoClass, Expectation Maximization, etc. one of the typical method is machine learning approach which are CLASSIT, and COBWEB, etc. and the last method is neural network method (Self-Organizing Feature Map).

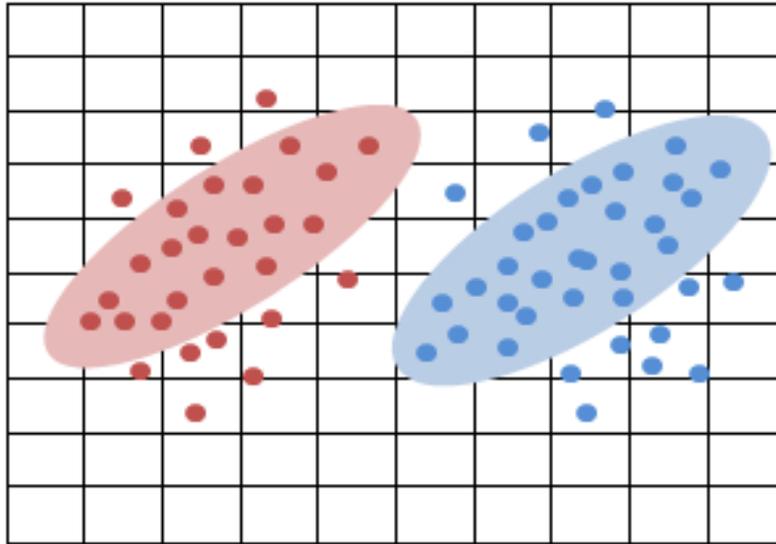


Figure 5.7 Model based Clustering

In this work, we give special attention to DENCLUE clustering algorithm. This is the family of density based clustering. Its improve the data efficiency in terms of clustering. Its effectively find clusters of arbitrary shapes and also detect the noisy data samples.

Proposed DENCLUE Algorithm

This is designed to cluster the similar data like pollution data, traffic data, parking data and city wather data. To perform clustering here we proposed *stability degree based DENCLUE clusering algorithm*.

In this section DENCLUE (DENsity based CLUstEring) clustering algorithm is proposed and discussed in detail. It shows, how it works as well as the input parameters takes. DENCLUE is a good candidate clustering algorithm for big data. This algorithm calculates

sum of influence functions of all of the data points. The influence function is defined by the impact of a data point within its neighbourhood.

The influence function between two points x and y is derived by,

$$f_{Gauss}(x, y) = \exp \frac{d(x,y)^2}{2\sigma^2} \quad (5.19)$$

Where $f_{Gauss}(x, y)$ defined as the distance between two points Gaussian function since many of the exist influence functions derived from Gaussian function. $d(x,y)$ is an euclidean distance between points x and y and σ represents the radius of point x neighbourhood. The density function is expressed as follows:

$$f_D(x) = \sum_{i=1}^N f_{Gauss}(x, x_i) \quad (5.20)$$

Where D represents the set of points on the database and N represents its cardinal

When a influence function defined, then density attractors can be determined as clusters. Density attractors means that the local maxima of the overall density function using Hill Climbing Algorithm. The DECLUE algorithm can be described by arbitrary shape using a simple equation with kernel density functions. The maximum of the density function is presented by

$$x = x^0, x^{i+1} = x^i + \delta \frac{\nabla f_{Gauss}^D(x^i)}{\|\nabla f_{Gauss}^D(x^i)\|} \quad (5.21)$$

The hill climbing process ends with when $f^D(x^k) < f^D(x^{k+1})$ with $k \in N$, then we consider $x^* = x^k$ as a new density attractor.

It has several advantages:

- It has solid mathematical foundation
- It has good clustering properties,
- It uses grid cells which produce the cell information about the cells that actually contain points, it uses compact mathematical description for high dimensional datasets that arbitrarily shaped clusters, it organize cells in a tree based access structure. Moreover, DENCLUE produce accurate clusters in datasets.

Pseudo-code for DENCLUE algorithm

Begin

Initialize dataset(d_s), minimum number of objects n , cluster radius r , grid G

1. Take d_s in G whose each side is of 2σ
2. Set the stability threshold S_{th}
3. While Cl_s is not stable
4. Find highly D_s //Find m for highly populated cells
5. If $d(m(c_1), m(c_2)) < 4a$
 - {
 - c_1, c_2 connected
 - else go to step 2
 - }
6. Find d_a using eqn ()
7. Pick p randomly

8. Compute the local 4σ density
9. Randomly pick $p+1$ by p
10. If $(\text{den}(p) < \text{den}(p+1))$ climb, then put $p, p+1$ within $(\sigma/2)$
11. Connect d_a based cluster
12. Data values assigned to clusters
13. End

The description of this DENCLUE algorithm are given follows: In this algorithm the dataset, cluster radius, minimum number of objects are considered as a input. In first step, dataset considered in the grid whose each side is of 2σ . Then determine highlyly densed cells in the grid. The mean value of cube 1 and 2 are less than value of $4a$. If the condition is satisfied, two cubes $c1, c2$ are connected where the highly populated cells will be considered in determining clusters. Then the density attractors determined using a Hill Climbing Algorithm. Randomly pick point p and compute the local density 4σ . To pick another point $p+1$, previous density compute. The density of point p should less than $p+1$, then put points within the path to the cluster. Finally connect the density attractor based cluster.

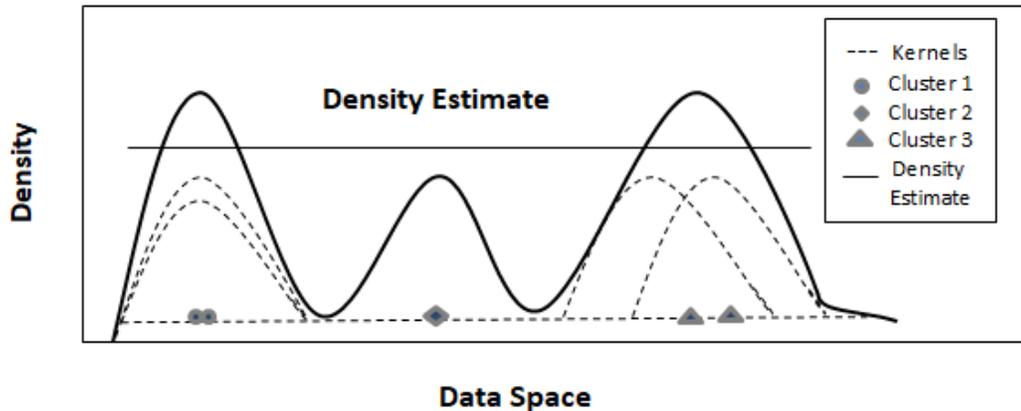


Figure 5.8 Illustration of DENCLUE density concept

5.3 PROBLEM DEFINITION

A content-plus-feature coding scheme ^[106] was proposed for joint compression and improve visual retrieval performances. Feature coding is comprised of three methods (i) block-partition guided feature selection, (ii) Multiple prediction modes and (iii) rate-accuracy optimization. Block –partition Guided feature selection method was used for selecting relevant key points. Prediction mode and quantization parameter is considered for video coding which was tested using a public dataset. Prediction mode includes both Interframe and intraframe which are predicted from static search set. The static set is not applicable for all type of videos and also the static set was low.

The participation of 3D images or videos, involves with the major problem of presence of noise. This noise is removed by using different filtering methods and result a noiseless image that shows better quality. A bilateral filter for faster implementation was

used for minimizing computation complexity that existed due to the Eigen vector construction and histogram formation ^[107]. In this work the authors have optimized the expansion function that has been used for histogram. Diagonal matrix are defined for removing noise, however it removes noise from 3D image and hence noises are extracted only diagonally i.e. some spaces are ignored which may contain noise in it.

In many video retrieval processes, classification algorithm is involved for classifying the relevant and irrelevant video separately ^{[108], [109]}. A video sequence can be represented as bag of regions, which is supported for video classification and action recognition. Unsupervised frame segmentation was performed for extracting regions. Regions are tracked and selected from graph model and then their features are extracted further the frames from training set and testing set is classified using SVM with RBF kernel function. Noises in the frames are not extracted before classification, since it leads to minimization of accuracy in classification. Storage was also one of the major issue in video retrieval concepts, since the size of video files are larger and also the involvement of number of users was also increased. Hence storage complexity issue also needs more attention while designing a video retrieval process. The problems defined in this section are solved in the proposed system which aims to achieve higher accuracy result in 3D video retrieval along with the solution of storage complexity that existed in previous works.

5.4 PROPOSED SYSTEM

This section details the proposed 3D content based video retrieval system along with Hadoop MapReduce environment. In this work Hadoop Distributed File System

servers are involved for managing the stored videos in database. A new video retrieval system is designed to increase the positive results from real-time large datasets. This system focuses more on

- Accurate video retrieval
- Joint similarity
- Multi-view content features

This process is performed by following four sequential processes as,

- Key frame selection
- Key frame denoising
- Feature extraction
- Similarity matching

Each process is performed one after the other and each process is essential for achieving higher performance result. Hence each process is defined with novel procedure that effectively resolves previous problem. For key frame selection, Relative Entropy based Fast Key Frame Selection (REFKFS) is performed by which optimal key frames is chosen from a video for further processing. Next denoising is supported by BM3D filter with Bayesian threshold for reducing Gaussian noises present in each frame that is selected as key frame. Feature extraction deals with extraction of four different features, they are,

- Shape feature
- Color feature

- Texture feature
- Motion feature

Among these features, motion feature plays a significant role, since a video involves with frequent changes in motion of the objects present. This feature is enables to represent the significant temporal information present in the video. On extracting features the frames are matched across the user query using Multi-Featured Light-Weight (MFLW) matching scheme.



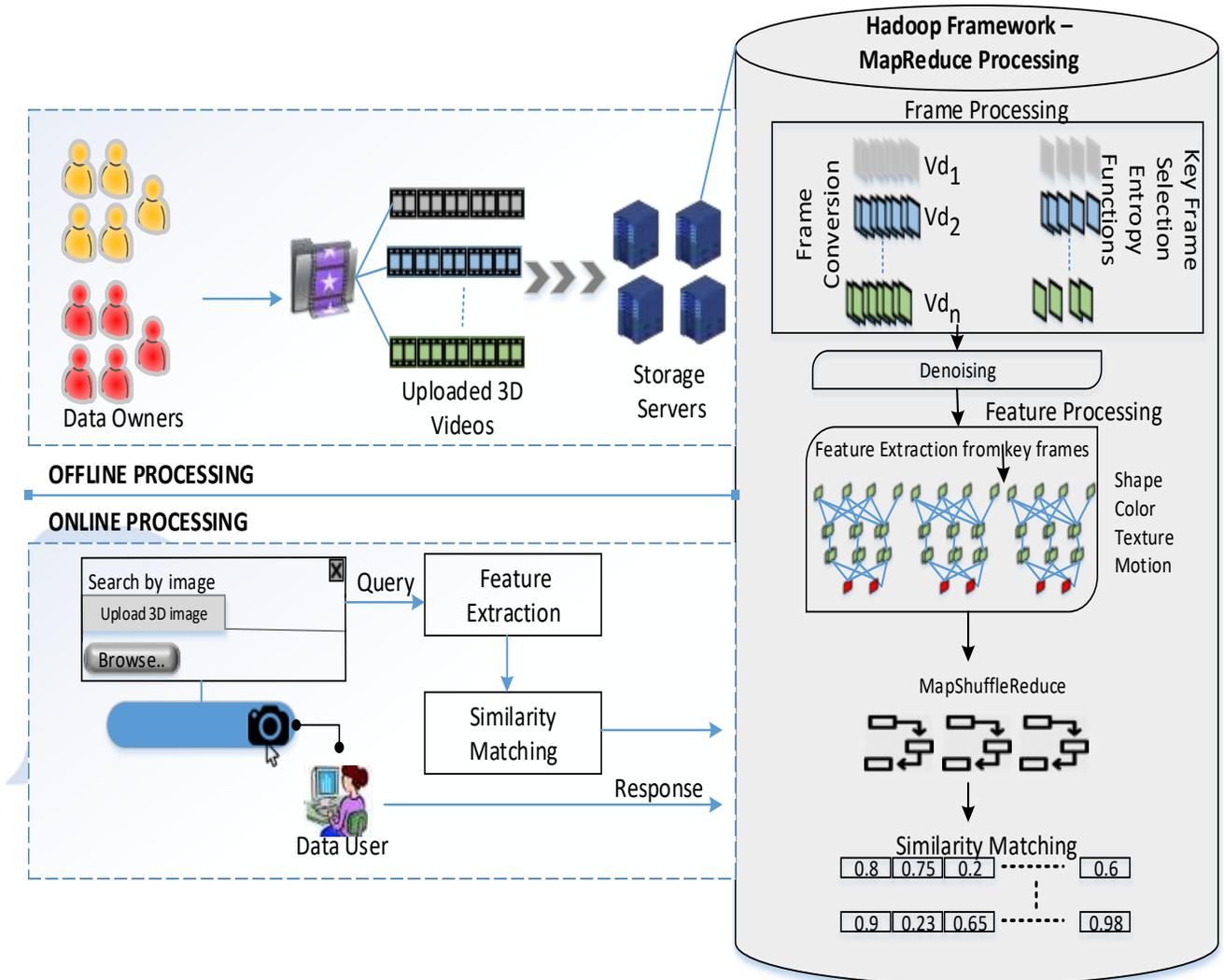


Figure 5.9 Proposed Architecture

These processing are supported in Hadoop environment for achievement of 98% accuracy due to the schemes and algorithms proposed. HDFS is incorporated along MapReduce that operates on the following functions as,

- Map ()
- Shuffle ()
- Reduce ()

The involvement of each process is considered to be significant to achieve higher accuracy; the design of this work is supported for future video retrieval search engine. The proposed HDFS server is defined as a social network similar to YouTube which is comprised with videos and user retrieve relevant videos. The change here is the main focus of this work is to use 3 Dimensional images and videos. Hence the user query is a 3 Dimensional image and the output will be retrieval of 3 Dimensional relevant videos from HDFS server.

The processes handled by Hadoop are denoising, feature extraction and similarity matching. Hadoop MapReduce framework is a recent development which is actively used for solving storage oriented challenges and limitations. Detailed study of Hadoop MapReduce have been discussed in previous sections, from which it can be assured to be chosen as an efficient framework to support 3D video retrieval process at higher speed. Hadoop MapReduce is also enabled to tolerate with the incoming of multiple queries at a time.

Hence the number of internet users has been greatly increased and also their participation in social networks is also tremendously increased. In this work, previous issues and challenges are addressed and solved. Different processes are addressed in this chapter and each process is discussed in following sections. Figure 5.9 depicts the overall proposed framework in HDFS server in which user feeds into a 3D image query.

Algorithm 1: Proposed Framework

Input: QI

Output: Relevant Videos

1: Start // Begin Process

2: DB Video $\rightarrow S_1, S_2 \dots S_N$

3: $S_1 \rightarrow F_1, F_2 \dots F_N$

4: Key frame selection: // Section 5.3.1

 For ($i=1 \dots N$)

$F_i, F_{i+1} \rightarrow$ Calculate DRE and DSRRE

$KF_i =$ Minimum (DRE and DSRRE)

 End For

 Obtain KF_i

 Begin Hadoop operations

5: Denoise (KF_i) // Section 5.3.2

6: Feature Extraction // Section 5.3.3

 For ($i = 1 \dots N$)

 Color (KF_i), Shape (KF_i), Texture (KF_i)

 Motion (KF_i)

 End For

7: Similarity Matching: // Section 5.3.4

$Sim - Value = (QI, KF_i)$

8: Map \rightarrow Pair (Sim-Value, F_i)

 Shuffle \rightarrow Sort (F_i , Sim-Value)

 Reduce \rightarrow Relevant Videos

9: End // Hadoop operations

10: End // Finish process

Admin process is begun, once data owners give video files, here different registered data owners are allowed to provide video to admin for uploading in HDFS. Admin proceeds with partitioning of the 3D videos into shots i.e. a series of frames. This partitioning is held on the basis of shot boundary detection mechanism. From the shots key frames are selected and then the frames are uploaded into HDFS server. The complete processing proposed in this framework is described in the following algorithm.

From the above algorithm, the terms insisted in the algorithms QI denotes the query image, $S_1, S_2 \dots S_N$ is the shots that is obtained from videos and $F_1, F_2 \dots F_N$ are the frames arrived from shots, KF_i is the key frames, DB represents database videos and then Sim -Value defines the similarity value obtained for the frames. Figure 5.9 represents the individual process performed in the retrieval of proposed novel framework. Our proposed framework works on three individual process such as admin process, user process and HDFS process. Data user process is simpler among the three, since its process is just to fed a 3D image as its query. Each user can fed only one 3D image at a time, then in admin process the videos are handled into sequential steps to reach HDFS. Admin process is performed simultaneous when the data owner sends video files to administrator. Admin process supported higher accuracy achievement and HDFS process supports faster processing and efficient matching based on the MapReduce in HDFS. Here all the processes are held only over online, since the time-to-time 3D videos and images are considered. Therefore the proposed novel approaches in each process are efficient and also applicable for real-time video retrieval environment. Hereby the entire

process follows this step-by-step procedure to achieve higher accuracy and also faster results. Each process is elaborated in the forthcoming sections with the formulations and figures that give clear idea over this work and also solutions for the previously existed retrieval procedure. Our contribution of using Hadoop is newer in the field of 3D based image processing.

5.4.1 Key Frame Selection

Key frame selection is a significant process involved in this proposed work, by which the most required content of the video are taken in account. The number of frames and number of shots are dependent on the size of the file. As per the increase in size of the video, the number of shots and frames also increases. This work has proposed a novel key frame selection approach called Relative Entropy based Fast Key Frame Selection (REFKFS). As per the name of this approach, it involves Relative Entropy and Square Root of Relative Entropy. The distance estimation between probability distributions is performed using Relative Entropy and an additional entity is measured for determining minor modification in the frames, which is estimated by means of Square Root of Relative Entropy. In this work, the frames are organized temporally in the direction of left towards right. As per this arrangements the first and last frames is preferred as key frames and then the intermediate frames are validated using Relative Entropy based Fast Key Frame Selection approach. For this approach the values of Relative Entropy and Square Root of Relative Entropy are determined from the following formulations,

$$D_{RE} (F_i, F_{i+1}) = \sum_{k=1}^n P_i (k) \log \frac{P_i (k)}{P_{i+1} (k)} \quad (5.22)$$

$$D_{SRRE} (F_i, F_{i+1}) = \sqrt{\sum_{k=1}^n P_i (k) \log \frac{P_i (k)}{P_{i+1} (k)}} \quad (5.23)$$

Equation 5.2 $D_{RE} (F_i, F_{i+1})$ denotes the distance measure of relative entropy and equation 5.3 $D_{SRRE} (F_i, F_{i+1})$ gives the distance measure of and Square Root of Relative Entropy between the frames. The term $P_i = \{P_i (1), P_i (2) \dots P_i (n)\}$ represents the Probability Distribution Function of frames present in a video i.e. given as F_i and F_{i+1} which is obtained from normalized intensity histogram consisting of n bins. The value of bin is $n = 256$ and k represents total number of frames present on each shot. On estimation of distance measures based on Relative Entropy and Square Root of Relative Entropy the comparison is made between them based on time direction. By this comparison, the alterations that have occurred in frames are identified and redundantly occurred frames are eliminated. Also the similar distance measure values of Relative Entropy are also ignored. After completion of this, the estimation of Relative Entropy and Square Root of Relative entropy is repeated for continuous frames that are remained in last step. The continuous frames that are left out are,

$$CF = \{ D (F_j, F_{j+1}), D (F_{j+1}, F_{j+2}), D (F_{j+n-1}, F_{j+n}) \} \quad (5.24)$$

CF represents the Continuous Frames, here the lastly present frame is defined that it has minimum difference between $D (F_{key}, F_{key+1})$ and hence the average distances of Relative Entropy and Square Root of Relative Entropy is defined as follows,

$$F_{key} = F \quad \left| \text{avg} \min_k \left[D (F_j, F_{j+1}) - \frac{\sum_{l=j}^{j+n-1} D (F_l, F_{l+1})}{n-1} \right] \right| \quad (5.25)$$

From the above defined formulation, final average key frames are obtained, hereby the first frame, key frames and last frame are taken into account for further processing. The similar frames are repeated in a video that is called redundancy and it is eliminated. By eliminating redundant frames, the amount of information is minimized and also the cost required for resources are reduced. Key frame selection is followed as a fundamental step in video retrieval process. The major redundancy present in a video is similar shots which are comprised of similar characteristics of a video. Hereby the selected key frames are capable to summarize all the essential salient characteristics present in a video. The main aim of this key frame selection is to minimize processing time and consumption of resources, if key frame selection process is ignored then these two parameters will be increased.



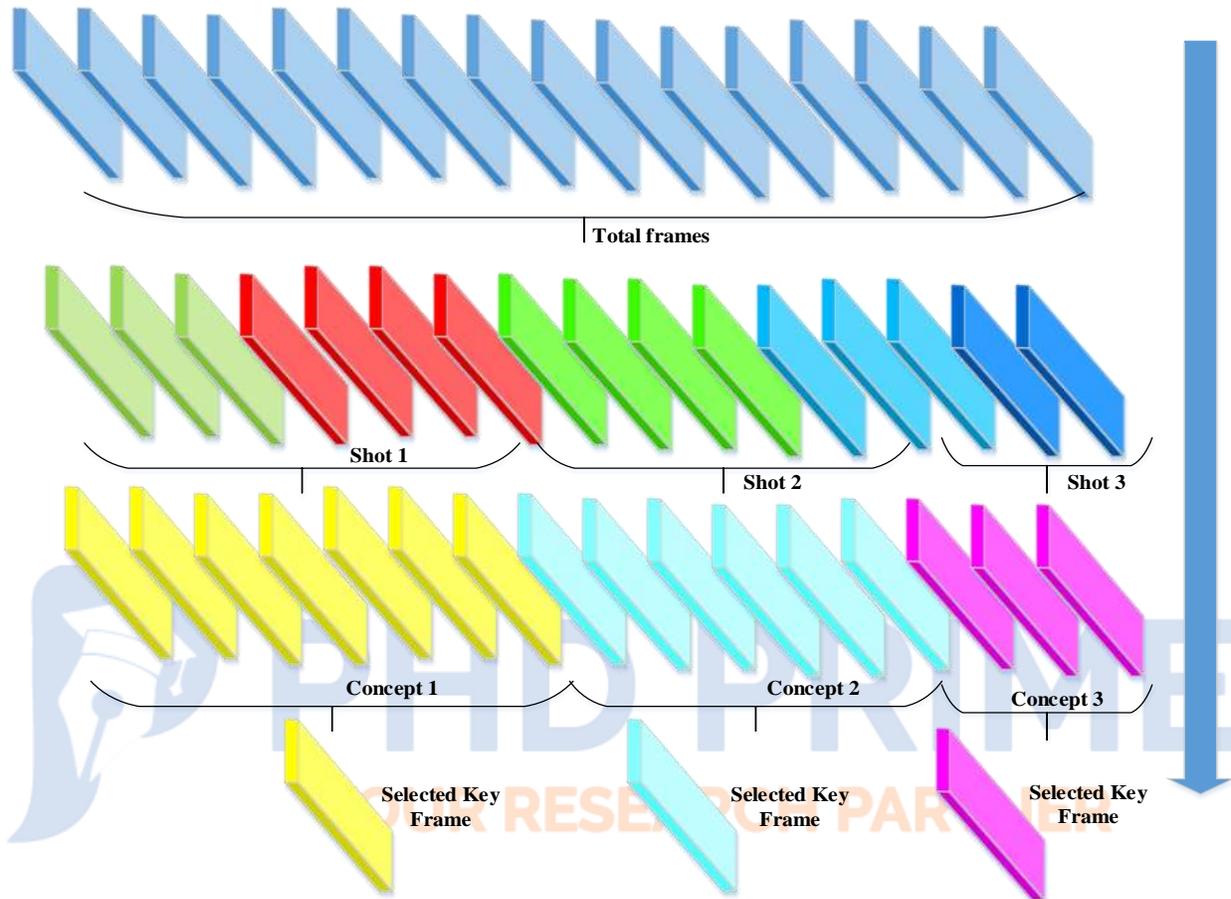


Figure 5.10 Key Frame Selection

Key frame selection process performed by Relative Entropy based Fast Key Frame Selection approach is illustrated in figure 5.10. For instance a 3D video is split into 3 shots and those shots are further divided into frames and then the optimal key frames are selected. According to the procedure the key frames are chosen and considered for next process of denoising.

5.4.2 Denoising

Denoising is the process of removing noise that are present in the selected 3D key frames for the purpose of enhancing the accuracy results and also improvising the quality of video before matching with the given query. Noise in videos and images occur due to the following reasons; (i) Capturing devices (mobile phones, digital camera, and surveillance camera), (ii) Movement of the objects, (iii) Channel noise, (iv) Low resolution devices and (v) Compressed storage. Denoising is considered to be a pre-processing step in image processing by which it preserves useful information present in the image. To achieve higher accuracy results the noise is removed and denoised frames are taken in account for next process.

Noise reduction is implemented, which minimizes the noise (Salt and Pepper) by adaptive median filtering. Image noise reduction direction is from Left to the Right. Adaptive Median Filtering is used for image denoising, which is computed by pixels median value. It is the best denoising filter, which is better than the conventional median filtering. An adaptive median filter implements the spatial processing to identify which pixels $P(i, j)$ is affected by impulse noise. It categorizes the pixels as noisy by comparing it with neighborhood pixels. After identification, the noisy pixels are replaced by the median value pixels. The advantages of adaptive median filtering are follows:

- (1). It removes the impulse noise
- (2). It performs well for smoothening of other noise
- (3). It reduces the distortion such thinning excessive / thickening of boundary in objects.

Algorithm 2. Pseudo code for Bilateral Adaptive Median Filtering

Require: Denoised image i

for $i \rightarrow$ Neighborhood size function $S(x, y)$

Stage 1:

$$v_1 = gl_{Med} - gl_{MIN}$$

$$v_2 = gl_{Med} - gl_{MAX}$$

If $(v_1 > 0) \ \&\& \ (v_2 < 0)$

Go to stage 2

Else maximize the size of window

If $(window \ size < S_{MAX})$ do stage 1

Else return gl_{xy}

Stage 2:

$$u_1 = gl_{xy} - gl_{MIN}$$

$$u_2 = gl_{xy} - gl_{MAX}$$

If $(u_1 > 0) \ \&\& \ (u_2 < 0)$

return gl_{xy}

Return denoised image

Description of Algorithm 2 is follows: If the $v_1 > 0) \ \&\& \ (v_2 < 0)$ is true, then the gl_{Med} is not an impulse noise. In stage 1, gl_{Med} is the median gray level value in $S(x, y)$, gl_{MAX} is the maximum gray level value in $S(x, y)$, gl_{MIN} is the minimum gray level in $S(x, y)$, gl_{xy} is the gray level value at pixel (x, y) . This step continue towards stage 2 to test when the gl_{xy} is an impulse, otherwise gl_{Med} is an impulse. For this

outcome, the window size is maximize and the stage 1 is repeated until the gl_{Med} is not impulse noise and go to stage 2 or S_{max} is reached, then the output is gl_{xy} . In stage 2, if the $(u_1 > 0)$ && $(u_2 < 0)$ then the gl_{xy} is not an impulse noise, then output is gl_{xy} (minimizes the distortion). Otherwise the result is gl_{med} , then gl_{med} is not an impulse noise. The outcome of the adaptive median filter is befits than the conventional median filter because it does not gives the best outcome if the pixel contain impulse noise. Our proposed denoising filtering is suited for two cases:

- Impulse noise range is increased by 2 for both horizontal and vertical directions and noise removed towards the left to right direction. Otherwise the same process is repeated until the impulse noise is removed and replaced by the median value.
- It preserves the image detail (lines, edges, corners, etc.) and smooth's the non-impulsive noise.

When compared to other filtering methods such as Normal Median Filter, Adaptive Median Filter, Impulse Noise Median Filter, Improved Median Filter, the proposed bilateral adaptive median filter performs better which provides higher PSNR for the given noisy image. Table illustrates the performance of the various filtering methods with respect to the PSNR.

Table 5.1 Performance of Filtering Methods

Filtering Method	PSNR
Normal Median Filter	18.47db
Adaptive Median Filter	30.845db

Impulse Noise Median Filter	33.45db
Improved Median Filter	35.92db
Bilateral Adaptive Median Filter	45.25db

5.3.3 Feature Extraction

All the denoised optimal frames and 3D images from user are denoised and proceeds into feature extraction. By extracting features, all the informative and non-redundant data are collected from each frame and 3D input image. In this work two different categories of features are taken in account, they are

- Visual low-level features
- Visual semantic features

Visual low-level features include color, texture and shape, similarly the visual semantic feature in motions present in the frames and images.

Visual Low-level features

In this section we discuss about the extraction of three features and the formulations defined for this feature extraction. Color feature is extracted by converting the image into grey scale image that is followed by the estimation of intensity values. Therefore the conversion of RGB into grey scale is expressed in the following equation as,

$$I_Y = 0.333 * F_R + 0.5 * F_G + 0.1666 * F_B \quad (5.26)$$

From equation (5.13), I_Y denotes the intensity values that is equivalent to grey level image of RGB and the intensities of R,G,B components are represented as $F_R, F_G,$

F_B respectively. Further the images and frames are converted into HSV and YCC color space for extracting the values of hue, saturation, luminance and intensity. Further color histograms are obtained and normalized, from the normalized color histogram the color feature is extracted.

Next texture features are extracted from the visual low-level features present in the image. Conventional method of estimating co-occurrence metric is involved for texture feature extraction. This method is supported by determination of 14 statistical texture measurements. Angle orientation based co-occurrence matrix is performed between pair of grey level and axis. The orientations taken in account are $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$. The grey level co-occurrence matrix is formulated as,

$$p(i, j | d, \theta) = \{(x_1, y_1), (x_2, y_2)\} \quad (5.27)$$

$i = I(x_1, y_1), j = I(x_2, y_2)$ respectively, $|x_1 - x_2| = 0^\circ, |y_1 - y_2| = d$ respectively. The terms d and θ denotes distance and angle, therefore the grey level co-occurrence matrix $p(i, j | d, 0^\circ)$ is defined with respect to d and θ . The formulation of probability value of grey level co-occurrence matrix is,

$$p(i, j | d, \theta) = \frac{p(i, j | d, \theta)}{\sum_{i=1}^{256} \sum_{j=1}^{256} p(i, j | d, \theta)} \quad (5.28)$$

Texture features involved in this work are

- Angular Second Moment
- Entropy
- Correlation

- Contrast
- Mean.

Angular second moment (ASM) is defined as sum of the squares of entries present in Grey Level Co-occurrence Matrix which is used for the measurement of image homogeneity. This texture feature is expressed as,

$$ASM = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P_{ij}^2 \quad (5.29)$$

i, j are the spatial coordinates of the defined function $p(i, j)$ and Ng is grey tone present in the image. ASM is also called as Uniformity or Energy.

Entropy implies the total contents present in the image that is requires for compressing. This entropy texture feature is formulated as follows,

$$Entropy = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} - P_{ij} * \log P_{ij} \quad (5.30)$$

Entropy is the measure of loss of information or data present in the transmitted signal and also it is capable to determine the image information. Then correlation feature is defined as the measurement of grey level linear dependency over neighboring pixels. This feature is given as,

$$Correlation = \sum_{i,j}^K \frac{(i - \mu_i)(j - \mu_j)}{\sigma_i \sigma_j} \quad (5.31)$$

μ and σ is the mean and standard deviation, this feature is enabled to measure pixels that are correlated with neighbors. Further contrast measures the local variations on Grey

Level Co-occurrence matrix in which the terms $i = j$ and k defines the row and column dimension i.e. 3×3 square matrix. The mathematical expression for contrast feature is,

$$\text{Contrast} = \sum_{i,j=0}^K P_{i,j} (i - j)^2 \quad (5.32)$$

Here $P_{i,j}$ is the probability of pixel pairs that satisfies offset. This feature in texture is also known as inertia. Lastly, mean is also one of the texture feature that is expressed in terms of a mathematical formulation,

$$\text{Mean} = \frac{1}{2} \sum_i^U \sum_j^V (iP [i,j]) + j P[i,j]) \quad (5.33)$$

This expression extract mean texture feature from the given frame or image. Mean is supported to measure the grey level present in the given image or frame. On obtained all the measurements the texture feature is extracted from the given denoised frames and query images. These measures are also used for similarity matching between query image and 3D frames.

Shape feature is comprised into two such as region and moments which are the significant entities to be focused in this feature. 3D videos are comprised with different shapes which are analyzed in terms of region and moments. For shape feature extraction, Shape Adaptive Discrete Wavelet Transform (SA-DWT) is applied for 3D in this research work. Extracting regions from 3D key frames follows SA-DWT applied over all sub bands. As per the 3D optimal key frame the sub bands are,

$$3D \text{ KF}_{(x, y, z)} = \text{subbands (LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH)} \quad (5.34)$$

On applying SA-DWT, decomposition of regions occur over the rows and columns. Sub bands are not selected since it is not suitable for region extraction. The obtained resulted pixel values are referred based on the pixel values that are present inside the segment regions. Edges are specified by including mask information from frames or images. As a result, convolution high pass and low pass filters are applied with respect to the directions (horizontal and vertical).

Moments of the 3D frames are determined using 3D Zernike moments, size of the objects is specified initially. 3-dimensional array is used in this work; hence the size is $N \times N \times N$ array of voxels. We formulate the intensity function of voxel points as,

$$x_i = \frac{2i-N-1}{N\sqrt{3}}, y_j = \frac{2j-N-1}{N\sqrt{3}}, z_k = \frac{2k-N-1}{N\sqrt{3}} \quad (5.35)$$

The sampling intervals in x , y and z directions $\Delta x_i = x_{i+1} - x_i$, $\Delta y_j = y_{j+1} - y_j$, $\Delta z_k = z_{k+1} - z_k$ along with i, j and $k = 1, 2, \dots, N$. The order of $(r + s + t)$ in 3D geometric moments is defined as,

$$G_{rst} = \sum_{k=1}^{\lfloor N/2 \rfloor} I_t(z_k) R_{rsk} \quad (5.36)$$

The term G defines the geometric moment in the order of $(r + s + t)$, rsk denotes the exponents at point z , I and R represents the intensity and augmented intensity function. From the mathematical equations, the moments for keys frames and input 3D image is determined.

2) Semantic Feature

Motion feature is comprised of motion information which is considered to be salient feature in retrieving videos. This information is significant since those values are obtained from foreground pixel which ignores background pixels. The motion information for k^{th} frame is determined from,

$$MI(k) = \sum P_{(i,j)}(k) , 0 \leq i < W, 0 \leq j < H \quad (5.37)$$

Let W and H be the width and height of a frame, $P_{(i,j)}(k)$ is the binary value of pixel in i^{th} row and j^{th} column over the k^{th} frame. Here, it is represented as 1 while the pixels are foreground else they represent 0 in the frame or image. As a result motion information for each frames and input image is extracted. The extraction process ends here, which is followed by final process of similarity matching.

Color, texture and motion features are extracted using Holistic CNN. This is the extended version of the CNN which provides higher feature extraction result in a minimum training time. Overall the model consists of seven layers as

- Four Convolution Layers
- Two Fully Connected Layers
- Softmax Layer (Output Layer)

Three kinds of Holistic CNN are constructed for extracting Color, Texture and Motion Features. The proposed model learns these features automatically with the use of these seven layers. However, FC layer has faces more computational complexities and thus only two layers are used for improving the feature importance and accuracy. Each convolution layer consists of Convolution, Batch Normalization (BN), ReLU activation

and pooling operations. The first convolution block consists of the RGB image of the matrix size is $100 \times 400 \times 1$ and its convolved with the kernel size of $16(11 \times 11 \times 1)$ and it results the feature map of the size is $90 \times 390 \times 16$. The ReLU and BN activation functions are performed across the convolutional layer's output. The max pooling of 2×2 which is performed the after to provide a new feature map of the size and the $45 \times 195 \times 16$. After which, the same process is repeated 3 times. Different set of kernels are utilized for the convolution layer of 3 succeeding blocks: $32 \ 7 \times 7 \times 16$ kernels (second block), $64 \ 5 \times 5 \times 32$ kernels (third block) and $128 \ 3 \times 3 \times 64$ kernels (4th block). In the final convolution block, average pooling of 2×2 with stride 2 is used and the fully connected layer consists of the 120 neurons and totally 240 neurons are used. In the final layer, softmax is used to compute the weight values for the features based on the feature importance. The overall holistic 3D CNN model for color, texture and motion features extraction is depicted in figure 5.11, 5.13, 5.14 respectively.

Red Channel				Green Channel			
35	18	28	10	35	18	28	10
14	22	16	53	14	22	16	53
6	9	22	14	6	9	22	14
12	14	12	15	12	14	12	15

Blue Channel			
35	18	28	10
14	22	16	53
6	9	22	14
12	14	12	15

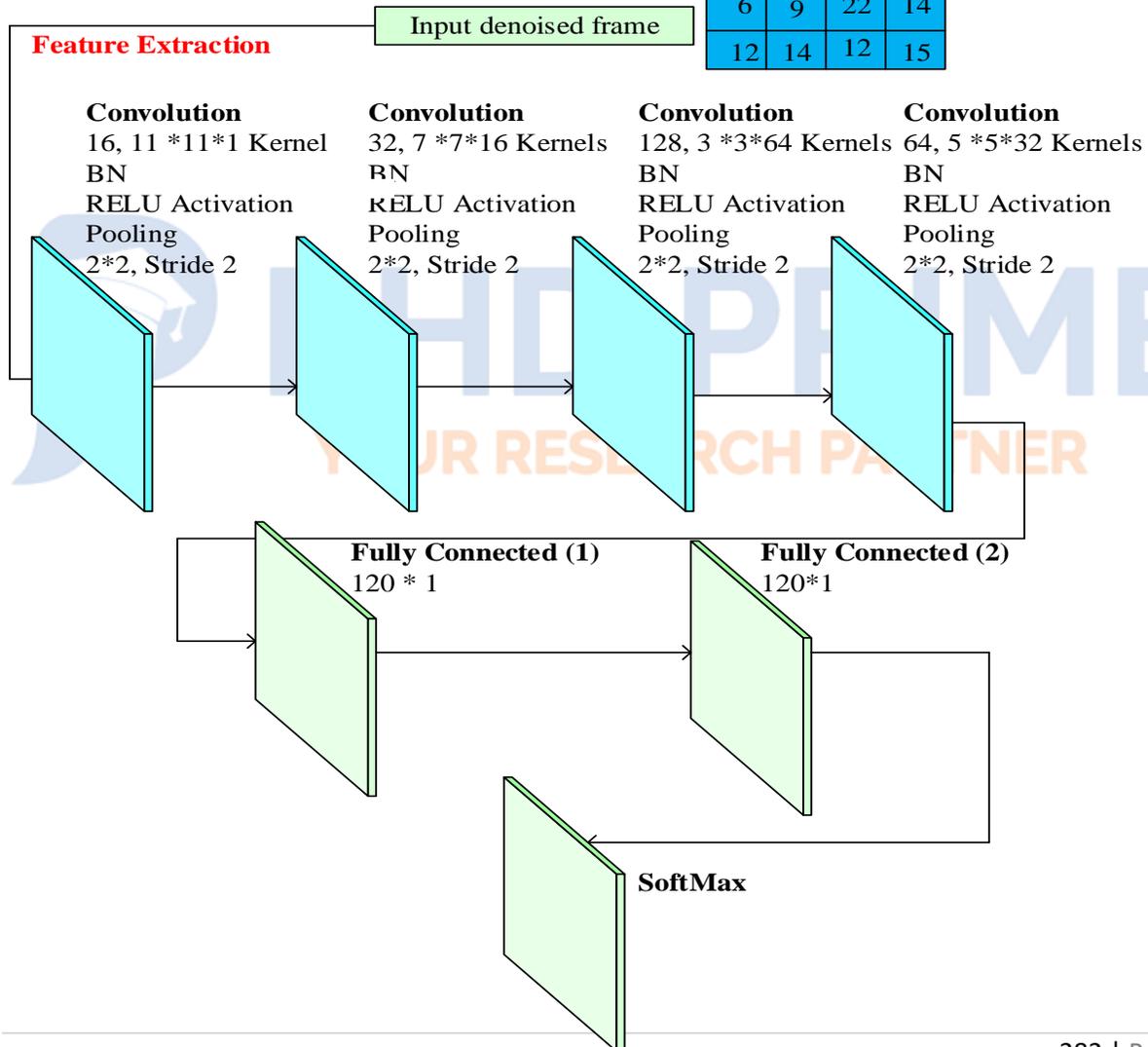


Figure 5.11 Color Feature Extraction



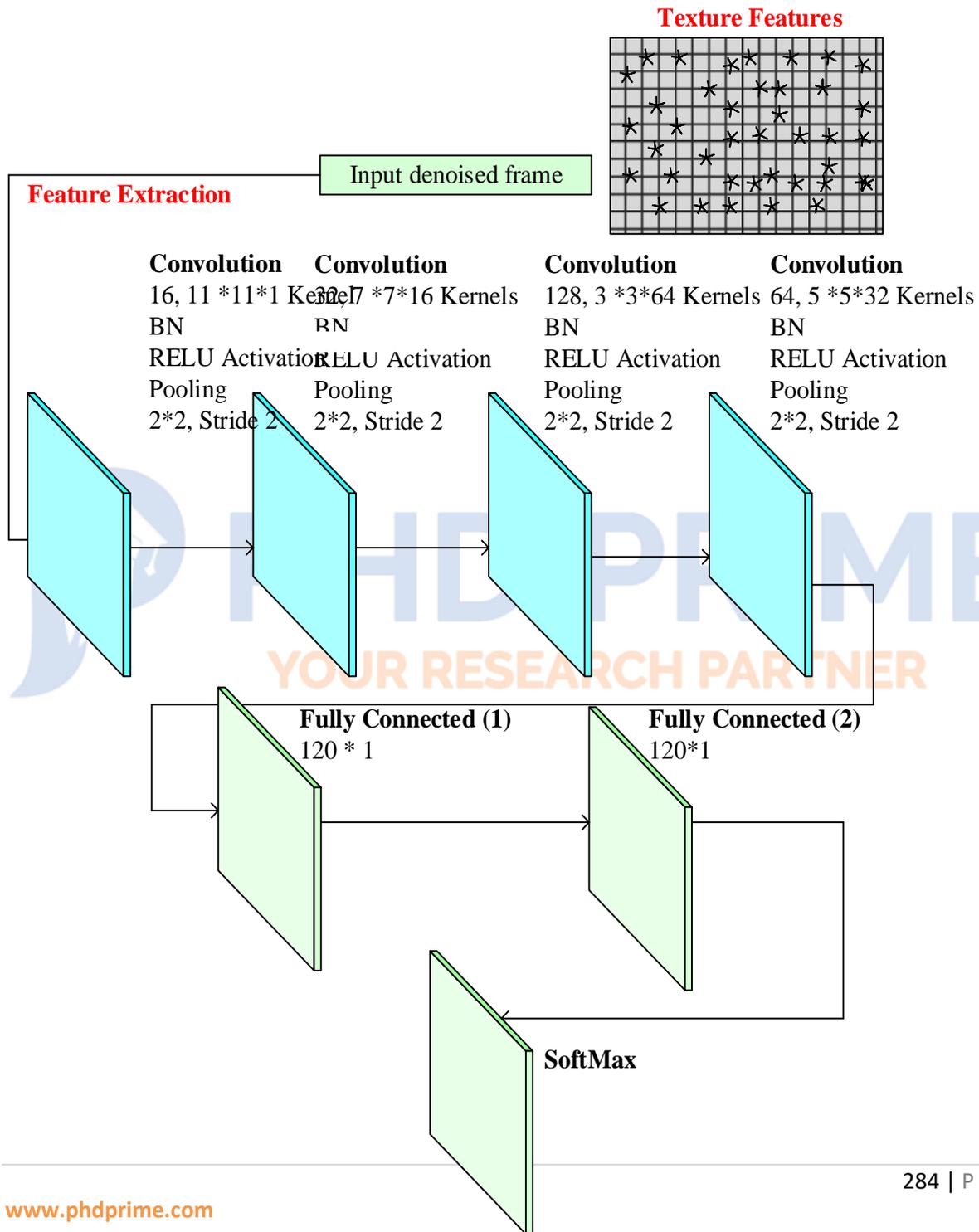


Figure 5.12 Texture Feature Extraction



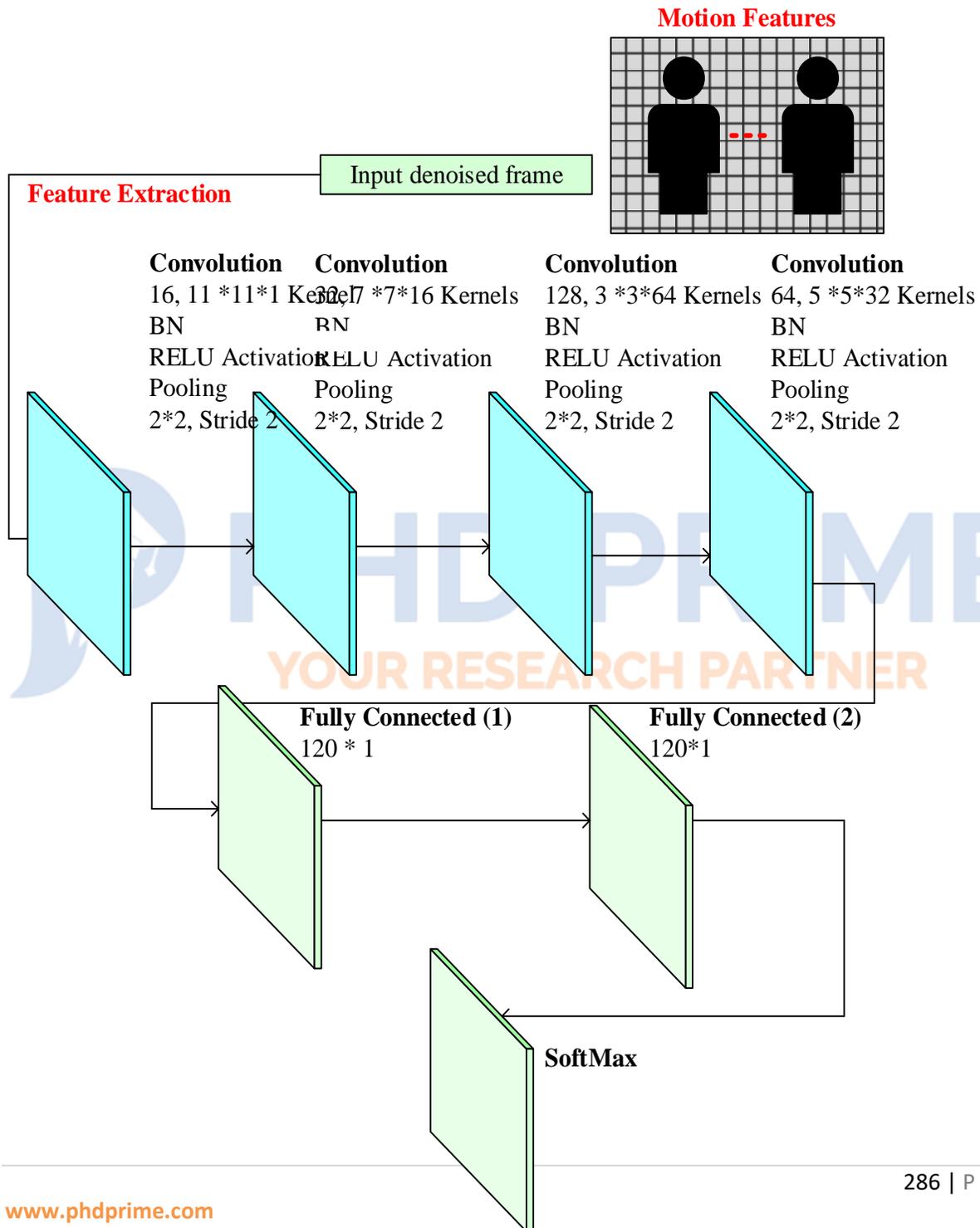


Figure 5.13 Motion Feature Extraction

Table 5.2 Holistic 3D CNN Layers

Layer type	# Filter	Kernel size	Stride	Output size	# Parameters
Image input	-	-	-	100 x 400 x 1	-
convolution	16	11 x 11 x 1	1	90 x 390 x 16	1952
BN	-	-	-	90 x 390 x 16	32
ReLU	-	-	-	90 x 390 x 16	0
Max pooling	1	2 x 2	2	45 x 195 x 16	0
Convolution	32	7 x 7 x 16	1	39 x 189 x 32	25,120
BN	-	-	-	39 x 189 x 32	64
ReLU	-	-	-	39 x 189 x 32	0
Max pooling	1	2 x 2	2	19 x 94 x 32	0
Convolution	64	5 x 5 x 32	1	15 x 90 x 64	51,264
BN	-	-	-	15 x 90 x 64	128
ReLU	-	-	-	15 x 90 x 64	0
Max pooling	1	2 x 2	2	7 x 45 x 64	0
Convolution	128	3 x 3 x 64	1	5 x 43 x 128	73,856
BN	-	-	-	5 x 43 x 128	256
ReLU	-	-	-	5 x 43 x 128	0
Average pooling	1	2 x 2	2	2 x 21 x 128	0
FC	-	-	-	1 x 1 x 120	645,240

Softmax	-	-	-	1 x 1 x 120	0
Class output	-	-	-	-	0
				Total	797,912

Figure 5.11, 5.12, 5.13 represents the feature extraction for color, texture and motion using holistic 3D CNN. In each holistic layers, features are extracted and the configuration and hyper parameters are described in table.

Table 5.3 Hyper Parameters of Holistic 3D CNN

Hyperparameter	Value
Batch size	10
Number of epochs	15
Momentum	0.9
Initial learn rate	0.001
L2Regulation	0.001

5.3.4 Similarity Matching

Similarity matching is an important process followed in this work by which the final relevant results are obtained. In this work, similarity matching is performed by novel multi-features matching scheme over Hadoop MapReduce framework. This similarity matching scheme includes texture, shape, color and motion features into considerations.

1) Color feature

Let A and B be the color objects in two 3D frames $f1$ and $f2$ respectively, it specifies the similar color present in two frames in terms of the number of pixels. This is given as,

$$u \in A, v \in B$$

(u, v) represents the presence of similar color pair. Then the colors are matched in accordance to Euclidean distance,

$$\|u - v\| < TH \quad (5.38)$$

TH is the threshold value in equation 5.21 which is assumed to be 3 for this work, u and v denotes the distance between HSV color space.

Let assume, $\Omega = \{(u, v) | (u, v) \in A * B, (u, v) \text{ is said to be a similar color pair}\}$, Hereby the color similarity between the frames $f1$ and $f2$ is given in the following formulation,

$$\text{Color - Similarity } (f1, f2) = \frac{1}{k} \sum_{(u,v) \in \Omega} \{W(D_s(u, v)) * \min(\overline{H_{C1}}(u), \overline{H_{C2}}(v))\} \quad (5.39)$$

Where $\overline{H_{C1}}$ and $\overline{H_{C2}}$ are the average color histogram obtained from frame $f1$ and frame $f2$ respectively. k is the image size, W represents the weight function i.e. Sigmoid function, D_s represents the spatial features present in color objects over the frames $f1$ and $f2$. The value of D_s is given as,

$$D_s(C(A), C(B)) = \frac{1}{4} (|\overline{f_{i1}} - \overline{f_{j1}}|) + (|\overline{f_{i2}} - \overline{f_{j2}}|) + (|\overline{f_{i3}} - \overline{f_{j3}}|) + (|\overline{f_{i4}} - \overline{f_{j4}}|) \quad (5.40)$$

Using these formulation, color is considered for similarity matching between query and frames in database.

2) Motion feature

Motion feature considers horizontal and vertical displacements of the motion vectors which range between $-M$ to M . Motion is significant, since in this work videos are retrieved with the corresponding query. In a video file motion changes are significant due to the involvement of moving objects.

Let, $R = \{ (x, y, z) \mid -M \leq x \leq M, -M \leq y \leq M, -M \leq z \leq M, (x, y, z) \text{ is said to be similar motion pair} \}$. Then the motion similarity is mathematically computed as,

$$\text{Motion - Similarity } (f1, f2) = \frac{1}{k} \sum_{(x,y,z) \in R} \min(\overline{H_{M1}}(x, y, z), \overline{H_{M2}}(x, y, z)) \quad (5.41)$$

From average histograms of $f1$ and $f2$ represented as $\overline{H_{M1}}$ and $\overline{H_{M2}}$ respectively, motion similarity is given in equation 5.28. Next shape similarity considers both regions and moments over the surface that ranges from $-S$ to S .

Let, $S = \{ (x, y, z) \mid -S \leq x \leq S, -S \leq y \leq S, -S \leq z \leq S, (x, y, z) \text{ is said to be similar motion pair} \}$.



Figure 5.14 Motion Feature

3) Shape feature

Shape similarity is formulated as,

$$\text{Shape - Similarity } (f1, f2) = \frac{1}{k} \sum_{(x,y,z) \in R} \min(\overline{H_{S1}}(x, y, z), \overline{H_{S2}}(x, y, z)) \quad (5.29)$$

This similarity is obtained from average shape histograms $\overline{H_{S1}}$ and $\overline{H_{S2}}$ for the frames $f1$ and $f2$. From the obtained equation (5.42) the shape similarity is predicted in similarity matching process.

4) Texture feature

For estimating texture similarity, Gower's Similarity coefficient is followed. The mathematical formulation is given as,

$$GS_{A,B} = \frac{1}{N} \sum_{T=1}^n \left(1 - \frac{|x_{AT} - x_{BT}|}{NM_T} \right) \quad (5.42)$$

$GS_{A,B}$ defines the estimated texture similarity, A, B denotes the 3D frames, N is the total number of texture features, T represents the texture and NM_T denotes the

normalized factor. This normalized factor is computed from the following mathematical equation,

$$NM_T = \text{Max} (x_{AT}) - \text{Min} (x_{BT}) \quad (5.43)$$

$x_{AT}, x_{BT}, T = 1, 2, \dots$ in equation (5.31) defines the values taken into consideration of each frame A,B,..... Finally all the similarity measurements are combined and a value is determined from the combination of considered color feature, motion feature, shape feature and texture feature. Hence the similarity value measured as,

$$\begin{aligned} \text{Similarity} = & \text{Color} - \text{Similarity} (f1, f2) + \text{Motion} - \text{Similarity} (f1, f2) + \\ & \text{Shape} - \text{Similarity} (f1, f2) + \text{Texture} - \text{Similarity} (A, B) \end{aligned} \quad (5.44)$$

If the obtained similarity value is higher, then those videos are listed as more relevance videos present in HDFS which is obtained on retrieval. The most similar video is provided with a rank 1 which is ranked top with more relevance. According to the similarity value the ranks are provided in the increasing order and sorted in similar manner. This similarity process is held in Hadoop MapReduce framework for achieving faster results.

5.3.5 Hadoop MapReduce Framework

Hadoop is defined as a java based programming framework designed for supporting enormously large data sets storage over the distributed computing environment. This programming model includes the following significant operations; (i) Map (), (ii) Reduce

(i) and (iii) shuffle (i). Similarity values of query images and frame are estimated in map operation, hereby the key value pairs are provided according to the frames and similarity values.

Hadoop framework's operation of MapReduce is demonstrated in figure 5.15 from which final video results are retrieved. Individual operation in Hadoop is discussed below,

1) Mapping Operation – Mapping operation is responsible to map the similarity values with the corresponding frames which is represented as,

$$[f_1 \rightarrow \text{sim}(f_1), f_2 \rightarrow \text{sim}(f_2), f_3 \rightarrow \text{sim}(f_3), \dots] \quad (5.45)$$

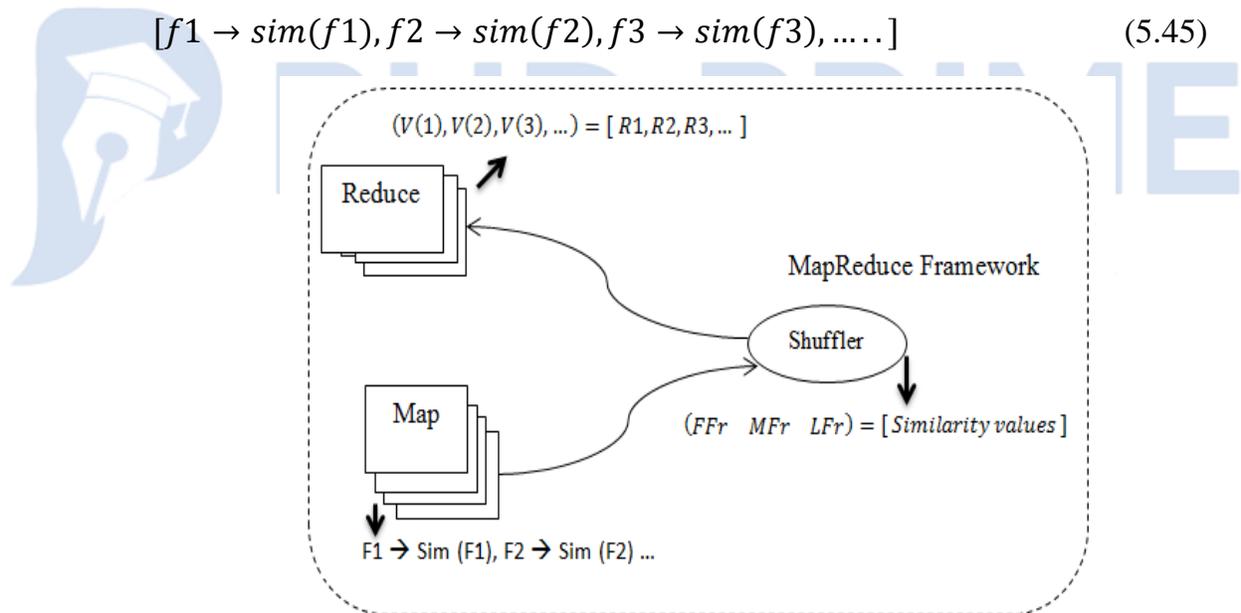


Figure 5.15 Hadoop Operations

Mapping process is proceeded with shuffle operation, the frames sequences as first, middle and last are arranged in the key frame selection process. Those frames are sorted in terms of their similarity values.

2) Shuffle Operation – Shuffle operation is provided with a key value pair, from sequence of frames and similarity values. Hereby the shuffle operation according to the proposed work is represented as,

$$(FFr \quad MFr \quad LFr) = [Similarity \ values] \quad (5.46)$$

FFr, *MFr* and *LFr* denotes the first frame, middle frame and last frame respectively, these are the frames which are obtained from the key frame selection process.

3) Reduce Operation – This is the final operation performed in Hadoop framework for retrieving relevant videos. The reducer consists of videos and their corresponding ranking values. From this key value pair is generated for reduction process, this operation is represented as,

$$(V(1), V(2), V(3), \dots) = [R1, R2, R3, \dots] \quad (5.47)$$

The results from reducer operation are the final results obtained for the user's given query. The most matched results are ranked at top position and also listed at top position of the retrieval. The use of Hadoop MapReduce framework provides higher efficiency in retrieval.

5.5. Results Discussion

This section evaluates the efficacy of the proposed DL which is based on the 3D Holistic CNN with Map-Shuffle-Reduce for CBVR. In this section, simulation results are compared with the existing methods in terms of multiple datasets. In our proposed method, Fast Bilateral Filter, Scalable Key Frame Selection, Multi-stage ESN-SVM classifier is used for CBVR across different characteristics.

Testing Scenarios

- Scenario 1: N_1 & N_5 are used
- Scenario 2: N_1 , N_2 & N_5 are used
- Scenario 3: All the five nodes ($N_1 - N_5$)

Where $N_1 - N_4$ are slave nodes and N_5 is the master node

4.5.1. Dataset Description

Our method utilizes three public 3D datasets such as Hollywood 3D (I), YouTube 8M segments (II), NAMA3DS1-CoSpaD (III). These databases are employed to estimate the performance of the descriptor. All these datasets consists of set of videos for different human actions and scene description.

- Encoding format: flv, wmv, avi, mpg, mp4, ram ...
- Frame rate: 15fps, 25fps, 29.97fps ...
- Frame resolution: 174x144, 320x240, 240x320 ..

Table 4.1 exemplifies the datasets used in deep learning based 3D videos retrieval for a given query image. The main of attributes of each dataset is about its description, number of videos, number of classes, video features and video type.

Table 5.4 Dataset Description

Dataset	Description	# of videos	# of classes	Video features	Video type
Hollywood 3D (I)	Action recognition	650	14	Viewpoint, lighting, background, unknown camera motion	HD
YouTube 8M segments (II)	Entertainment (games, news, shopping)	1000	> 20	Colorfulness, object & camera motion and texture	HD, full HD, ultra HD
NAMA3DS1-CoSpaD (III)	Scene descriptions (natural, sports, media)	110	10	Spatial, temporal and in-depth contents	Full HD



Query Image



Figure 5.16 Relevant Results for Action Query Image (Dataset 1)





Query Image



Figure 5.17 Relevant Results for Games Query Image (Dataset 2)



Query Image



Figure 5.18 Relevant Results for Sports Query Image (Dataset 3)

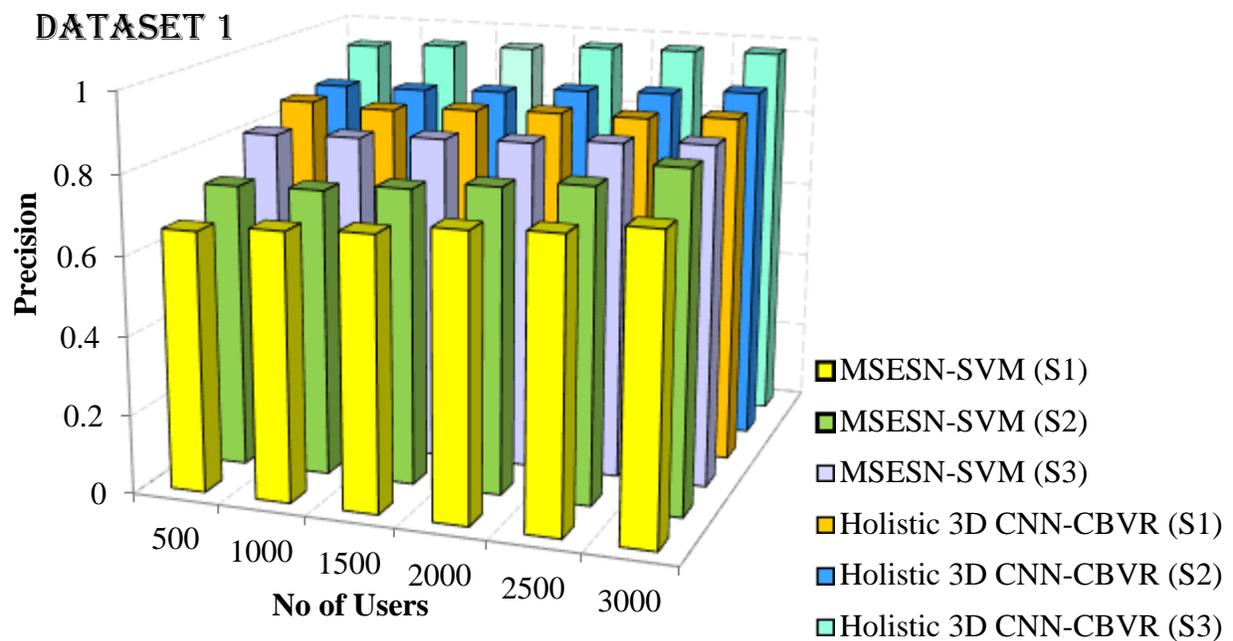
5.5.3 COMPARATIVE ANALYSIS

The performances of this proposed work are analyzed for various performance metrics and all the metrics are related to the proposed concept, so that the better achievements of this proposed research work can be analyzed. Each metric is compared with previous research work to analyze the betterment of this proposed work.

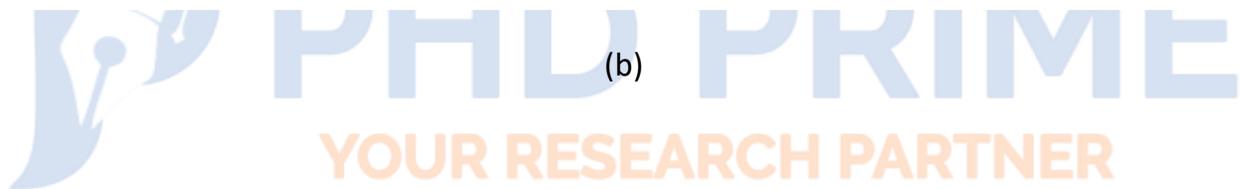
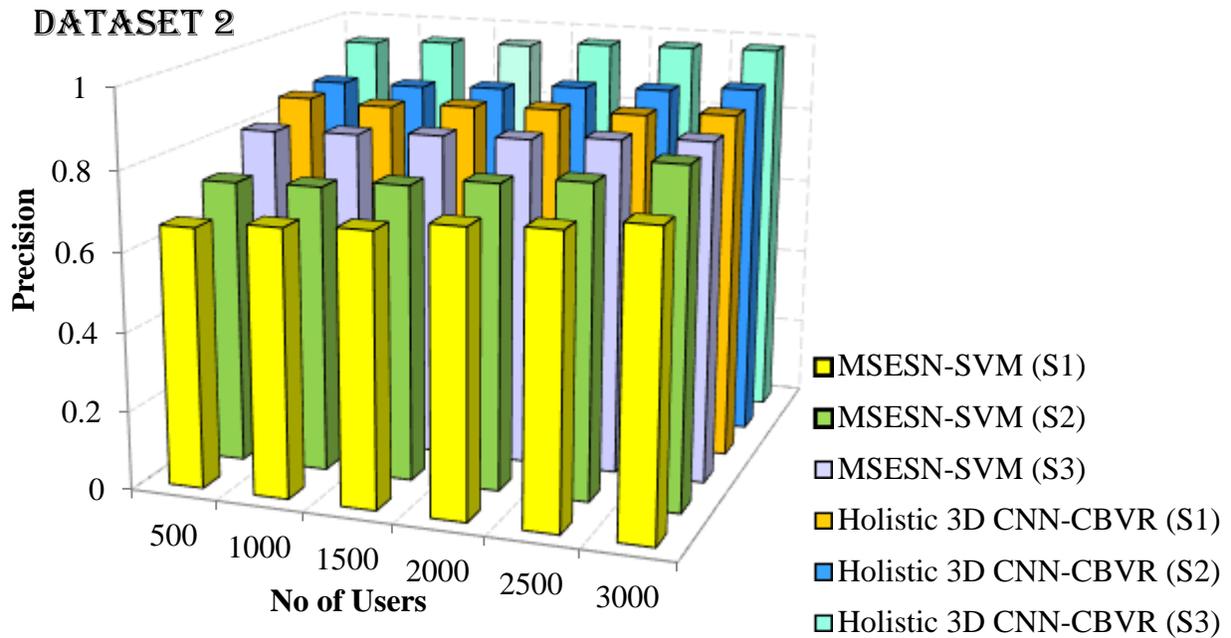
5.5.3.1 Precision

The performance of the precision is computed for accurate results and not accurate precise when feature extraction, selection and classification becomes failure. However, precision is the measure of how much detailed information was provided and it is the degree to which the exactness was applied. Figure 5.19 shows that the result of precision

in which the result of the holistic 3D CNN is applied for four different features extraction such as color, texture, motion and shape. Then the similar features are extracted using map shuffle reduce. In previous work i.e. MSESN-SVM, multiclass based SV is used which classifies the input features into various classes.



(a)



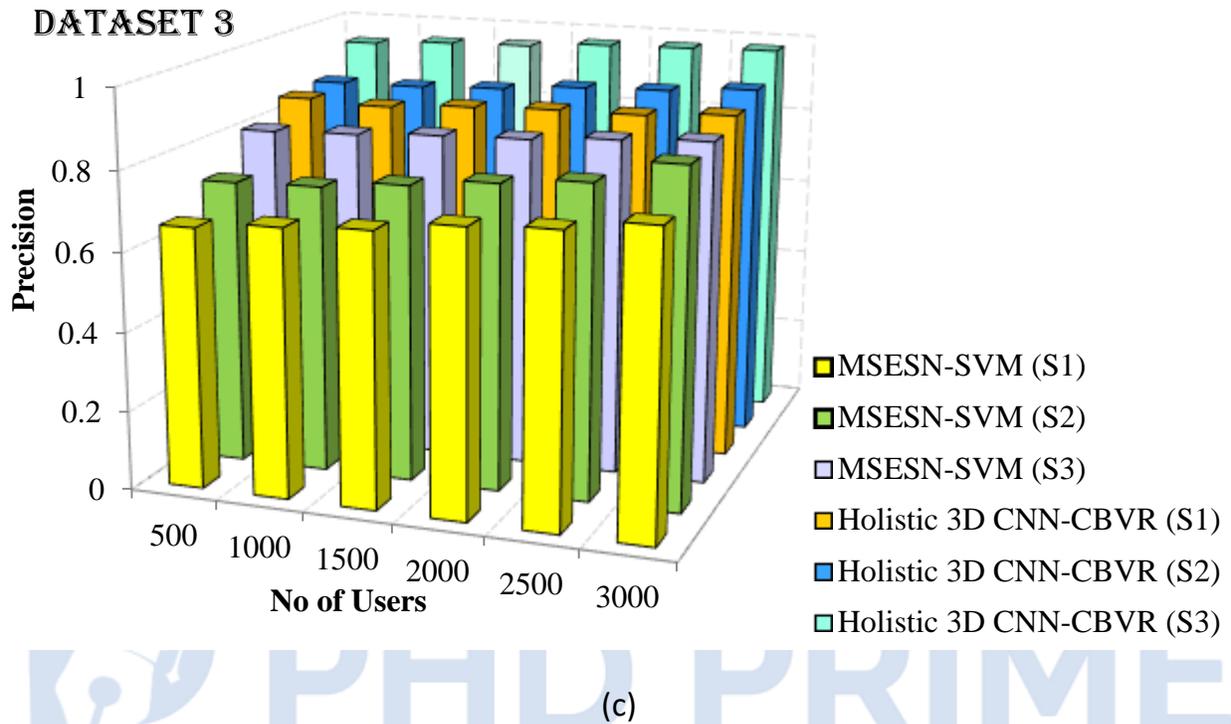


Figure 5.19 Performance of Precision

Table illustrates the performance of the precision that represents the set of observations. Precision is the degree of conformity to a known reference value. The true value and observed values are known as the reference value.

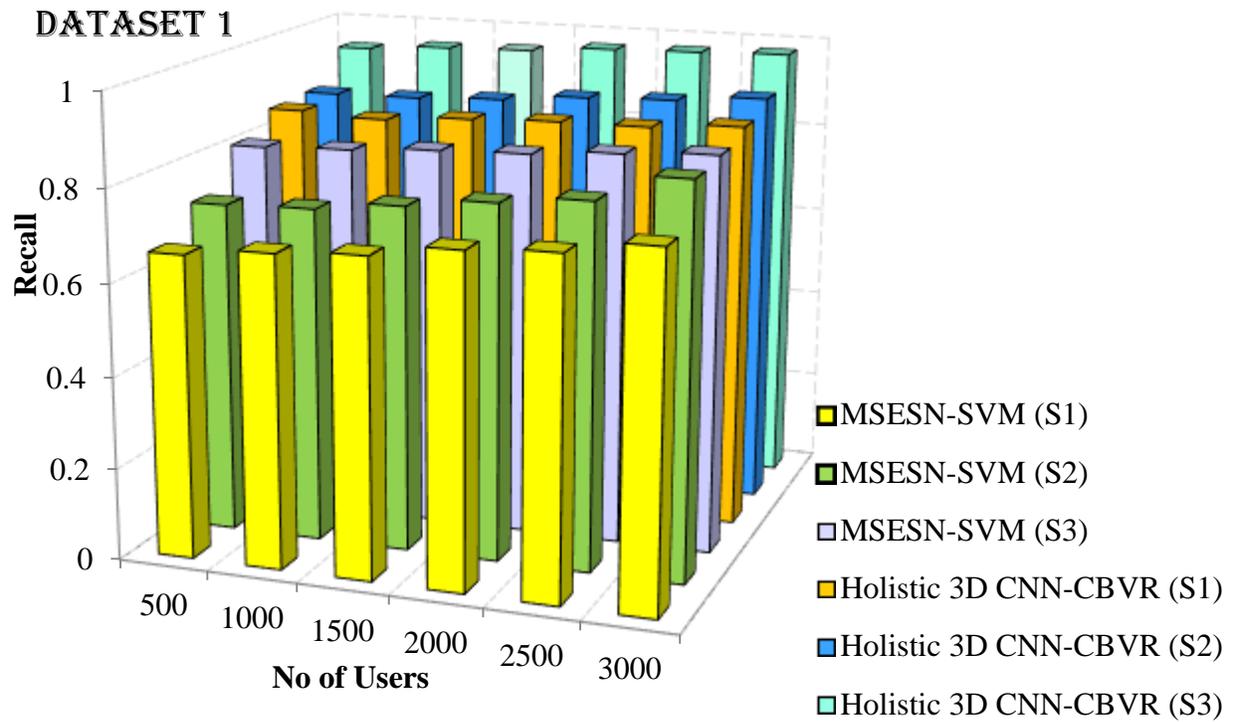
Table 5.5 Statistical Analysis on Precision

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN - CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM

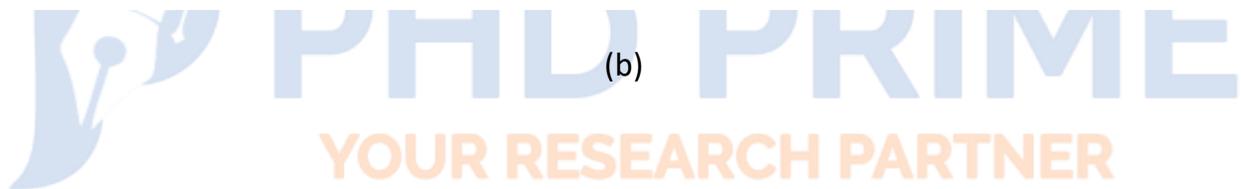
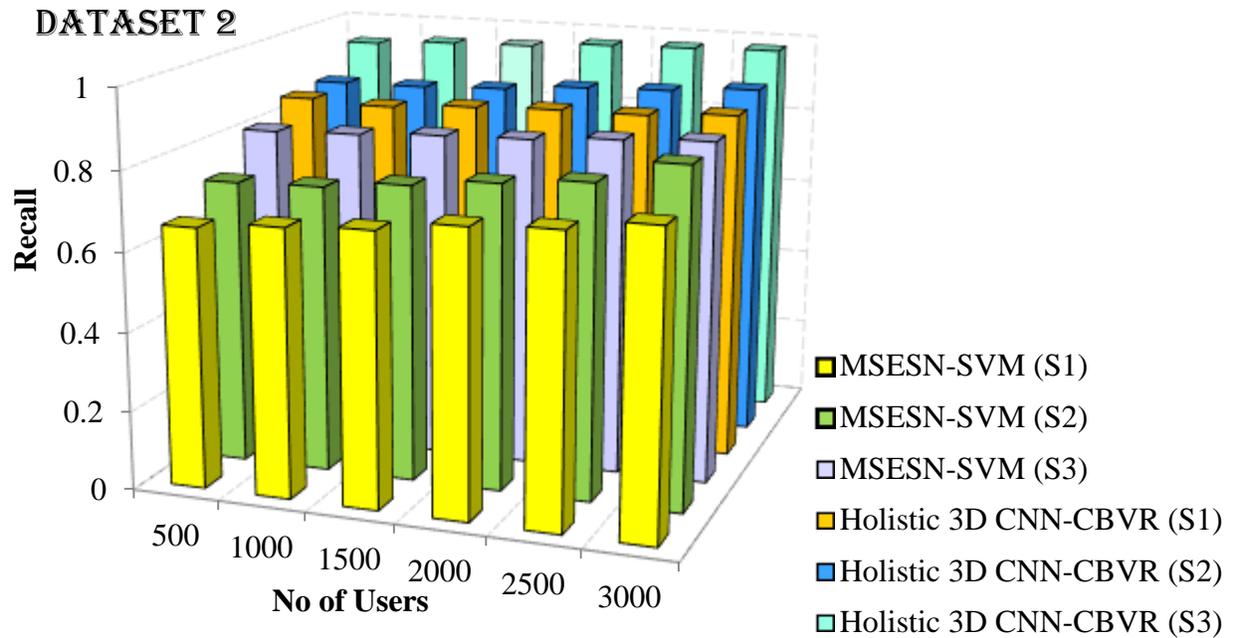
Hollywood 3D (I)	0.84	0.77	0.88	0.8	0.985	0.867
YouTube 8M segments (II)	0.85	0.775	0.89	0.82	0.981	0.875
NAMA3DS1-CoSpaD (III)	0.857	0.789	0.91	0.835	0.982	0.885

5.5.3.2 Recall

Recall is represented by a tight grouping of shots (they are finely tuned) by the color, texture, motion and shape. It is represented by providing the correct dimension of the components. The ratio of the total relevant videos in a database retrieved by your given query in the search is called as Recall. When you know that there are 1000 videos in a database and your search retrieved videos are 100 means that the recall value for a given query is 10%. The result of recall is depicted for the number of users which are compared for the three different scenarios. In third scenario, holistic 3D CNN is used to obtain the higher performance. In database, features are grouped using modified mapreduce paradigm that provides the higher recall value. In previous works, lack of two processes such as key frame selection and denoising results the poor recall value. The performance of recall for all scenarios is better than the previous works.



YOUR RESEARCH PARTNER (a)



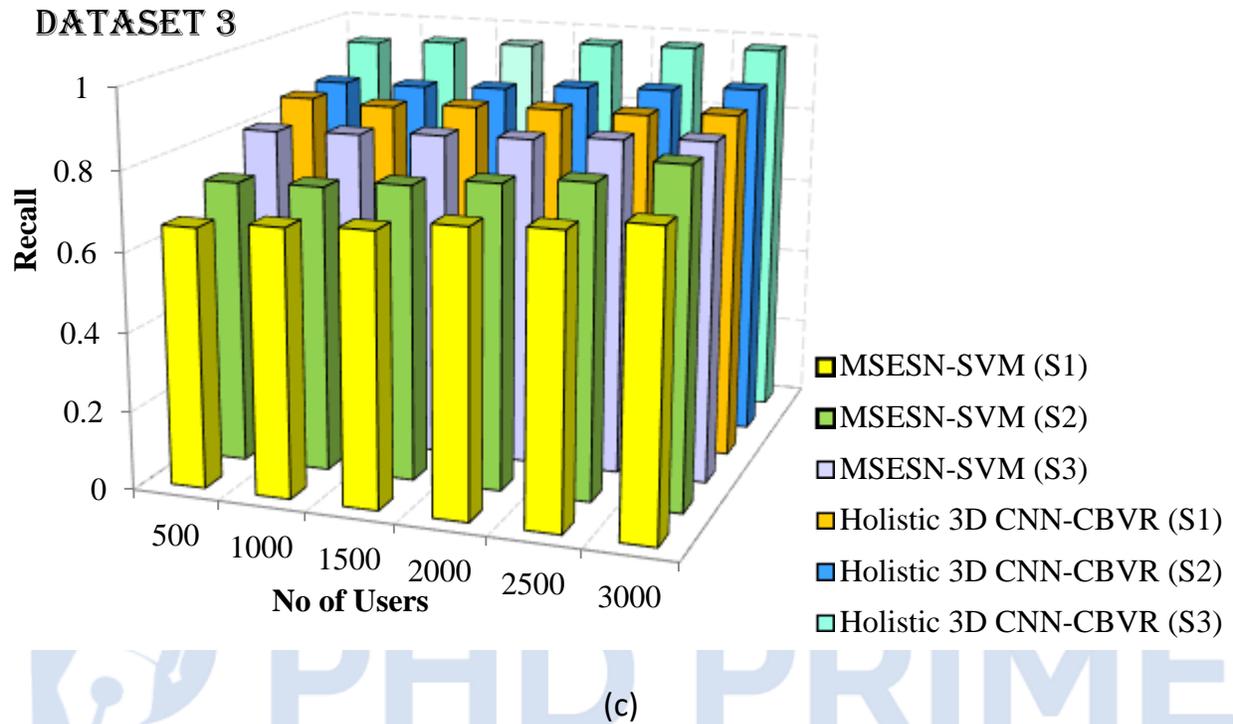


Figure 5.20 Performance of Recall

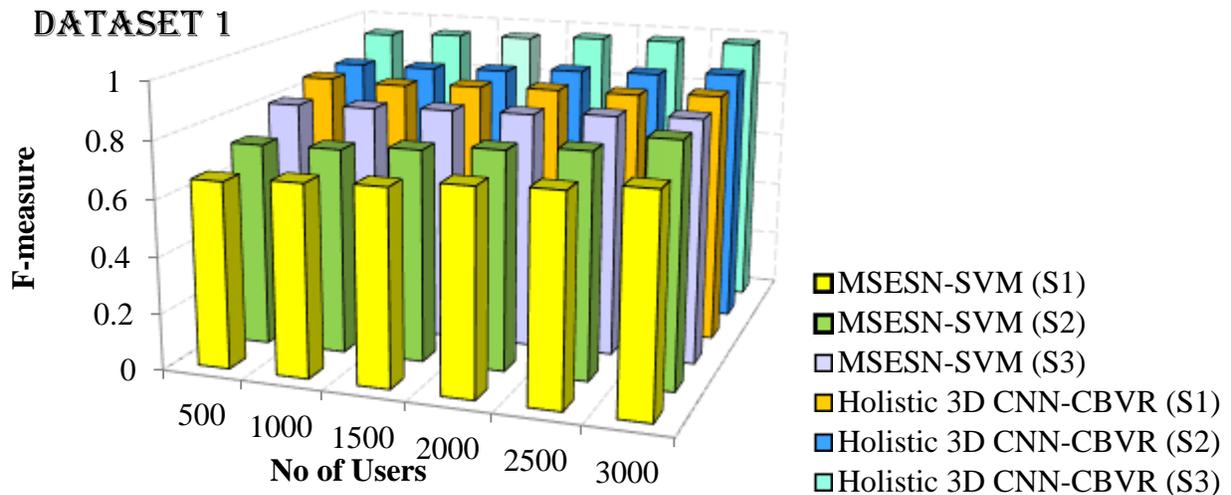
Table 5.6 Statistical Analysis on Recall

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN - CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM
Hollywood 3D (I)	0.81	0.74	0.86	0.78	0.965	0.847
YouTube 8M segments (II)	0.83	0.745	0.87	0.8	0.961	0.855

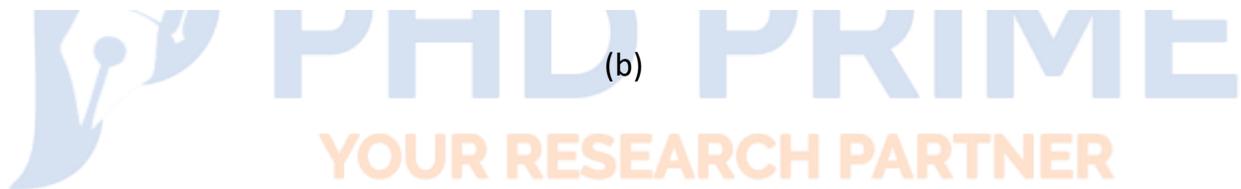
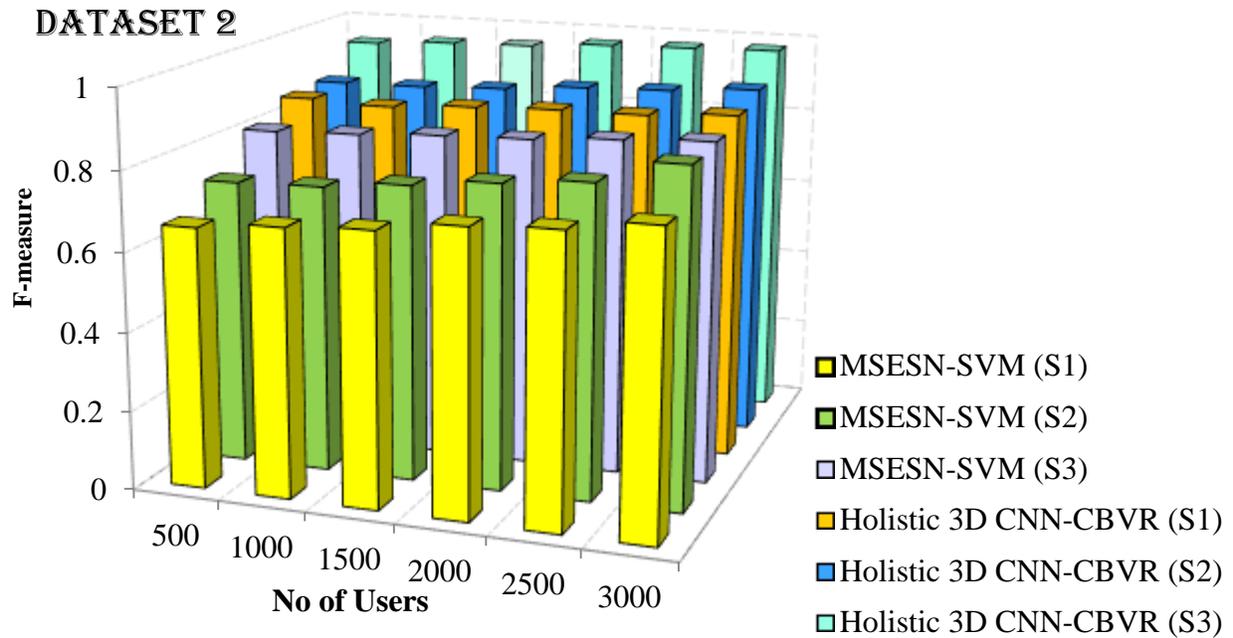
NAMA3DS1-CoSpaD (III)	0.837	0.759	0.89	0.815	0.962	0.865
-----------------------	-------	-------	------	-------	--------------	--------------

5.5.3.3 F-measure

F-measure is the mutual value that indicates the performance of precision and recall for the given query. In other words, it is known as the second external quality measure since it is computed after the precision and recall. For the proposed holistic 3D CNN for any class is the maximum value that obtains the node in the tree and an overall value for the f-measure is computed by the average weight of all precision and recall values.



(a)



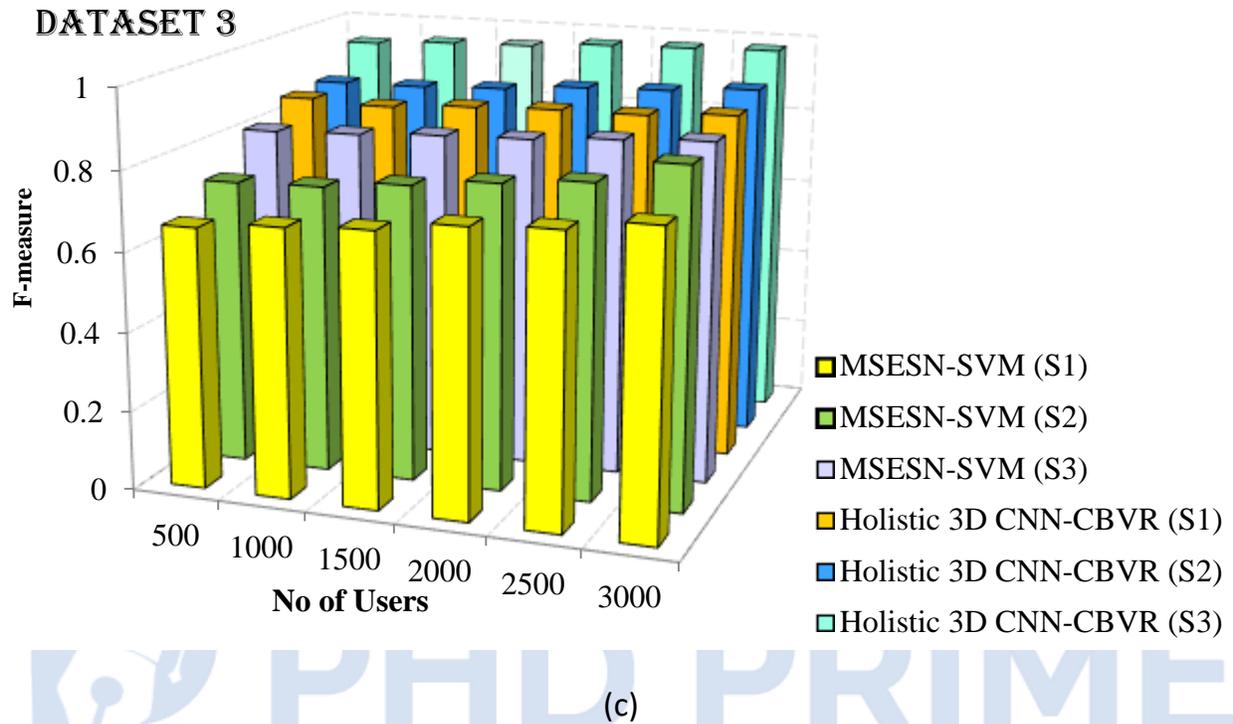


Figure 5.21 Performance of F-measure

The performance of the f-measure is depicted for three different scenarios with respect to the number of users. This speaks that the representation of holistic 3D CNN models has the capability to model the dependencies between the frames of a video that related to the given query. This capability results the simpler result for feature extraction and classification.

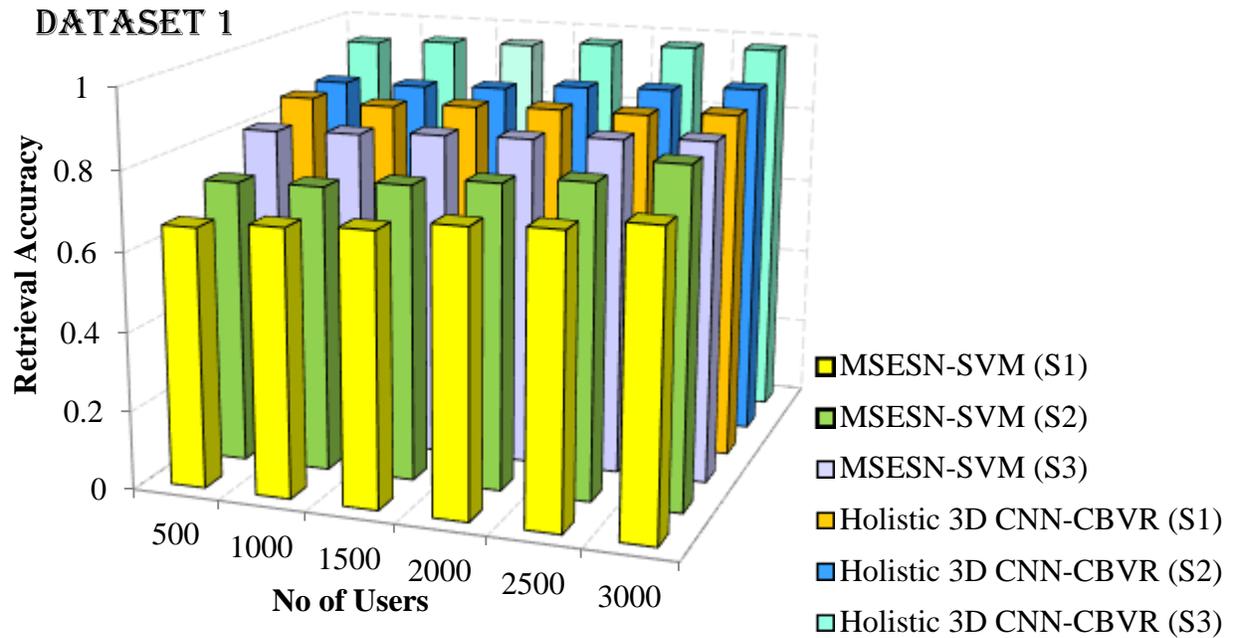
Table 5.7 Statistical Analysis on F-measure

	Scenario 1	Scenario 2	Scenario 3

Datasets	Holistic 3D CNN - CBVR	MSESN- SVM	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN- SVM
Hollywood 3D (I)	0.84	0.77	0.88	0.8	0.985	0.867
YouTube 8M segments (II)	0.85	0.775	0.89	0.82	0.981	0.875
NAMA3DS1- CoSpaD (III)	0.857	0.789	0.91	0.835	0.982	0.885

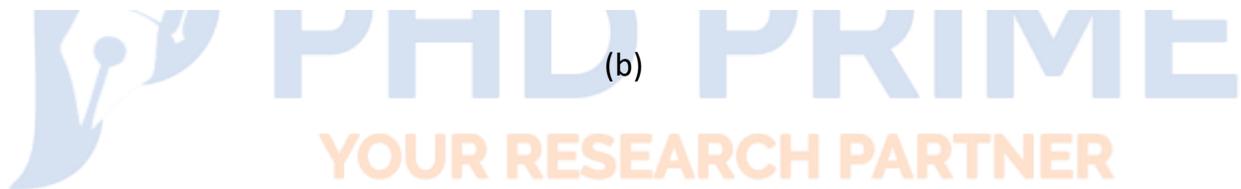
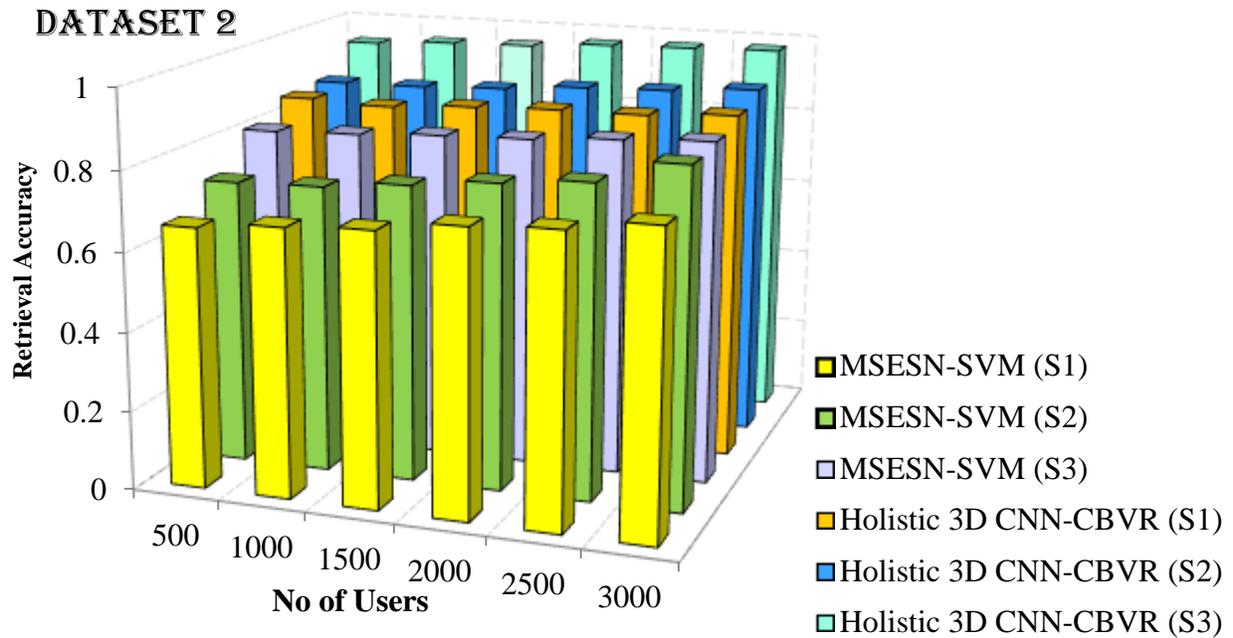
5.5.3.4 Retrieval Accuracy

Figure 5.22 confirms the benefit of the retrieval accuracy over the wide range of previous and existing models. When feature extraction, noise filtering, classification applied then the performance of the retrieval accuracy becomes higher. In this work, the proposed holistic 3D CNN has applied the different features for various categories. A simple trick to improve the retrieval accuracy is that maximize the performance of information retrieval. In earlier works, query results are not satisfied by the users due to no standard of how queries must be processed for search. Figure 5.22 represents the advantage of the proposed holistic 3D CNN which uses no SQL database for storage and query results have approximately 98% of the retrieval accuracy. Further, the statistical analysis of the proposed and previous works is illustrated in table.



(a)

YOUR RESEARCH PARTNER



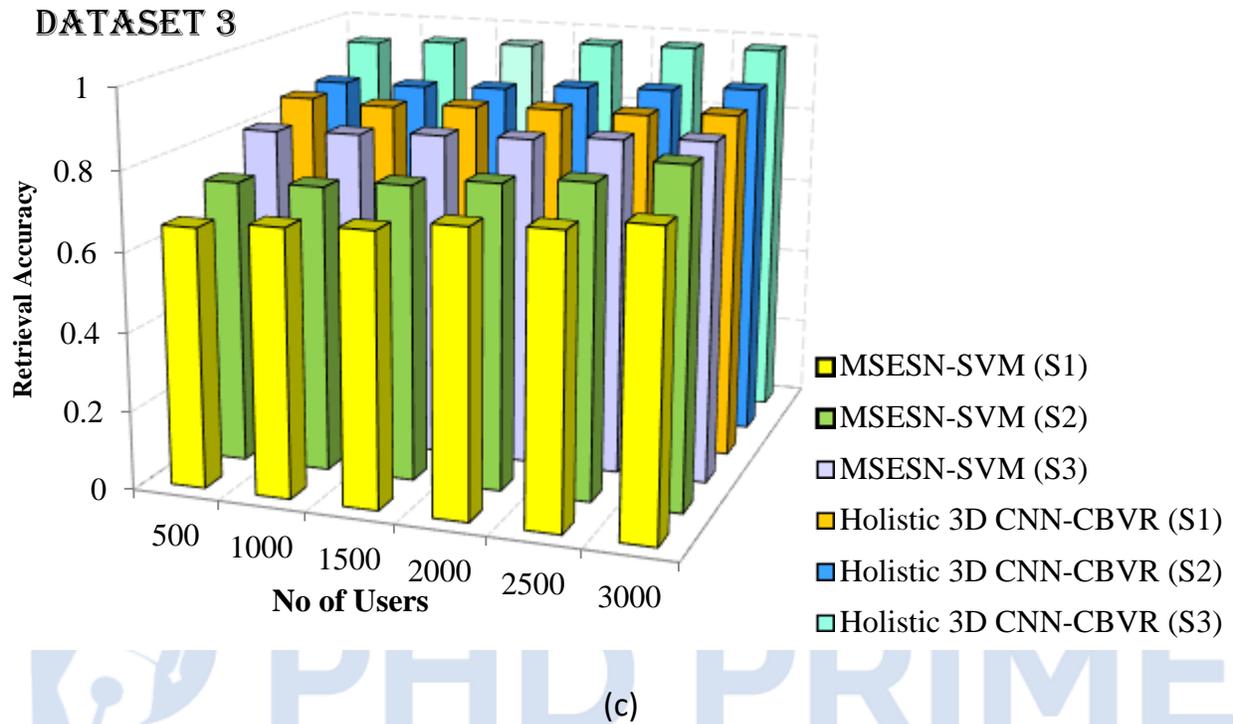


Figure 5.22 Performance of Retrieval Accuracy

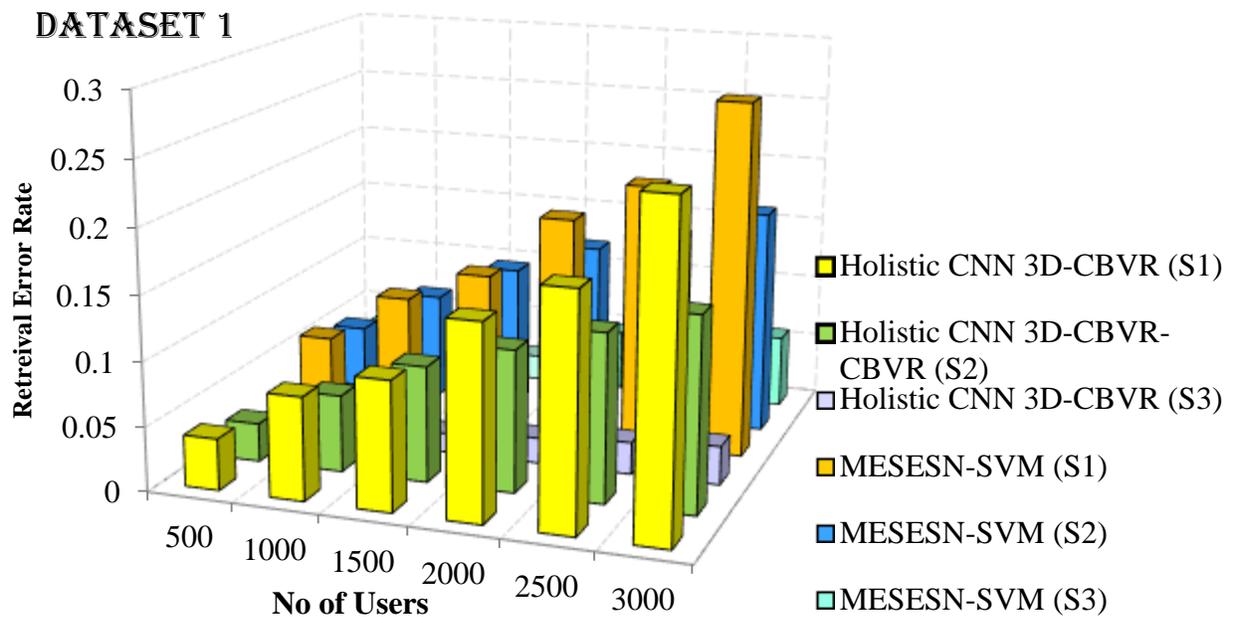
Table 5.8 Statistical Analysis on Retrieval Accuracy

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN -CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM
Hollywood 3D (I)	91.7	88.15	95.5	89	99.2	91
YouTube 8M segments (II)	91.8	88.45	95.9	89.2	99.3	92.1

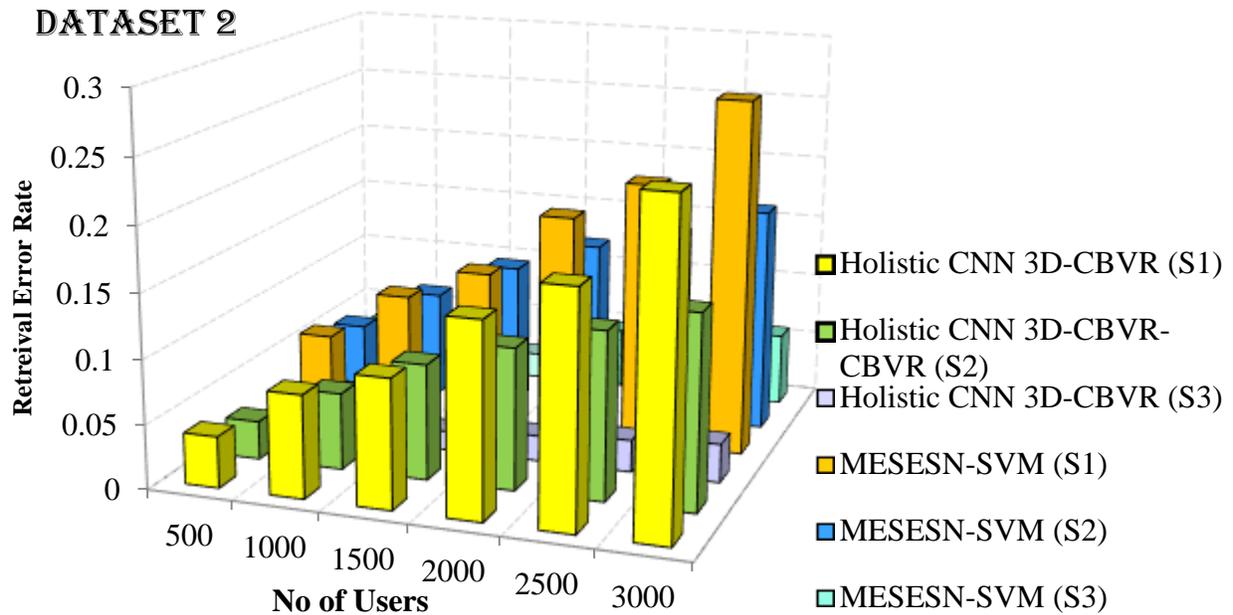
NAMA3DS1- CoSpaD (III)	92	88.76	96	89.5	99.6	92.45
---------------------------	----	-------	----	------	------	-------

5.5.3.5 Retrieval Error Rate

It deals when the representation of the unstructured queries are handled for the storage and retrieval system. Classical retrieval systems are mainly deals with the text based queries.

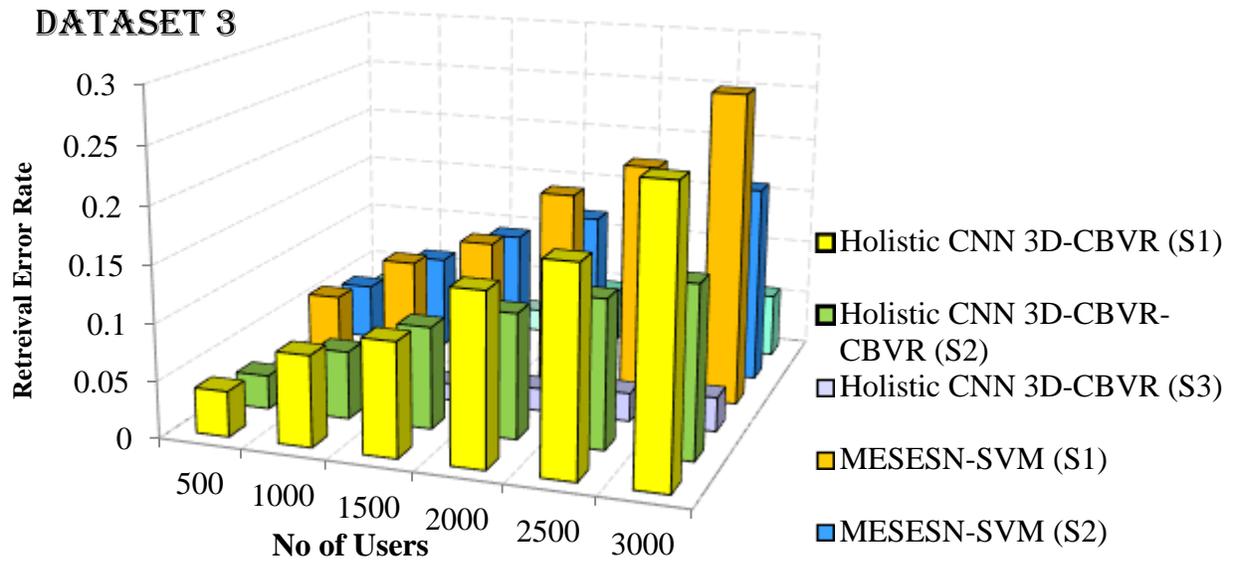


(a)



(b)

The higher the error rate shows the poor retrieval performance and proves the deficiency of the feature extraction, selection and classification procedure. Figure 5.23 shows the performance of the retrieval error rate with respect to the number of users and three scenarios of the retrieval error rate reflect the lower error rate for the proposed work performance.



(c)

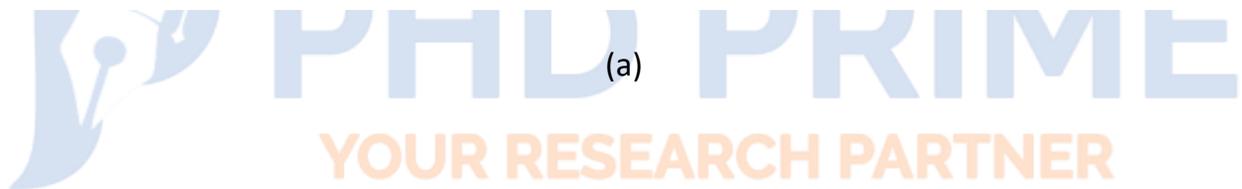
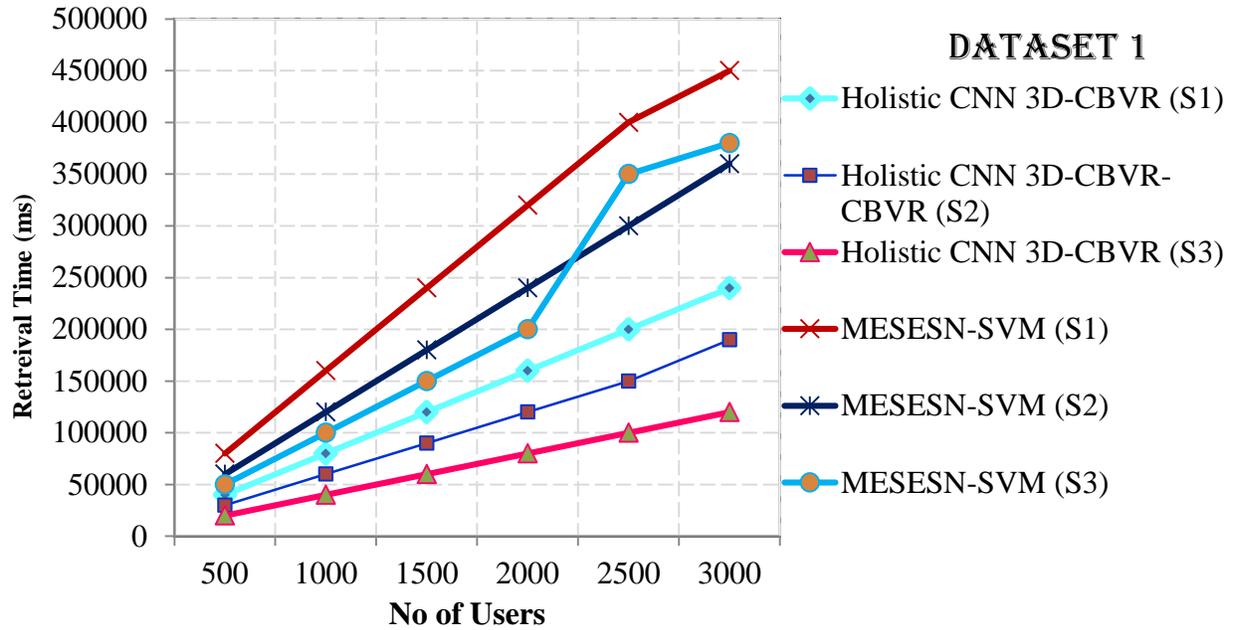
Figure 5.23 Performance of Retrieval Error Rate

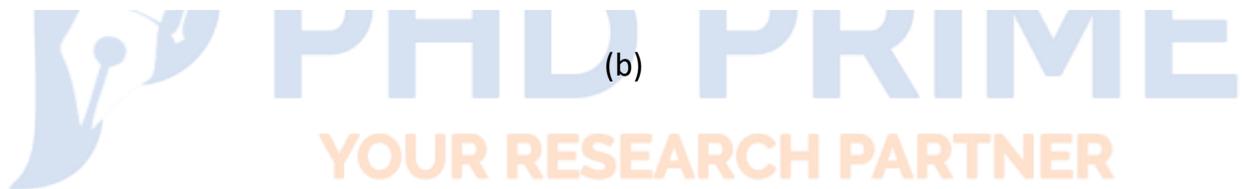
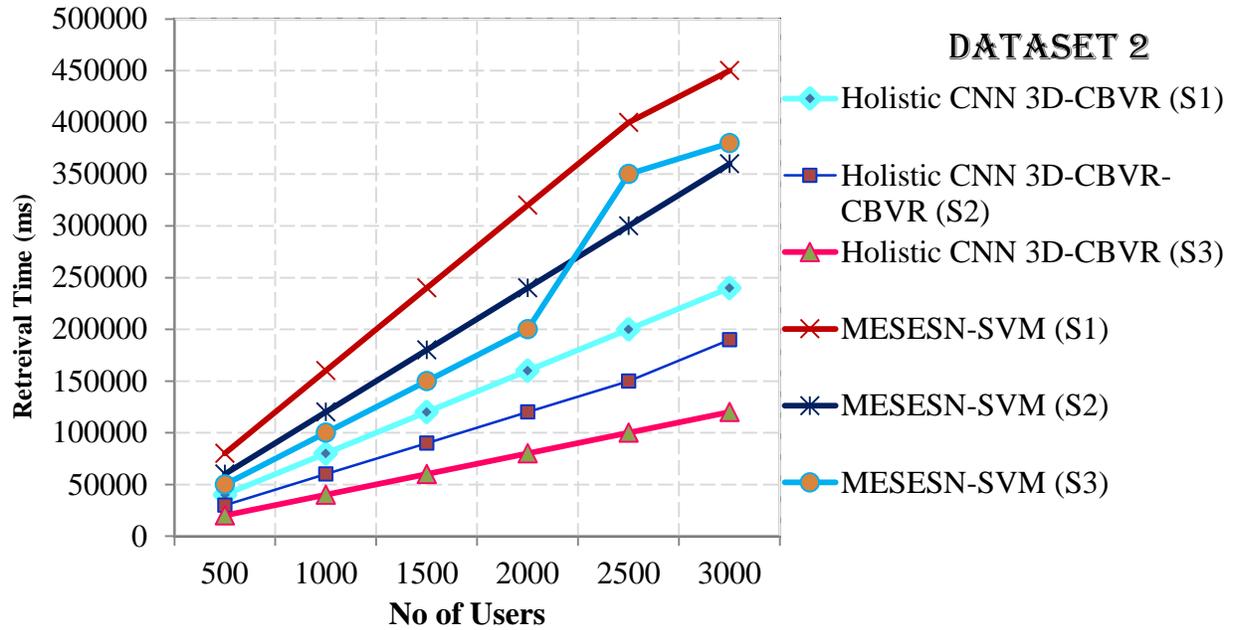
Table 5.9 Statistical Analysis on Retrieval Error Rate

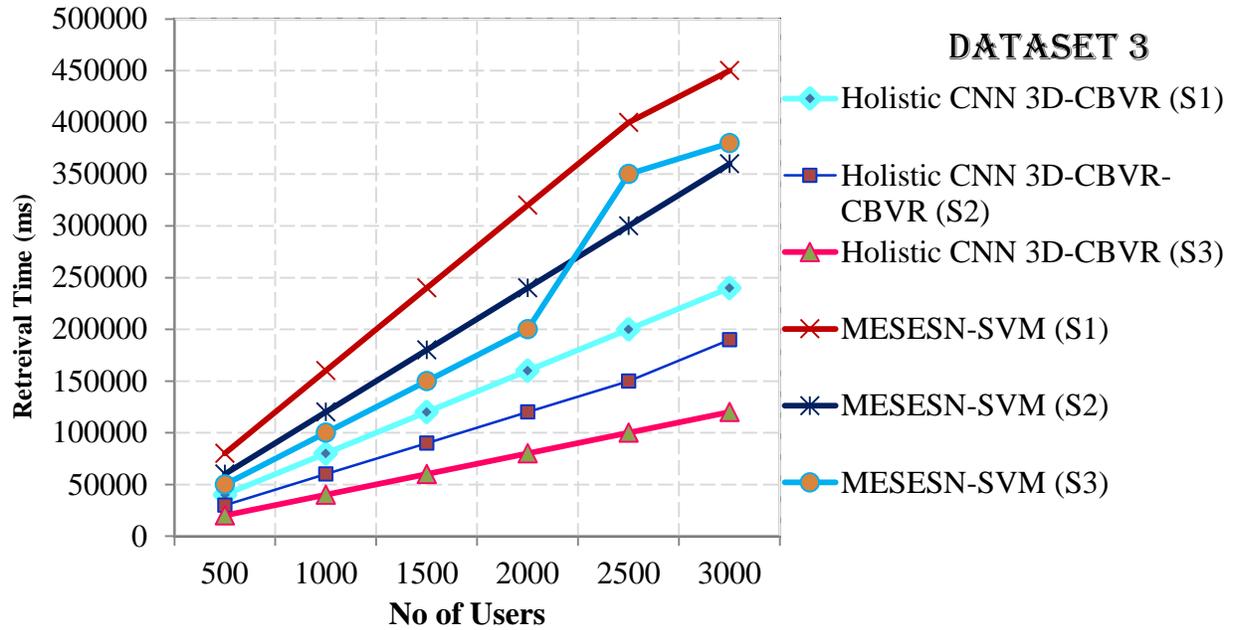
Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM
Hollywood 3D (I)	0.036	0.039	0.03	0.09	0.012	0.08
YouTube 8M segments (II)	0.036	0.045	0.03	0.15	0.014	0.16
NAMA3DS1-CoSpaD (III)	0.04	0.047	0.038	0.17	0.02	0.2

5.5.3.6 Retrieval Time

For the given user given query, the required information is that the relevant videos in a short span of time. In order to address this issue, feature extraction, importance of each feature determination, indexing, clustering and classification are introduced. When the work is uses all these methods which lead to the overhead of computation also. Hence, lightweight mechanisms are used in the proposed holistic CNN 3D CBVR. Using light weight algorithms reduce the time and also the proposed holistic uses less number of layers for feature extraction and also mapreduce model runs in the parallel mode and thus the retrieval time is reduced two times than the previous works. Retrieval time is measured in milliseconds and it is plotted in a graph by figure 5.24 and also the statistical analysis is given in table.







(c)

Figure 5.24 Performance of Retrieval Time

Table 5.10 Statistical Analysis on Retrieval Time

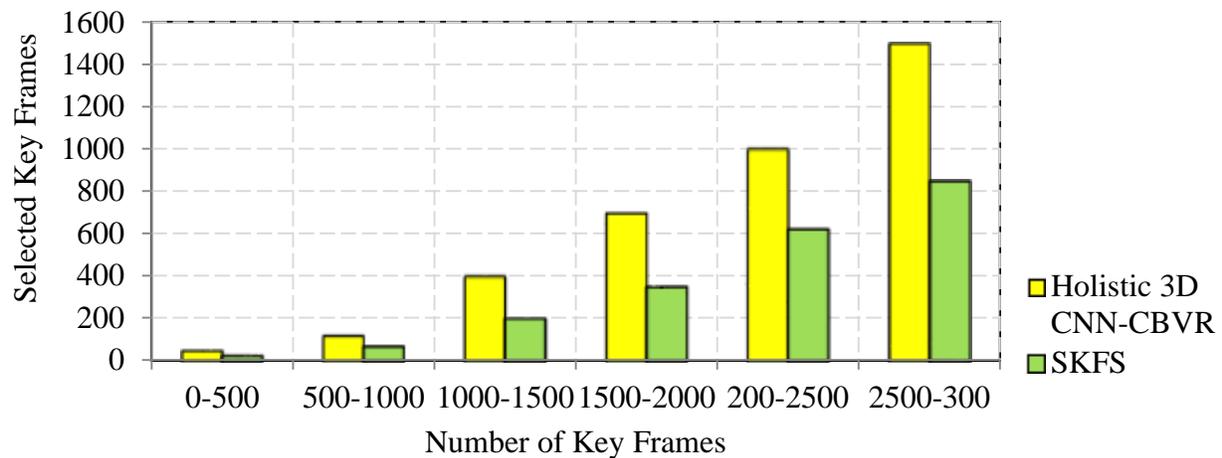
Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM
Hollywood 3D (I)	402000	1201050	301000	801020	202010	401050
YouTube 8M	402024	1201062	302014	801052	202084	401062

segments (II)						
NAMA3DS1- CoSpaD (III)	402020	1201075	302010	801075	204000	401075

5.5.3.7 Key Frame Selection

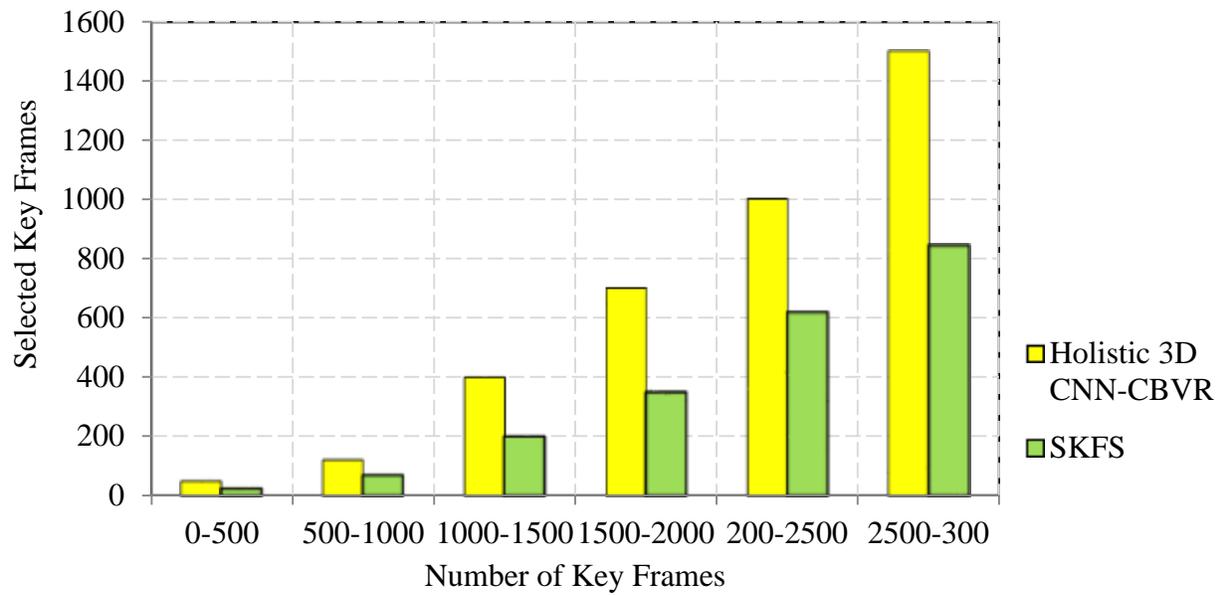
The performance of the key frame selection is evaluated for the total number of frames. In this process, the key frames are selected using entropy values. Figure 5.25 shows the advantage of the proposed work with respect to the selected key frames. The precision of a key frame is computed and the higher precision key frames are selected for retrieval of a relevant video.

DATASET 1

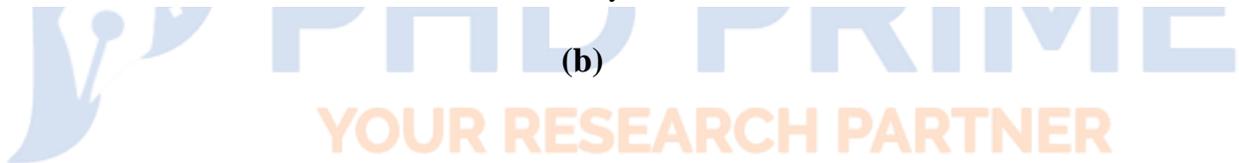


(a)

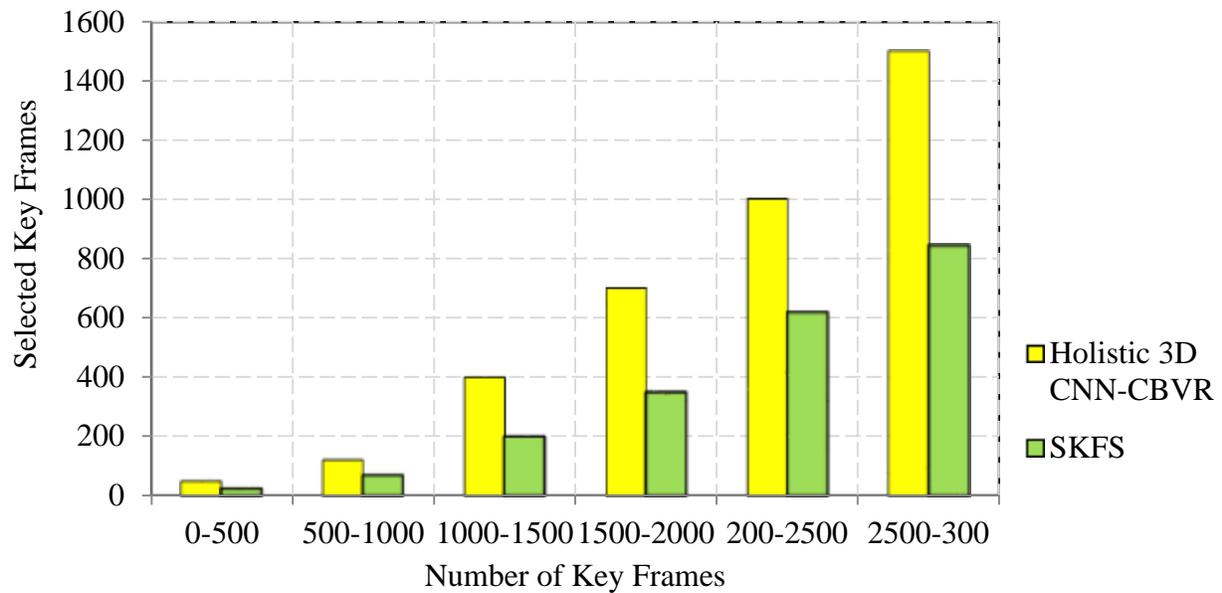
DATASET 2



(b)



DATASET 3



(c)

Figure 5.25 Performance of Key Frame Selection

Table 5.11 Statistical Analysis on Key Frame Selection

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM	Holistic 3D CNN -CBVR	MSESN-SVM
Hollywood 3D (I)	120	70	700	350	1500	845

YouTube 8M segments (II)	118	72	685	325	1480	850
NAMA3DS1-CoSpaD (III)	125	70	695	350	1490	865

5.5.3.8 Noise Filtering Accuracy

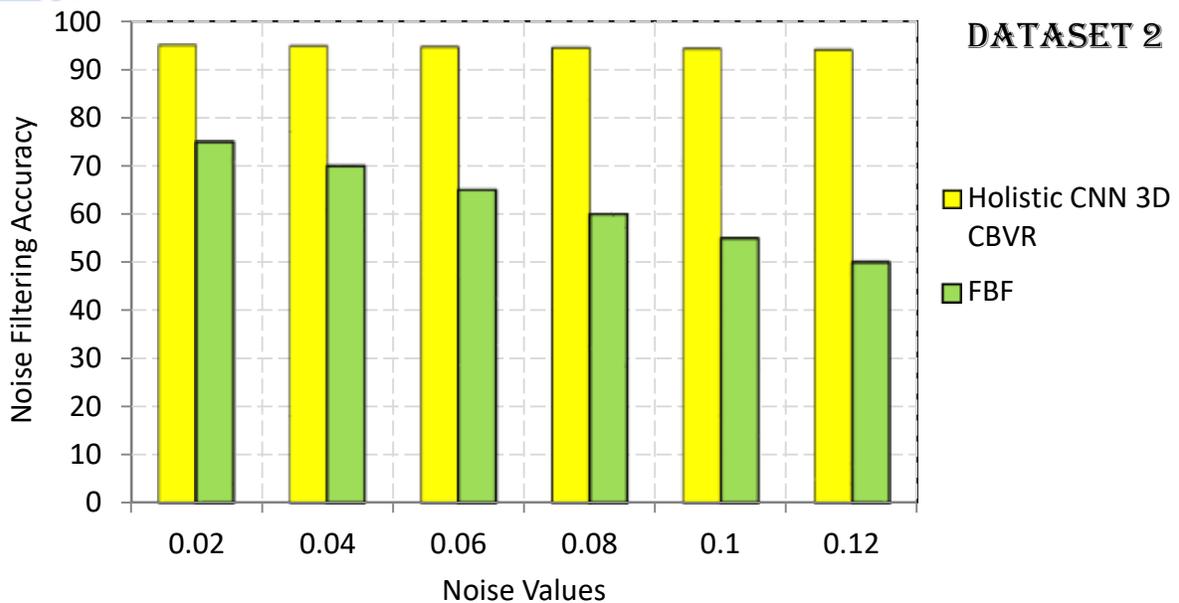
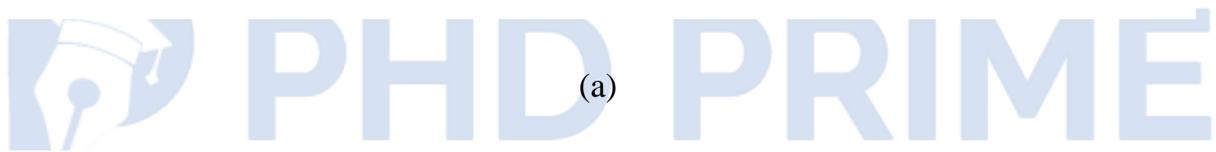
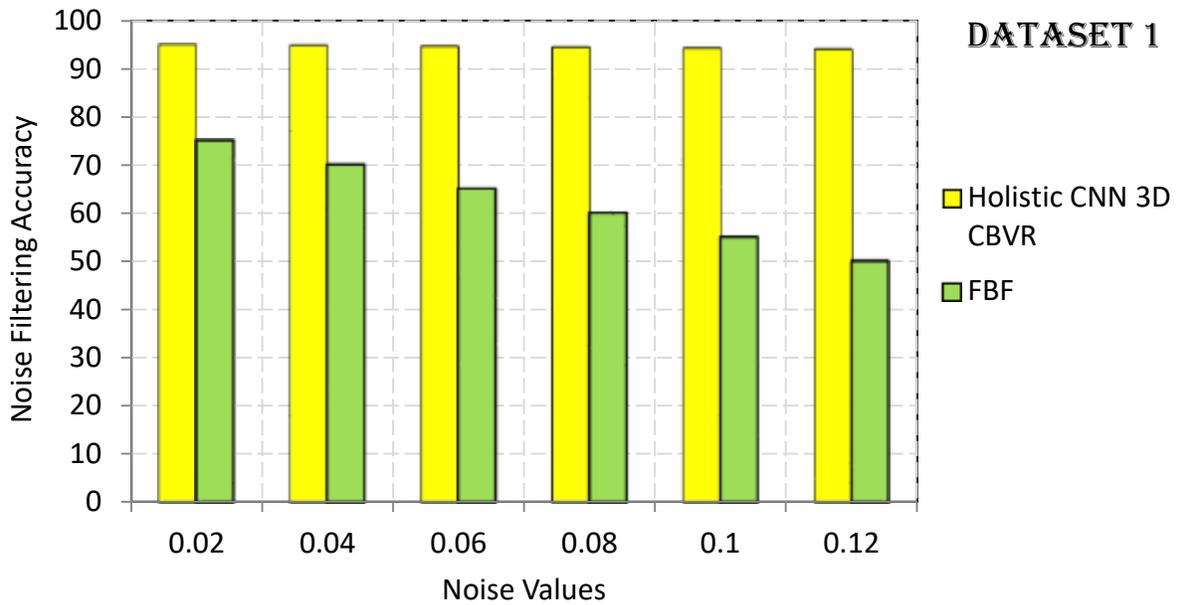
This measure finds the rate of accuracy in noise reduction. However, noise filtering accuracy is computed by the added noise level. Further, it is computed by Mean Square Error (MSE) which is done by,

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (29)$$

Where $\frac{1}{n} \sum_{i=1}^n$ is the mean, $(\hat{Y}_i - Y_i)^2$ is the square of errors.

$$NFA = \left(\frac{1}{MSE} \right) * 100 \quad (30)$$

The noise filtering accuracy is an additional credit measure that gives the retrieval accuracy for the given query. However, video has white Gaussian noise that consists of noisy pixels that minimize the video quality. Hence it is removed before to store into the Hadoop database. The advantage of the denoising is shown in figure 5.26. In this work, the performance of the bilateral median filter is clearly represented. After reducing the noise level and normalize the pixel values of the video, it is stored into the Hadoop database.



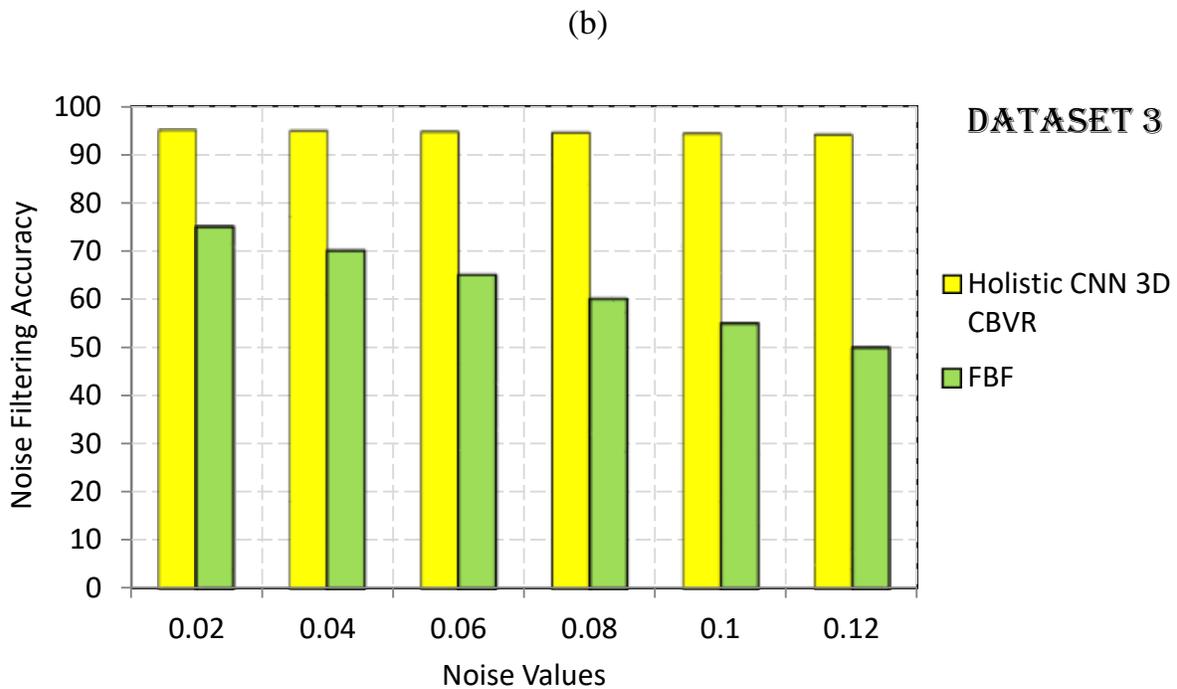


Figure 5.26 Performance of Noise Filtering Accuracy

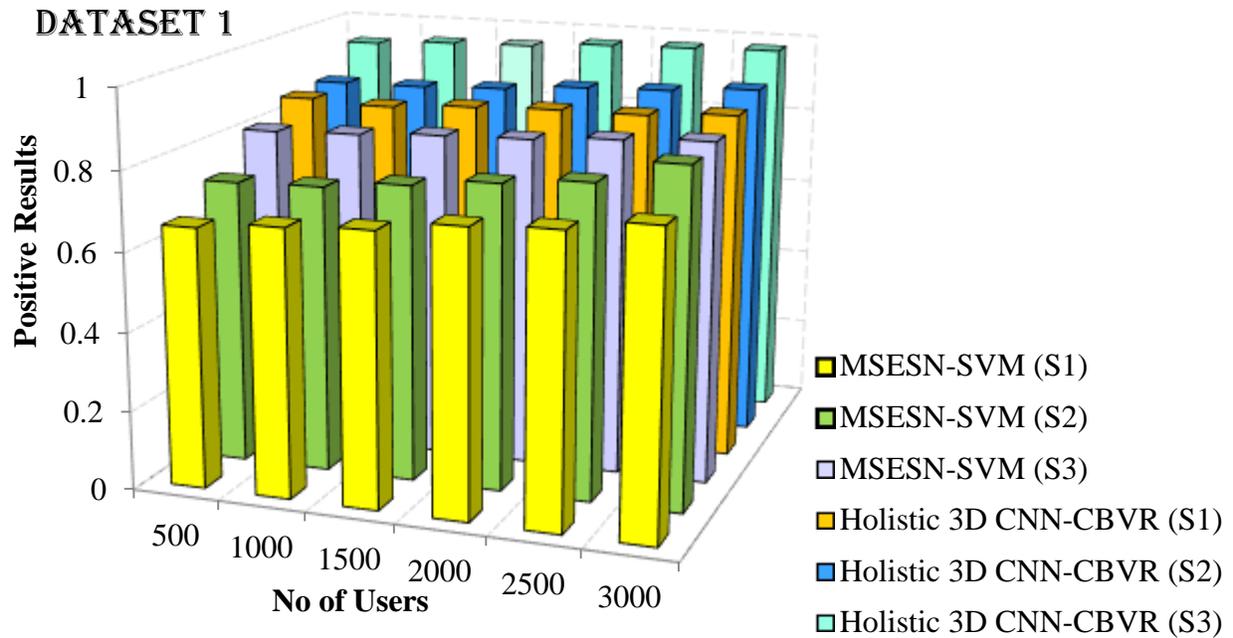
Table 5.12 Statistical Analysis on Noise Filtering Accuracy

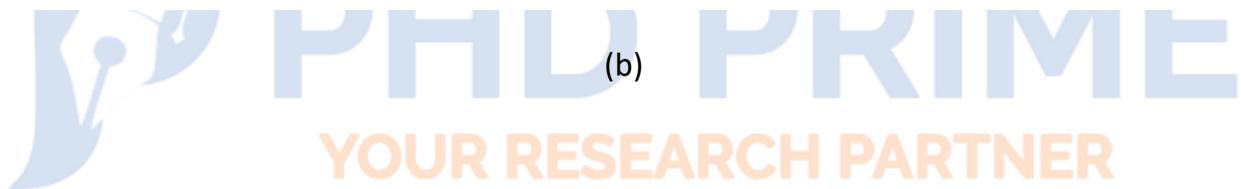
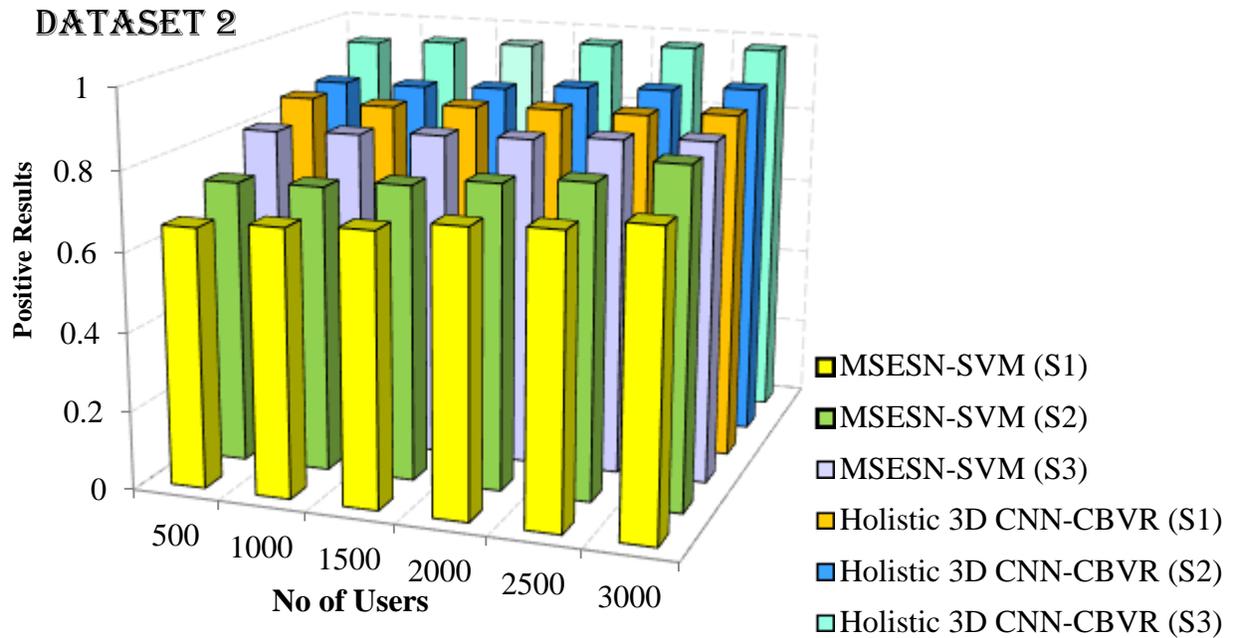
Datasets	0.04		0.08		0.12	
	Holistic 3D CNN - CBVR	FBF	Holistic 3D CNN -CBVR	FBF	Holistic 3D CNN -CBVR	FBF
Hollywood 3D (I)	94.8	70	94.4	60	94	50
YouTube 8M segments (II)	95	75	94.6	74	94.2	73

NAMA3DS1-CoSpaD (III)	95.2	75.3	94.9	74.8	94.6	74
-----------------------	------	------	------	------	------	----

5.5.3.9` Positive Results

A general information retrieval system requires higher positive results which mean that for a given query all results must be satisfied by the user and thus all collection should be actually relevant. An evaluation of relevant results for the search will result higher performance in CBVR. This suited for all 2D and 3D videos. Figure 5.27 represents the result of the positive result ratio with respect to the number of users. In holistic 3D CNN, scene, and object based features are extracted in pixelwise and the previous works extract the features by blockwise which does not preserve the object boundaries and poor semantic and visual information. Hence the performance of the proposed work will be higher.





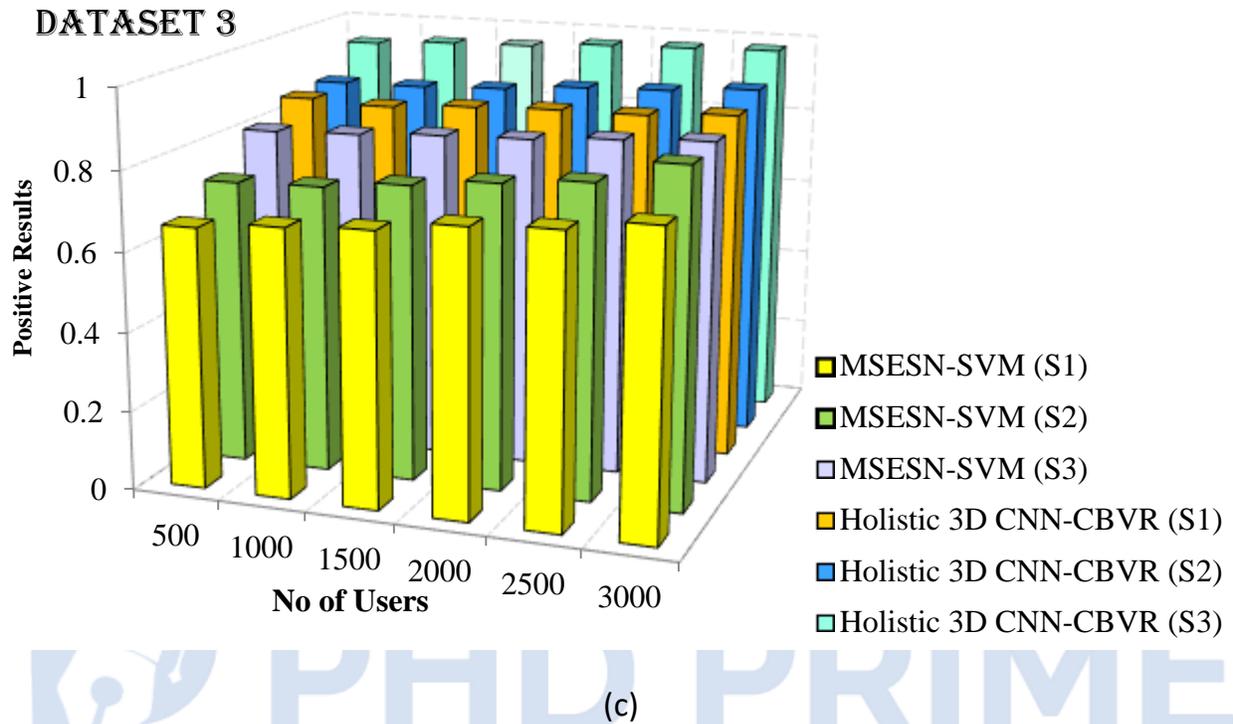


Figure 5.27 Performances of Positive Results

Table 5.13 Statistical Analysis on Positive Results

Datasets	Scenario 1		Scenario 2		Scenario 3	
	Holistic 3D CNN -CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM	Holistic 3D CNN -CBVR	MSESNSVM
Hollywood 3D (I)	92.45	88.32	94.55	89.15	99.89	90.68
YouTube 8M segments (II)	93.95	88.55	94.58	89.4	99.9	90.72

NAMA3DS1-CoSpaD (III)	93.88	88.76	94.77	89.7	99.9	90.78
-----------------------	-------	-------	-------	------	------	-------

5.5.3.10 Processing Time

MapReduce is modified in this work i.e. shuffle operation is additionally inserted between the Map and Reduce phases. Due to shuffle operation, the number of map and reduce operations are reduced. Table represents the MapReduce processing time for various scenarios to the training and testing.

Table 5.14 MapReduce Processing Time

Scenario	MapReduce Processing time (s)	
	Testing	Training
1	2500	3758
2	1800	2608
3	900	1200

MapReduce plays a major part in this proposed framework, table 2 depicts the processing time consumed in each scenario while performing MapReduce process. From this, it is clear that scenario three shows minimum processing time since all the nodes are involved in participation. Scenario 2 is twice the time of scenario 3, hence to tolerate a

huge number of data users scenario three is the best choice. In this analyzes 1500 samples are used and the possible number of users are served with relevant 3D videos.

Table 5.15 MapReduce Processing Time

Scenario	MapShuffleReduce Processing Time (s)	
	Testing	Training
1	1700	2758
2	900	1608
3	600	780

5.5.4 Obtained Output

The performance of this proposed 3D-CBVR is evaluated by measuring different metrics in three different scenarios. In simple, this process can be illustrated as user query based result retrieval system. According to the given query, relevant video are obtained and ranked in terms of most accurately matched videos. User given query is subjected to sequential processing and it matches with the training data for obtaining related 3D videos. The relevant videos are listed along with the length of each video i.e. their duration. Most relevant video is represented by providing higher number of stars and gradually the numbers of stars are minimized with respect to the relevancy of given query to that of the particular video.

Confusion matrix is computed for the proposed and the previous work for a scenario 3. It makes the prediction with respect to the outcome values i.e. the number of correct predictions for each category. Table represents the performance analysis of confusion matrix for the holistic 3D CNN and MSESN-SVM, respectively.

Table 5.16 Confusion Matrix for Holistic 3D CNN-CBVR (Scenario 3)

1	95.00	0.00	0.00	0.00	0.00	0.00	5.42	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	1.24	0.00	
2	0.00	95.65	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00	0.00	0.00	
3	1.45	3.12	97.25	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00	
4	0.00	0.00	4.15	97.84	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	3.14	
5	2.35	1.35	0.00	6.42	97.97	3.14	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	
6	4.15	0.00	6.42	0.00	4.15	97.85	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	5.42	
7	6.42	0.00	0.00	6.42	0.00	0.00	97.95	0.00	4.15	0.00	0.00	0.00	0.15	2.35	1.35	0.15	2.35	1.35	1.45	
8	0.15	2.35	1.35	2.35	1.35	0.00	0.00	98.1	0.00	4.15	1.35	1.35	4.15	0.15	2.35	1.35	4.15	0.15	2.35	
9	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.36	98.2	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.15	2.35	1.56	
10	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	98.7	1.35	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	
11	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.78	4.15	0.00	98.8	0.00	0.00	1.35	2.35	4.15	1.35	2.35	4.15	
12	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	2.31	98.9	0.00	4.15	2.35	4.15	0.15	2.35	1.35	
13	4.15	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	98.9	5.47	0.15	0.15	2.35	0.00	4.15	
14	0.00	6.42	0.00	2.35	1.35	0.00	2.35	1.35	0.00	1.47	3.47	0.00	7.98	99	0.00	0.00	0.15	2.35	0.00	
15	3.14	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.56	6.47	7.98	99	1.35	4.15	0.15	2.35	
16	2.35	1.35	0.00	1.02	4.15	0.00	0.15	2.35	0.00	0.00	0.15	2.35	0.00	0.00	5.78	99.5	1.35	4.15	2.35	
17	4.15	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.15	2.35	1.35	0.00	99.7	0.00	0.15	2.35	
18	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	1.03	0.00	99.7	0.00	0.15	
19	0.00	0.15	2.35	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	1.35	4.15	0.15	99.8	
20	4.15	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	2.00	3.45	1.36	0.00	99.9	0.15	2.35	1.35	2.35	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Table 5.17 Confusion Matrix for MSESN-SVM (Scenario 3)

1	59.00	0.00	0.00	0.00	0.00	0.00	5.42	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	1.24	0.00
2	0.00	59.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00	0.00	0.00
3	1.45	3.12	60.75	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	1.26	0.00	0.00
4	0.00	0.00	4.15	62	4.15	0.00	6.42	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	3.14	
5	2.35	1.35	0.00	6.42	63.45	3.14	0.00	6.42	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	
6	4.15	0.00	6.42	0.00	4.15	63.89	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	1.45	5.42
7	6.42	0.00	0.00	6.42	0.00	0.00	64.58	0.00	4.15	0.00	0.00	0.00	0.15	2.35	1.35	0.15	2.35	1.35	1.45
8	0.15	2.35	1.35	2.35	1.35	0.00	0.00	78.56	0.00	4.15	1.35	1.35	4.15	0.15	2.35	1.35	4.15	0.15	2.35
9	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.36	82.45	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.15	2.35	1.56
10	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	82	1.35	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.15	2.35	1.35	1.78	4.15	0.00	86.2	0.00	0.00	1.35	2.35	4.15	1.35	2.35	4.15
12	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	2.31	87.5	0.00	4.15	2.35	4.15	0.15	2.35	1.35
13	4.15	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	88.6	5.47	0.15	0.15	2.35	0.00	4.15
14	0.00	6.42	0.00	2.35	1.35	0.00	2.35	1.35	0.00	1.47	3.47	0.00	7.98	89.8	0.00	0.00	0.15	2.35	0.00
15	3.14	6.42	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	4.56	6.47	7.98	90.5	1.35	4.15	0.15	2.35

16	2.35	1.35	0.00	1.02	4.15	0.00	0.15	2.35	0.00	0.00	0.15	2.35	0.00	0.00	5.78	91.52	1.35	4.15	2.35	2.35
17	4.15	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	91.25	0.00	0.15	2.35
18	0.00	6.42	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	0.00	0.00	0.00	0.00	0.00	1.03	0.00	92	0.00	0.15
19	0.00	0.15	2.35	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	0.00	0.00	0.00	0.00	1.35	4.15	0.15	92	0.00
20	4.15	0.00	0.00	0.00	0.00	0.00	0.00	0.15	2.35	1.35	2.00	3.45	1.36	0.00	99.9	0.15	2.35	1.35	2.35	93.7
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

5.6 CHAPTER SUMMARY

Holistic CNN 3D-CBVR is proposed with REFKFS (key frame selection), Adaptive Median Bilateral Filtering (key frame denoising), Holistic 3D CNN MapShuffleReduce (multi-feature extraction) and Multi-feature lightweight method (similarity matching) for accurate video retrieval to the given query. The performance of proposed Holistic CNN 3D-CBVR is evaluated with the following metrics precision, recall, accuracy, retrieval error rate, time overhead, noise filtering accuracy, key frame selection and positive results. The performance of Holistic CNN 3D-CBVR is improved by appropriate resource utilization in key frame selection by REFKFS, effortless key frame denoising by filtering with Bayesian threshold due to the process applied only on selected key frames, robustness of the multiple features are evaluated under noisy conditions, highly relevant retrieval response for users by guaranteed multi feature extraction, fast retrieval response by MapShuffleReduce, simple similarity matching by lightweight matching. The compatibility of proposed Holistic 3D CNN MapShuffleReduce is verified against MSESN-SVM, FBF and SKFS.



CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT



In semantic web, Content Based Video Retrieval (CBVR) provides the visual co-occurrence relationships between contents with the support of semantic word similarity. In past days, various technologies are developed for effective storage and retrieval of digital 3D video content. Hence, the conventional techniques have retrieved large number of irrelevant information and are not capable to satisfy user's requirements. We have proposed a novel 3D CBVR in Map Reduce based Hadoop environment for designing innovative and new 3D video retrieval mechanisms which overcome the limitations and issues of existing schemes and video retrieval techniques. Our proposed system is designed intelligently for retrieving the most relevant 3D videos to user based on their given query image. The involvement of Map Reduce framework in this system essentially speeds up the processing time and reduces storage complexity of 3D video retrieval process.

3D CBVR has following stages which include key frame extraction, Denoising, Feature extraction, code book generation and similarity matching. On performing these processes, 3D video is retrieved effectively and also solves problems / challenges existed in previous approaches. The achievements of 3D CBVR is listed below:

- Complexity reduction and processing time minimization during key frame extraction even in case of increased number of video frames
- Effectively extracts features (shape, color and texture) with topology and geometry and uses four features (shape, color, motion and texture)

- Increased accuracy of retrieved result from combinational similarity matching process

Finally the experimental results confirmed that this proposed research work significantly reduces computational time and storage space compared with existing 3D video retrieval process. Also, this system outperforms most of the existing 3D video retrieval algorithms in terms of precision and recall. This overall process improves the research results with respect to accuracy, searching time, storage and computation complexity. We have affirmed that this system achieved better results due to Hadoop Map Reduce based 3D CBVR which consumes less time for video retrieval process.

- In future, we have planned to concentrate on high resolution videos in this research of 3D CBVR process since it improves the retrieval processing in real time environment.
- Further this research work will include automatic key frame extraction which is required to choose optimal key frames and balance the representation size (frame rate) against distortions
- As a future direction we can incorporate the duplicate video detection process for further reducing the processing time.

REFERENCES

Tzelepi, M., & Tefas, A. (2018). Deep convolutional learning for Content Based Image Retrieval. *Neurocomputing*, 275, 2467-2478.

Aziz, M.A., Ewees, A.A., & Hassanien, A. (2018). Multi-objective whale optimization algorithm for content-based image retrieval. *Multimedia Tools and Applications*, 77, 26135-26172.

Saritha, R., Paul, V., & Kumar, P.G. (2018). Content based image retrieval using deep learning process. *Cluster Computing*, 22, 4187-4200.

Mistry, Y., Ingole, D., & Ingole, M.D. (2018). Content based image retrieval using hybrid features and various distance metric. *Journal of Electrical Systems and Information Technology*, 5, 874-888.

Wan, J., Wang, D., Hoi, S., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014). Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. *MM '14*.

Zhu, L., Shen, J., Xie, L., & Cheng, Z. (2017). Unsupervised Visual Hashing with Semantic Assistant for Content-Based Image Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 29, 472-486.

Xia, Z., Xiong, N., Vasilakos, A., & Sun, X. (2017). EPCBIR: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing. *Inf. Sci.*, 387, 195-204.

Murala, S., Maheshwari, R.P., & Balasubramanian, R. (2012). Local Tetra Patterns: A New Feature Descriptor for Content-Based Image Retrieval. *IEEE Transactions on Image Processing*, 21, 2874-2886.

Mirza, T., Indoria, R., Dalwani, G., & Jain, K. (2016). Content based Image Retrieval using Color and Texture. *International Journal of Computer Applications*, 141, 5-8.

Guo, J., & Prasetyo, H. (2015). Content-Based Image Retrieval Using Features Extracted From Halftoning-Based Block Truncation Coding. *IEEE Transactions on Image Processing*, 24, 1010-1024.

Piras, L., & Giacinto, G. (2017). Information fusion in content based image retrieval: A comprehensive overview. *Inf. Fusion*, 37, 50-60.

Fadaei, S., Amirfattahi, R., & Ahmadzadeh, M. (2017). New content-based image retrieval system based on optimised integration of DCD, wavelet and curvelet features. *IET Image Process.*, 11, 89-98.

Alsmadi, M. (2017). An efficient similarity measure for content based image retrieval using memetic algorithm. *Egyptian Journal of Basic and Applied Sciences*, 4, 112 - 122.

AlZu'bi, A., Amira, A., & Ramzan, N. (2017). Content-based image retrieval with compact deep convolutional features. *Neurocomputing*, 249, 95-105.

Srivastava, P., & Khare, A. (2017). Integration of wavelet transform, Local Binary Patterns and moments for content-based image retrieval. *J. Vis. Commun. Image Represent.*, 42, 78-103.

Ashraf, R., Ahmed, M., Jabbar, S., Khalid, S., Ahmad, A., Din, S., & Jeon, G. (2017). Content Based Image Retrieval by Using Color Descriptor and Discrete Wavelet Transform. *Journal of Medical Systems*, 42, 1-12.

Liang, R., Shi, L., Wang, H., Meng, J., Wang, J., Sun, Q., & Gu, Y. (2016). Optimizing top precision performance measure of content-based image retrieval by learning similarity function. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2954-2958.

Dubey, S., Singh, S., & Singh, R. (2016). Multichannel Decoded Local Binary Patterns for Content-Based Image Retrieval. *IEEE Transactions on Image Processing*, 25, 4018-4032.

Liu, G., & Yang, J. (2013). Content-based image retrieval using color difference histogram. *Pattern Recognit.*, 46, 188-198.

Liu, P., Guo, J., Wu, C., & Cai, D. (2017). Fusion of Deep Learning and Compressed Domain Features for Content-Based Image Retrieval. *IEEE Transactions on Image Processing*, 26, 5706-5717.

Bala, A., & Kaur, T. (2016). Local texton XOR patterns: A new feature descriptor for content-based image retrieval. *Engineering Science and Technology, an International Journal*, 19, 101-112.

Memon, M., Li, J., Memon, I., & Arain, Q. (2016). GEO matching regions: multiple regions of interests using content based image retrieval based on relative locations. *Multimedia Tools and Applications*, 76, 15377-15411.

Kundu, M., Chowdhury, M., & Bulò, S.R. (2015). A graph-based relevance feedback mechanism in content-based image retrieval. *Knowl. Based Syst.*, 73, 254-264.

Ashraf, R., Bajwa, K., Irtaza, A., & Mahmood, M. (2015). Content Based Image Retrieval Using Embedded Neural Networks with Bandletized Regions. *Entropy*, 17, 3552-3580.

Yue, J., Li, Z., Liu, L., & Fu, Z. (2011). Content-based image retrieval using color and texture fused features. *Math. Comput. Model.*, 54, 1121-1127.

Younus, Z.S., Mohamad, D., Saba, T., Alkawaz, M.H., Rehman, A., Al-Rodhaan, M., & Al-Dhelaan, A. (2014). Content-based image retrieval using PSO and k-means clustering algorithm. *Arabian Journal of Geosciences*, 8, 6211-6224.

Rashedi, E., Nezamabadi-pour, H., & Saryazdi, S. (2013). A simultaneous feature adaptation and feature selection method for content-based image retrieval systems. *Knowl. Based Syst.*, 39, 85-94.

Ferreira, B., Rodrigues, J., Leitão, J., & Domingos, H. (2019). Practical Privacy-Preserving Content-Based Retrieval in Cloud Image Repositories. *IEEE Transactions on Cloud Computing*, 7, 784-798.

Rahimi, M., & Moghaddam, M. (2015). A content-based image retrieval system based on Color Ton Distribution descriptors. *Signal, Image and Video Processing*, 9, 691-704.

Bagri, N., & Johari, P.K. (2015). A Comparative Study on Feature Extraction using Texture and Shape for Content Based Image Retrieval. *International journal of advanced science and technology*, 80, 41-52.

Pedronette, D.C., Almeida, J., & Torres, R. (2014). A scalable re-ranking method for content-based image retrieval. *Inf. Sci.*, 265, 91-104.

Agarwal, S., Verma, A., & Singh, P. (2013). Content Based Image Retrieval using Discrete Wavelet Transform and Edge Histogram Descriptor. *2013 International Conference on Information Systems and Computer Networks*, 19-23.

Agarwal, S., Verma, A., & Dixit, N. (2014). Content Based Image Retrieval using Color Edge Detection and Discrete Wavelet Transform. *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 368-372.

Shrivastava, N., & Tyagi, V. (2014). Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching. *Inf. Sci.*, 259, 212-224.

Yildizer, E., Balci, A., Hassan, M., & Alhadj, R. (2012). Efficient content-based image retrieval using Multiple Support Vector Machines Ensemble. *Expert Syst. Appl.*, 39, 2385-2396.

Murala, S., & Wu, Q. (2014). Expert content-based image retrieval system using robust local patterns. *J. Vis. Commun. Image Represent.*, 25, 1324-1334.

Mühling, M., Meister, M., Korfhage, N., Wehling, J., Hörth, A., Ewerth, R., & Freisleben, B. (2018). Content-based video retrieval in historical collections of the German Broadcasting Archive. *International Journal on Digital Libraries*, 20, 167-183.

Ansari, A., & Mohammed, M. (2015). Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges. *International Journal of Computer Applications*, 112, 13-22.

Mohamadzadeh, S., & Farsi, H. (2016). CONTENT BASED VIDEO RETRIEVAL BASED ON HDWT AND SPARSE REPRESENTATION. *Image Analysis & Stereology*, 35, 67-80.

Hong, S., Im, W., & Yang, H. (2018). CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*.

Bhaumik, H., Bhattacharyya, S., Nath, M.D., & Chakraborty, S. (2016). Hybrid soft computing approaches to content based video retrieval: A brief review. *Appl. Soft Comput.*, 46, 1008-1029.

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A Survey on Visual Content-Based Video Indexing and Retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41, 797-819.

Ansari, A., & Mohammed, M. (2015). Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges. *International Journal of Computer Applications*, 112, 13-22.

Yang, H., & Meinel, C. (2014). Content Based Lecture Video Retrieval Using Speech and Video Text Information. *IEEE Transactions on Learning Technologies*, 7, 142-154.

Iyer, R.R., Parekh, S., Mohandoss, V., Ramsurat, A., Raj, B., & Singh, R. (2016). Content-based Video Indexing and Retrieval Using Corr-LDA. *ArXiv, abs/1602.08581*.

Rossetto, L., Giangreco, I., Schuldt, H., Dupont, S., Seddati, O., Sezgin, T.M., & Sahillioglu, Y. (2015). IMOTION - A Content-Based Video Retrieval Engine. *MMM*.

Raftopoulos, Konstantinos A., Klimis S. Ntalianis, Dionyssios D. Sourlas, and Stefanos D. Kollias. "Mining User Queries with Markov Chains: Application

to Online Image Retrieval." Knowledge and Data Engineering, IEEE Transactions, vol. 25, no. 2, pp 433-447, 2013.

Leung, KW-T., Wilfred Ng, and Dik Lun Lee. "Personalized concept-based clustering of search engine queries." Knowledge and Data Engineering, IEEE Transactions, vol. 20, no. 11 1505-1518, 2008.

Shaoting Zhang Ming Yang Timothee Cour Kai Yu Dimitris N. Metaxas, "Query Specific Fusion for Image Retrieval", Computer vision-ECCV, pp 660-673, 2012.

Hoi, Steven CH, Wei Liu, and Shih-Fu Chang. "Semi-supervised distance metric learning for collaborative image retrieval." In Computer Vision and Pattern Recognition, CVPR IEEE Conference on, pp. 1-7, 2008.

J.Premkumar, Mr.P.Prasenna, "Image Retrieval using Markovian Semantic indexing (MSI)", International Journal Of Engineering And Computer Science, vol. 4, issue 4, pp 11430-11433, 2015.

V.S.V.S. Murthy, E.Vamsidhar, J.N.V.R. Swarup Kumar, P.Sankara Rao, "Content Based Image Retrieval using Hierarchical and K-Means Clustering Techniques", International Journal of Engineering Science and Technology, vol. 2, issue. 3, pp 209-212, 2014.

Valiollahzadeh, S. M., A. Sayadiyan, and F. Karbassian. "Adaptive Boosting of Support Vector Machine Component Classifiers Applied in Face Detection", IEEE Conference Publications, 2007.

A.Kannan, Dr.V.Mohan, Dr.N.Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques, IEEE International Conference on Computational Intelligence and Computing Research, 2010.

Ahanger, Gulrukh, and Thomas DC Little. "Data semantics for improving retrieval performance of digital news video systems." Knowledge and Data Engineering, IEEE Transactions, vol. 13, no. 3, pp 352-360, 2001.

Yang Y, Zha ZJ, Gao Y, Zhu X, Chua TS. Exploiting web images for semantic video indexing via robust sample-specific loss. IEEE Transactions on Multimedia, vol. 16, issue. 6, pp 1677-1689, 2014.

Fan, Jianping, Hangzai Luo, Yuli Gao, and Ramesh Jain. "Incorporating concept ontology for hierarchical video classification, annotation, and visualization." Multimedia, IEEE Transactions, vol. 9, no. 5, pp 939-957, 2007.

Susu Shan, Hailiang Xu and Feng Su, "A New Method for Spatiotemporal Textual Saliency Detection in Video", IEEE 23rd International Conference on Pattern Recognition, pp 3240-3245, 2016.

Noel C. F. Codella, Gang Hua,, Liangliang Cao, “Large-Scale Video Event Classification Using Dynamic Temporal Pyramid Matching Of Visual Semantics”, IEEE conference, 2877-2881, 2013.

Ahanger, Gulrukh, and Thomas DC Little. "Data semantics for improving retrieval performance of digital news video systems." Knowledge and Data Engineering, IEEE Transactions, vol. 13, no. 3 pp 352-360, 2001.

Shuichi Shiitani, Takaydci Baba, Susumu Endo, Yusuke Uehara, Daiki Masumoto, and Shigemi Nagata, “Efficient Video Retrieval System Using Virtual 3’D Space”, IEEE Southwest symposium in Image analysis and Interpretation, pp 206-210, 2004.

Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman and David Dobkin, “A Search Engine for 3D Models,” ACM Transactions on Graphics, vol. 22, pp. 83-105, 2003.

Zongmin Li, Zijian Wu, Zhenzhong Kuang, Kai Chen, Yongzhou Gan and Jianping Fan, “Evidence-based SVM fusion for 3D model retrieval,” Multimedia Tools and Applications, vol. 72, pp. 1731-1749, 2013

Wichian Premchaiswadi, Anucha Tungkatsathan, Sarayut Intarasema, “Improving Performance of Content-Based Image Retrieval Schemes using Hadoop MapReduce,” in Proc. of IEEE International Conference, pp. 615 – 620, 2013.

Amirthalingam Ramanan and Mahesan Niranjan, “A Review of Codebook Models in Patch-Based Visual Object Recognition,” *Journal of Signal Processing System*, vol. 68, pp. 333-352, 2011.

Bernard Zenko, , Saso Dzeroski and , Jan Struyf, “Learning Predictive Clustering Rules,” *Knowledge Discovery in Inductive Databases*, vol. 3933, pp. 234-250, 2006.

Frank Moosmann, Eric Nowak and Frederic Jurie, “Randomized Clustering Forests for Image Classification,” *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 30, no. 9, pp. 1632 – 1646, 2008.

Athanasios Mademlis, Petros Daras, Apostolos Axenopoulos, Dimitrios Tzovaras and Michael G. Strintzis, “Combining Topological and Geometrical Features for Global and Partial 3-D Shape Retrieval,” *IEEE Transactions On Multimedia*, vol. 10, no. 5, pp. 819 – 831, 2008.

Yu-Gang Jiang, Chong-Wah Ngo and Jun Yang, “Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval,” in *Proc. of 6th ACM International Conference on Image and Video Retrieval*, pp.494-501, 2007.

Jonathon S. Hare, Sina Samangooei and Paul H. Lewis, “Practical scalable image analysis and indexing using Hadoop,” *Multimedia Tools and Applications*, vol.71, pp.1215-1248, 2012.

Brandyn White, Tom Yeh, Jimmy Lin, and Larry Davis, “Web-Scale Computer Vision using MapReduce for Multimedia Data Mining,” in Proc. of 10th ACM International Workshop on Multimedia Data mining, 2010.

Deepika Nagthane, “Content Based Image Retrieval Using K-means clustering technique”, International Journal of CAIT, Vol. 3, Issue I June-July 2013.

Huanbo Luan, Yan-Tao Zheng, Meng Wang and Tat-Seng Chua, “VisionGo: Towards video retrieval with joint exploration of human and computer”, Information Sciences, Vol. 181, No. 19, pp. 4197-4213, 2011.

Leung, KW-T., Wilfred Ng, and Dik Lun Lee. "Personalized concept-based clustering of search engine queries." Knowledge and Data Engineering, IEEE Transactions, 20, No. 11, pp 1505-1518, 2008.

Tuia, Devis, and Gustavo Camps-Valls. "Semisupervised remote sensing image classification with cluster kernels." Geoscience and Remote Sensing Letters, IEEE, vol. 6, No. 2 pp 224-228, 2009.

Ren, Chuan-Xian, Dao-Qing Dai, and Hong Yan. "Coupled kernel embedding for low-resolution face image recognition." Image Processing, IEEE Transactions , vol. 21, No. 8, pp 3770-3783, 2012.

Xia, Hao, and Steven CH Hoi. "Mkboost: A framework of multiple kernel boosting." Knowledge and Data Engineering, IEEE Transactions, vol. 25, No. 7 pp 1574-1586, 2013.

Hertz, Tomer, Aharon Bar-Hillel, and Daphna Weinshall. "Boosting margin based distance functions for clustering." In Proceedings of the twenty-first international conference on Machine learning, pp. 50. 2004.

He, Ben, and Iadh Ounis. "A Query-based Pre-retrieval Model Selection Approach to Information Retrieval." , In RIAO, pp. 706-719. 2004.

Zhang, Lining, Lipo Wang, Weisi Lin, and Shuicheng Yan. "Geometric Optimum Experimental Design for Collaborative Image Retrieval." IEEE Trans. Circuits Syst. Video Techn. Vol. 24, No. 2, pp 346-359, 2014.

Umesh K K and Suresha, "Web Image Retrieval Using Visual Dictionary," International Journal on Web Service Computing, vol. 3, No. 3, 2012.

Brandyn White, Tom Yeh, Jimmy Lin, and Larry Davis, "Web-Scale Computer Vision using MapReduce for Multimedia Data Mining," 10th ACM International Workshop on Multimedia Data mining, 2010.

Jialu Liu, "Image Retrieval based on Bag-of-Words model," Information Retrieval, 2013.

Hanli Wang, Yun Shen, Lei Wang, Kuangtian Zhufeng, Wei Wang and Cheng Cheng, "Large-Scale Multimedia Data Mining Using MapReduce Framework," IEEE 4th International Conference on Cloud Computing Technology and Science, pp 287 – 292, 2012.

Eric Brachmann, Marcel Spehr and Stefan Gumhold, “Feature Propagation on Image Webs for Enhanced Image Retrieval,” ACM International Conference on Multimedia Retrieval, pp 25-32, 2013.

KehuaGuo, WeiPan, MingmingLu, XiaokeZhou and JianhuaMa, “An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval,” Journal of System and Software, vol. 102, pp 207–216, 2014.

Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Saso Dzeroski, “Fast and Scalable Image Retrieval Using Predictive Clustering Trees,” Lecture Notes in Computer Science, vol. 8140, pp 33-48, 2013.

Frank Moosmann, Eric Nowak and Frederic Jurie, “Randomized Clustering Forests for Image Classification,” IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 30, no. 9, pp. 1632 – 1646, 2008.

William Robson Schwartz and Hélio Pedrini, “Color Textured Image Segmentation Based on Spatial Dependence Using 3D Co-occurrence Matrices and Markov Random Fields,” Journal of Science, vol. 53, no. 3, pp 693-702, 2012.

Arati S. Kurani, Dong-Hui Xu, Jacob Furst and Daniela Stan Raicu, “Co-Occurrence Matrices for Volumetric Data,” in Proc. of 7th International Conference on Computer Graphics and Imaging, 2014.

Chunlai Yan, “Accurate Image Retrieval Algorithm Based on Color and Texture Feature,” Journal Of Multimedia, vol. 8, no. 3, pp 277-283, 2013.

Md. Baharul Islam, Krishanu Kundu and Arif Ahmed, “Texture Feature based Image Retrieval Algorithms,” International Journal of Engineering and Technical Research, vol. 2, pp 71 – 75, 2014.

Pereira JC, Coviello E, Doyle G, Rasiwasia N, Lanckriet GRG, Levy R, Vasconcelos N, “On the role of correlation and abstraction in cross-modal multimedia retrieval”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 3, pp 521–535, 2014.

Zou W, Bai C, Kpalma K, Ronsin J, “Online glocal transfer for automatic figure-ground segmentation”, IEEE Transactions on Image Processing, vol. 23, No. 5, pp 2109–2121, 2014.

Shen X, Lin Z, Brandt J, Wu Y, “Spatially-constrained similarity measure for large-scale object retrieval”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No, 6, pp 1229–1241, 2014.

Zhang S, Yang M, Cour T, Yu K, Metaxas DN, “Query specific rank fusion for image retrieval”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, No. 4, pp 803–815, 2014.

Zhang S, Tian Q, Huang Q, Gao W, Rui Y, “Cascade category-aware visual search”, IEEE Transactions on Image Processing, vol. 23, No. 6, pp 2514–2527, 2014.

Wang D, Hoi SCH, He Y, Zhu J, “Mining weakly labeled web facial images for search-based face annotation” IEEE Transactions on Knowledge and Data Engineering, vol. 26, No. 1, pp 166–179,2014.

Liu Z, Li H, Zhang L, Zhou W, Tian Q, “Cross-indexing of binary SIFT codes for large-scale image search”, IEEE Transactions on Image Processing, vol. 23, No, 5, pp 2047–2057, 2014.

Hoi SCH, Lyu MR, “A semi-supervised active learning framework for image retrieval”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 302–309, 2005.

Andre B, Vercauteren T, Buchner AM, Wallace MB, Ayache N, “Learning semantic and visual similarity for endomicroscopy video retrieval”, IEEE Transactions on Medical Imaging, vol. 31, No, 6, pp 1276–88, 2012.

Hanjalic A, Lagendijk RL, Biemond J, “Automated high-level movie segmentation for advanced video-retrieval systems”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, No. 4, pp 580–588, . 1999.

Yuan-Hao L, Chuan-Kai Y, “Video object retrieval by trajectory and appearance”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, No. 6, pp 1026–1037, 2015.

Aslandogan YA, Yu CT, “Techniques and systems for image and video retrieval”, IEEE Transactions on Knowledge and Data Engineering, vol. 11, No. 1, pp 56–63, 1999.

Aasif Ansari, Muzammil H Mohammed, “Content based Video Retrieval Systems – Methods, Techniques, Trends and Challenges”, International Journal of Computer Applications, vol. 112, no. 7, pp 13 – 22, 2015.

Xiang Zhang, Siwei Ma, , Shiqi Wang, Xinfeng Zhang, Huifang Sun, Wen Gao, “A Joint Compression Scheme of Video Feature Descriptors and Visual Content”, IEEE Transactions on Image Processing, vol. 26, no. 2, pp 633 – 647, 2016.

Giuseppe Papari, Nasiru Idowu, Trond Varslot, “Fast bilateral filtering for denoising large 3D images”, IEEE Transactions on Image Processing, vol. 26, no. 1, pp 251 – 261, 2016.

Min-Kook Choi, ZiyuWang, Hyun-Gyu Lee, Sang-Chu, Lee, “A bag-of-regions representation for video classification”, Multimedia Tools and Applications, Springer, vol. 75, no. 5, pp 2453 – 2472, 2016.

Zhong-Min Huangfu, Shu-Sheng Zhang, Luo-Heng Yan, “A method of 3D CAD model retrieval based on spatial bag of words”, Multimedia Tools Application, Springer, pp.1 -29, 2016.

Amit Fegade, Vipul Dalal, “A Survey on Content Based Video Retrieval”, International Journal of Engineering and Computer Science, vol. 3, no. 7, pp 7271 – 7279.

Thanh Duc, Duy Dinh Le, Shinichi Satoh, “Scalable approaches for content based video retrieval”, The future of Multimedia Analysis and Mining, no. 11, pp 31 – 39, 2014.

Pradeep Chivadshetti, Kishor Sadafale, Kalpana Thakare, “Content Based Video Retrieval Using Integrated Feature Extraction and Personalization of Results”, International Conference on Information Processing, pp 170 – 175, 2015.

Thepade, S., & Yadav, N. (2015). Novel Efficient Content Based Video Retrieval Method Using Cosine-Haar Hybrid Wavelet Transform with Energy Compaction. *2015 International Conference on Computing Communication Control and Automation*, 615-619.

Nagaraja, G.S., Murthy, S.R., & Deepak, T.S. (2015). Content based video retrieval using support vector machine classification. *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 821-827.

Thepade, S., & Tonge, A.A. (2014). An optimized key frame extraction for detection of near duplicates in content based video retrieval. *2014 International Conference on Communication and Signal Processing*, 1087-1091.

Thepade, S., Subhedarpage, K.S., Mali, A.A., & Vaidya, T.S. (2013). Color Content based Video Retrieval using Block Truncation Coding with Different Color Spaces. *International Journal of Computer Applications*, 64, 35-38.

Kulkarni, P., Patil, B., & Joglekar, B. (2015). An effective content based video analysis and retrieval using pattern indexing techniques. *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, 87-92.

Asha, S., & Sreeraj, M. (2013). Content Based Video Retrieval Using SURF Descriptor. *2013 Third International Conference on Advances in Computing and Communications*, 212-215.

Tahboub, K., Gadgil, N., Comer, M., & Delp, E. (2014). An HEVC compressed domain content-based video signature for copy detection and video retrieval. *Electronic Imaging*.

Agharwal, A., Kovvuri, R., Nevatia, R., & Snoek, C.G. (2016). Tag-based video retrieval by embedding semantic content in a continuous word space. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1-8.

Thepade, S., Subhedarpage, K.S., & Mali, A.A. (2013). Performance rise in Content Based Video retrieval using multi-level Thepade's sorted ternary Block Truncation Coding with intermediate block videos and even-odd videos. *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 962-966.

Megrhi, S., Souidène, W., & Beghdadi, A. (2013). Spatio-temporal salient feature extraction for perceptual content based video retrieval. *2013 Colour and Visual Computing Symposium (CVCS)*, 1-7.

