Content-Based Video Retrieval Based on Integration of Wavelet Transform, Color and Texture Features

by

JAHANVI THAKKAR
202111058

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May,2023

Declaration

I hereby declare that

i) The thesis comprises my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) Due acknowledgment has been made in the text to all the reference material used.

JThakkar

Jahanvi Thakkar

Certificate

This is to certify that the thesis work entitled **Content-Based Video Retrieval Based on Integration of Wavelet Transform, Color, and Texture Features** has been carried out by **Jahanvi Thakkar** for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my supervision.

M Khare
28/7/23

Dr. Manish Khare
Thesis Supervisor

Acknowledgments

I would like to express my deepest gratitude to my supervisor, [Dr. Manish Khare], for his guidance, support, and encouragement throughout my research. Without his invaluable insights, expertise, and feedback, this thesis would not have been possible.

I would also like to thank the faculties and staff of the ICT Department at Dhirubhai Ambani Institute of Information and Communication Technology for their support and resources, as well as my colleagues and friends for their encouragement and assistance.

I am grateful to my study participants for their time and cooperation, without whom this research would not have been possible.

Finally, I thank my family for their unwavering support and encouragement throughout my academic journey.

Thank you all for your contributions to this thesis.

Contents

Abstract **v**

List of Tables **vi**

List of Figures **vii**

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Retrieval Systems 2

1.2.1 Text Retrieval Systems 2

1.2.2 Audio Retrieval Systems 3

1.2.3 Image Retrieval Systems 3

1.2.4 Video Retrieval Systems 4

1.3 Challenges 5

1.4 Contributions 6

1.5 Organization of Thesis 6

2 Related Work **7**

2.1 Image Retrieval 7

2.2 Video Retrieval 10

2.2.1 Colour Feature 10

2.2.2 Texture Feature 11

2.2.3 Wavelet Transform 14

2.3 Limitations of Existing Work 16

3 Dataset **17**

4 Proposed Method **20**

4.1 Training Process 21

4.2 Testing Process 23

4.3 Color Layout Descriptor 23

4.4 Edge Histogram Descriptor 24
4.5 Performance Measurements 25

5 Experiment and Results 27

6 Conclusion and Future Scope 37

References 38

Abstract

High-resolution, large-sized videos can now be transferred due to the fast development of information and communication technology, and video applications have developed in line with data quality levels. The applications of content-based video retrieval (CBVR) in various fields, like surveillance, education, sports, medicine etc., make it a crucial video application. The efficient Content-Based Video Retrieval (CBVR) algorithm described in this thesis is based on the MPEG-7 features of the Discrete Wavelet Transform (DWT), Dual-Tree Complex Wavelet Transform (DTCWT), Edge Histogram Descriptor (EHD), Linear Binary Pattern (LBP) and colour Layout Descriptor (CLD).

This thesis proposes a content-based video retrieval system that integrates wavelet transform, colour, and texture features for efficient and accurate video retrieval. The proposed system aims to address the limitations of traditional video retrieval systems that rely solely on low-level features, such as colour and texture, and do not consider the video's structural information. The system utilizes the wavelet transform to extract the structural information of the video. It combines it with colour and texture features to create a robust and accurate feature set for retrieval.

List of Tables

5.1 Comparisons of different Algorithms on Youtube Dataset[10] . . . 29
5.2 Performance comparison of the proposed method (CLD + EHD) . . 29
5.3 Performance comparison of the proposed method(CLD + LBP) . . . 30
5.4 Performance comparison of the proposed method (CLD + EHD) for UCF101[22] 34
5.5 Performance comparison of the proposed method (CLD + LBP) for UCF101[22] 34

List of Figures

2.1 Algorithm for CLD in Image Retrieval 8

2.2 Training process of Image Retrieval 9

2.3 Algorithm based on Query to image 11

2.4 ICTSLIC Method 12

2.5 The neighbourhood of surrounding pixels shown with different radii in (a), (b) and (c) 13

2.6 General algorithm for LBP 13

2.7 2D DWT with three levels of decomposition 14

3.1 UCF YouTube Action Dataset[10] 18

3.2 Various Categories in UCF101[22] Dataset 19

4.1 A generalized CBVR System 20

4.2 Training process of proposed Algorithm 22

5.1 Testing Video Clips of Youtube Dataset[10] 27

5.2 Retrieved Results for different algorithms tested on YouTube Dataset[10] 28

5.3 Retrieved Results for CLD + EHD algorithm tested on Youtube Dataset[10] 30

5.4 Retrieved Results for CLD + LBP algorithm tested on Youtube Dataset[10] 31

5.5 Testing video clips of UCF101[22] Dataset 32

5.6 Retrieved Results for CLD + EHD algorithm for UCF101[22] 33

5.7 Retrieved Results for CLD + LBP algorithm for UCF101[22] 35

CHAPTER 1
Introduction

This chapter introduces the research topic of content-based video retrieval (CBVR) and provides a comprehensive background on the challenges and limitations of traditional video retrieval systems. The research objectives are outlined in this chapter, emphasizing the aim of developing a robust CBVR system based on integrating wavelet transform, colour, and texture features. The chapter also discusses the potential applications and benefits of the proposed CBVR system in areas such as video surveillance, video search engines, and multimedia databases.

1.1 Background and Motivation

Due to the rapid growth of information and communication technology, large amounts of data can now be accessed more easily and quickly. From the time of COVID-19, daily YouTube uploads and views increased by roughly 700% and 210%, respectively, from the year 2020 to 2022, compared to the corresponding levels before the arrival of the pandemic. There are different sources of video content:

- Professionally created content like TV, News, Sports, Documentaries, etc.,
- User Created Content like youtube videos
- Security video Content like CCTV Footage

The number of Internet users will reach billions by 2025 due to increasing transmitted video traffic and quality standards. As a result, research into video applications seems to have been active. A video combines motion, audio, and text for more complex information than a single image. As a result, video application research requires an integrated strategy to consider the various sorts of information. One field of video application research is content-based video retrieval (CBVR).

The video retrieval system aims to retrieve videos similar to the user-defined query. By matching and extracting features, similarity may be evaluated. The algorithms for retrieving videos take queries in the form of either a sample image or a succession of sample photos (frames of video clips). It is common to use text-based algorithms for video retrieval. YouTube is a prime example of a text-based retrieval system. A text-based video retrieval system allows users to enter queries describing the video's characteristics. The algorithm looks up videos tagged with the user-described bag-of-words in the database. The key concern with text-based retrieval systems is a significant semantic gap when transferring rich user-defined semantics to low-level video data. This leads to several undesirable outcomes. There are three levels to describe the visual features of an image or video described below:

- **Low-level description** of the visual contents is based on colour, texture, and shape features. The low-level visual features of an image are directly related to the image content

- **Middle-level description** concerned with the background and spatial attributes of the concerned object

- **High-level description** representation deals with the human brain and perception. Examples are events, scenes, and human thinking, such as emotions.

1.2 Retrieval Systems

The history of retrieval systems spans various media types, including text, audio, images, and eventually video. Here is a brief overview of the evolution of retrieval systems across these media types:

1.2.1 Text Retrieval Systems

Text retrieval systems have a long history, dating back to the 1960s with the advent of early information retrieval systems. These systems were primarily based on keyword-based indexing and retrieval techniques. Documents were indexed based on their textual content, and users could search for specific terms or phrases to retrieve relevant documents. The Cranfield experiments, conducted in the 1960s and 1970s, played a crucial role in evaluating and advancing text retrieval techniques.

Early text retrieval systems employed techniques such as inverted indexes, which mapped each unique term in a document collection to the documents containing that term. Ranking algorithms, such as the Term Frequency-Inverse Document Frequency (TF-IDF), were used to measure the relevance of documents to a given query. These systems formed the foundation for modern search engines, which have evolved significantly with advancements in natural language processing, machine learning, and information retrieval algorithms.

1.2.2 Audio Retrieval Systems

The development of audio retrieval systems followed the increasing availability of digital audio recordings. In the 1990s, research efforts in audio information retrieval emerged, driven by the need to organize and search through large audio collections. These systems aimed to extract meaningful information from audio signals and enable users to search and retrieve specific audio segments or whole audio files.

Audio retrieval systems utilize various techniques, including speech recognition, audio signal processing, and acoustic feature extraction. Speech recognition algorithms convert spoken words into text, enabling keyword-based searching within audio recordings. Audio signal processing techniques analyze the spectral and temporal characteristics of audio signals to extract features that capture relevant information for retrieval. Acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms, are commonly used for audio retrieval.

1.2.3 Image Retrieval Systems

The rise of digital imaging technology led to the development of image retrieval systems. Content-based image retrieval (CBIR) systems emerged in the late 1990s and early 2000s, enabling users to search and retrieve images based on their visual content. These systems aimed to overcome the limitations of text-based image annotation and organization.

CBIR systems employ techniques such as feature extraction, similarity measurement, and indexing to facilitate image retrieval. Features such as colour histograms, texture descriptors (e.g., local binary patterns), and shape descriptors are extracted from images to represent their visual content. Similarity measurement methods, such as Euclidean distance or cosine similarity, are used to compare the features of query images with those in the image database. Indexing structures, such as inverted files or binary trees, efficiently store and retrieve images based on their visual features.

1.2.4 Video Retrieval Systems

Video retrieval systems have evolved by building upon the image and audio retrieval advancements. These systems aim to enable users to search and retrieve relevant videos based on their content, including visual and auditory aspects. Video retrieval presents unique challenges due to the temporal nature of videos, where sequences of frames and audio need to be analyzed and compared.

Video retrieval systems employ keyframe extraction, shot detection, video segmentation, and feature extraction techniques. Keyframes, representative video frames, are extracted to represent the video's content efficiently. Shot detection algorithms identify boundaries between different shots or scenes within a video. For more granular retrieval, video segmentation techniques divide videos into meaningful segments, such as scenes or objects. Features from both visual and auditory domains, such as colour histograms, texture descriptors, audio spectrograms, and speech recognition, are extracted to capture the content of the video.

These features are then used to measure similarity or relevance between videos, enabling users to search for specific video content or retrieve videos similar to a given query. Indexing structures, such as spatiotemporal indexes or video fingerprints, efficiently store and retrieve videos based on their content features.

Advancements in machine learning, deep learning, and multimedia processing techniques have further improved video retrieval systems. Cross-modal retrieval techniques aim to bridge the gap between different modalities, such as text, audio, images, and video, enabling comprehensive retrieval across multiple media types. Additionally, user interaction and relevance feedback mechanisms allow users to provide feedback on retrieved videos, improving the system's performance and adaptability to individual preferences.

Overall, the evolution of retrieval systems from text to audio, image, and video has been driven by advancements in data storage, processing power, and algorithmic techniques. These systems continue to evolve, exploring new research areas to enhance the retrieval experience and cater to the growing demands of multimedia content organization and access.

1.3 Challenges

Content-based video retrieval (CBVR) faces several challenges that need to be addressed to develop effective retrieval systems. Some of these challenges include:

- **Semantic Gap:** The semantic gap refers to the disparity between low-level visual features extracted from videos and the high-level semantic concepts that humans use to interpret and understand video content. Bridging this gap is a significant challenge in CBVR, as it requires developing techniques that can accurately capture and represent the semantic meaning of videos based on their visual features.

- **Scalability:** As the volume of video data grows exponentially, CBVR systems must be scalable to handle large-scale video collections. Efficient indexing and retrieval algorithms are required to ensure fast and accurate search results, even when dealing with vast amounts of video content.

- **Computational Complexity:** Extracting and processing visual features from video frames can be computationally demanding. CBVR systems must address the challenge of managing the computational complexity associated with feature extraction, indexing, and retrieval processes to ensure real-time or near-real-time performance.

- **Robustness to Variations:** Videos exhibit variations in lighting conditions, camera viewpoints, object appearances, and other factors. CBVR systems should be robust to these variations to ensure accurate and consistent retrieval results. Complete feature extraction and matching techniques are necessary to handle variations in video content and maintain retrieval performance across diverse conditions.

- **Multimodal Integration:** Videos often contain multiple modalities, such as audio, text, and visual information. Integrating these modalities effectively is a challenge in CBVR. Developing techniques that can effectively fuse and leverage different modalities to improve retrieval accuracy and comprehensiveness is crucial.

1.4 Contributions

The thesis proposes a novel fusion approach that combines colour and texture features in the context of content-based video retrieval. By integrating these complementary features, a more comprehensive representation of video content is achieved, leading to improved retrieval performance and more accurate results. We have conducted a thorough evaluation and analysis of different feature fusion methods and their impact on content-based video retrieval. It compares the performance of various fusion techniques, such as CLD and EHD fusion or CLD and LBP fusion, using quantitative measures like precision, recall, accuracy, and mean average precision (MAP). We used Youtube Action Dataset [10], and UCF101 [22] dataset, a widely recognized benchmark dataset in the field of video analysis and retrieval. The thesis presents comprehensive performance evaluations of the proposed fusion methods using appropriate metrics. It visually illustrates the retrieval results through figures and tables, providing a clear demonstration of the efficacy and superiority of the proposed fusion techniques. The performance analysis and results contribute to the overall understanding of content-based video retrieval and offer practical insights for future research and system development.

1.5 Organization of Thesis

The organization of the thesis is as follows. Chapter 2 deals with the related work in the field of CBVR; It covers related work done using primitive features such as colour, texture, and wavelet transform. A brief mention of available CBVR methods is described. It also includes limitations of existing work. Chapter 3 deals describing data sets used in this thesis to run our proposed method. Chapter 4 deals with the proposed method, which includes the Training and testing process followed by Performance Measurements, used to evaluate our proposed system. Chapter 5 provides an Experiment of the proposed method and shows some interesting simulation results and a comparison with existing methods. Chapter 6 concludes the thesis and discusses the future scope of this thesis.

CHAPTER 2
Related Work

This chapter will discuss related work carried out regarding CBVR System. Past research in Content-Based Video Retrieval Systems has been carried out on the feature extraction part. The primitive features used in the past include colour, texture, shape, edge, structure etc. First, we will go through research on Image Retrieval and then in the field of Video Retrieval concerning colour and texture features.

A retrieval system is a software-based system that organises, stores and retrieves information based on user queries or search criteria. It is commonly used to search and retrieve relevant documents, data, or media files(Images or Videos) from an extensive collection of resources.

2.1 Image Retrieval

Image retrieval systems were created due to the advancement of digital imaging technology. Content-based image retrieval (CBIR) systems grew popular in the late 1990s and early 2000s. Users of CBIR systems might search for and retrieve photos based on their visual content by using techniques including feature extraction (for example, colour, texture, and shape), similarity measurement, and indexing. These technologies enabled performing operations like looking for comparable photos or fetching images using visual searches.

We classify the different types of visual search based on the information in the database and the challenge of indexing and querying visual data. The image-to-image (I2I) problem, where a picture is queried against a database of images, is the topic of the majority of research in visual search. Agarwal et al. [1] suggested a method based on wavelet transform and colour layout descriptor for image retrieval. Fig. 2.1 depicts the overall feature extraction procedure of this approach.

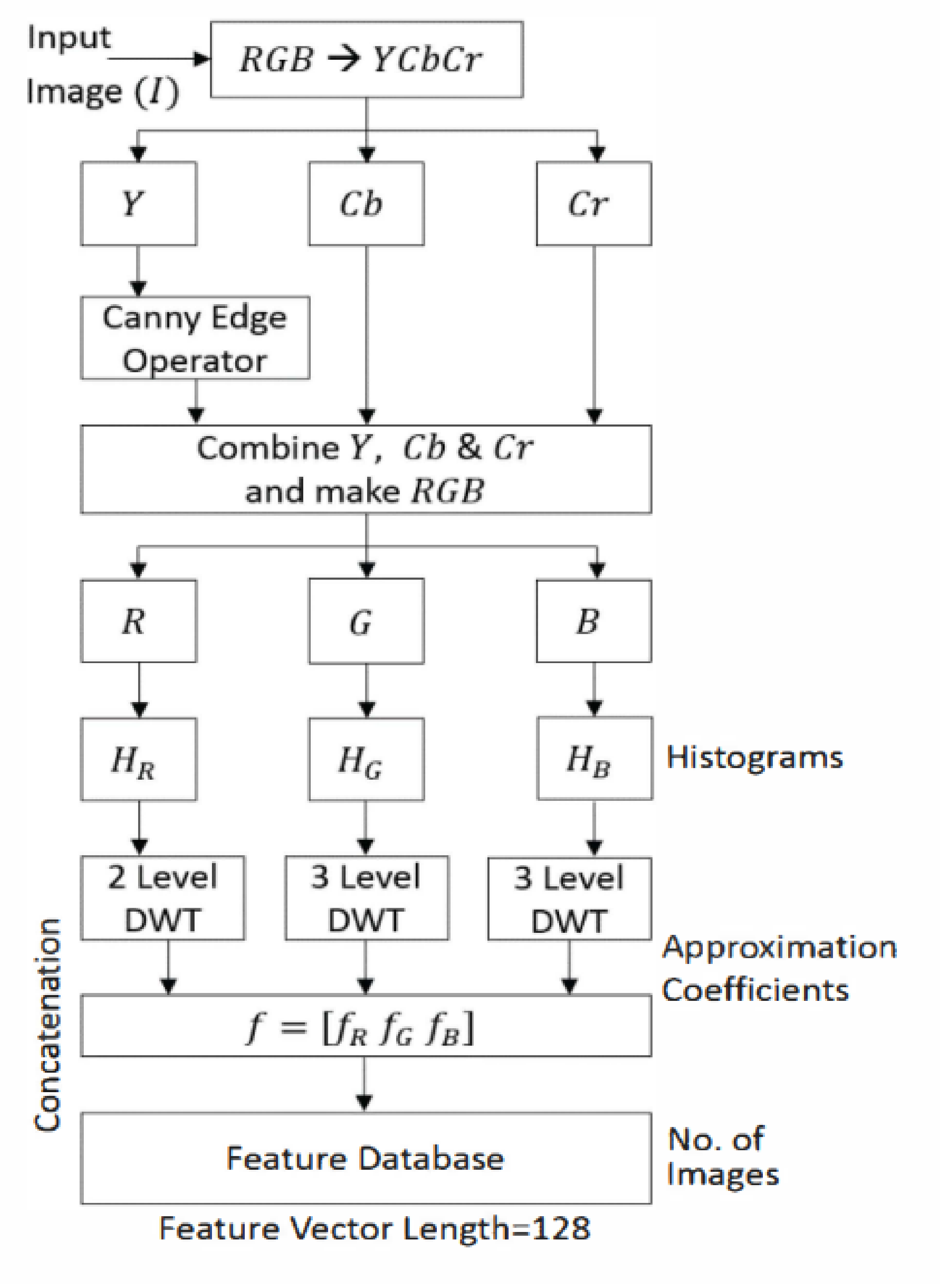


Figure 2.1: Algorithm for CLD in Image Retrieval

Agarwal et al. [2] proposed that The Edge Histogram Descriptor (EHD) and Discrete Wavelet Transform (DWT) are combined in the suggested approach to improve the reliability and accuracy of image retrieval. The technique offers a more thorough representation of image content by combining EHD, which captures edge information, and DWT, which captures structure features. This integration improves retrieval speed across different picture databases and applications by considering local edge features and global structure information. In Fig. 2.2, the whole training procedure is depicted.

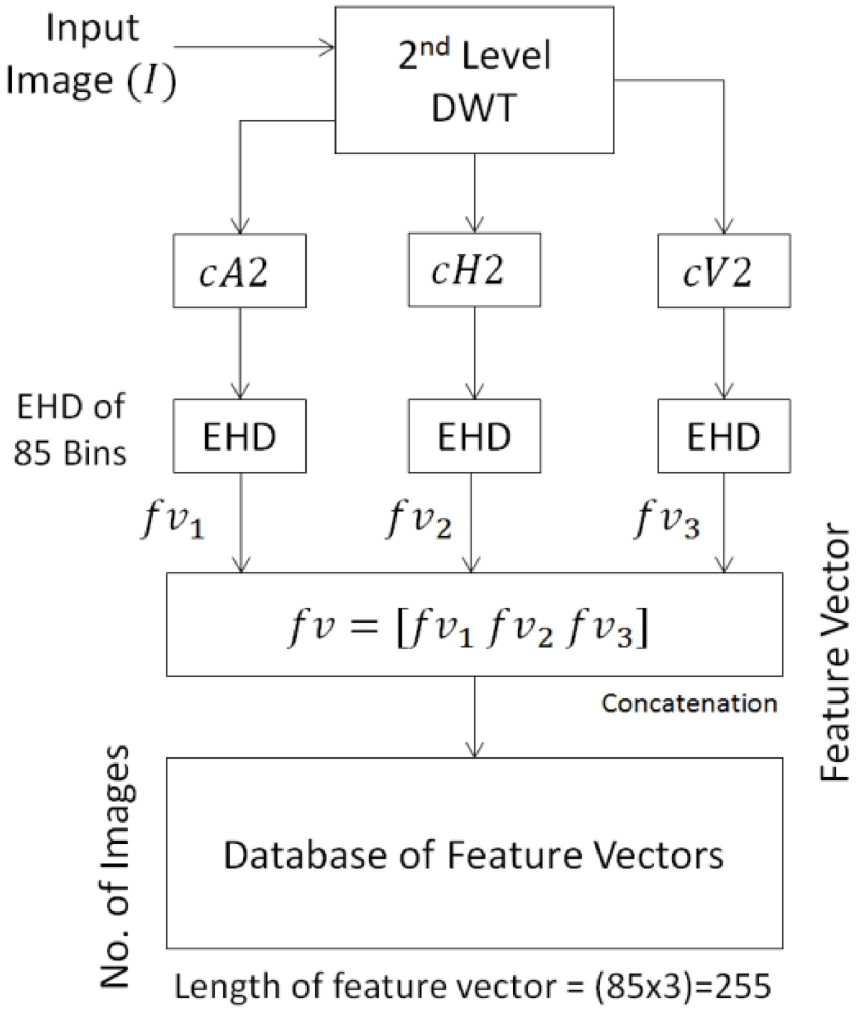


Figure 2.2: Training process of Image Retrieval

Verma et al. [20] proposed that the wavelet-based colour feature extraction technique outperforms existing Mpeg-7 colour descriptors. The edge histogram created from the wavelet coefficients of the image at various resolution levels is used to extract texture information. Compared to existing MPEG-7 descriptors, the multi-resolution technique helps capture texture details from finer to coarser levels.

2.2 Video Retrieval

The detailed reviews offer thorough insights into the methods, difficulties, evaluation procedures, and new developments in video retrieval systems. Newton et al. [23] reviewed several proposed methodologies from different kinds of literature and found different eligibility criteria, such as dimensionality reductions, feature extractions, machine learning approaches, or segmentation approaches, etc. Bhaumik et al. [5] aimed to study and analyse the appropriate hybrid soft computing methods in the content-based picture and video retrieval systems. They understood the need to increase retrieval efficiency and accuracy by combining the advantages of several soft computing techniques. To improve the precision and effectiveness of video retrieval, Padmakala et al. [19] introduced a content-based video retrieval system that successfully blends texture, colour, and ideal keyframe properties. Results of the retrieval process are more accurate and swifter when numerous features are combined and the best keyframes are used. The suggested method shows potential in several applications, including video surveillance, video recommendation systems, and multimedia databases.

2.2.1 Colour Feature

Colour is one of the most used visual features in multimedia retrieval systems. It is robust to background complications and independent of the size of the video frame and orientation. Many colour representation techniques have been introduced. Daga et al. [12] proposed an integrated approach for content-based video retrieval using colour features that enhance accuracy and efficiency. Combining multiple colour features allows the system to capture a more comprehensive representation of video content and provide more accurate retrieval results. The integration approach holds promise for various applications, including multimedia databases, video recommendation systems, and video surveillance.'

Potluri et al. [21] proposed an approach for content-based video retrieval using dominant colour and shape features to enhance the accuracy and efficiency of video retrieval. By incorporating both dominant colour and shape information, the system can capture distinct visual characteristics of the video content and provide more accurate retrieval results. Mallick and Mukopadhyay [16] proposed a video retrieval framework based on the colour co-occurrence feature of adaptive low-rank extracted keyframes, and graph pattern matching enhances the accuracy and efficiency of video retrieval. By leveraging the relationships between colour co-occurrence patterns and keyframe representations, the framework provides a more comprehensive picture of the video content and achieves more accurate retrieval results.

2.2.2 Texture Feature

Texture is another essential property of multimedia. Unlike colour, texture occurs over a region rather than a point. It can be defined as the intrinsic property of all surfaces that describe visual patterns and contains essential information about the structural arrangement of the surface. Texture has been proven to be very effective in content-based video retrieval systems. Another work in multimedia is carried out with the concern of video-to-image(V2I), where the query is a video and the images as a database. Another variant is video-to-video(V2V), where both query and datasets are videos. Early work in the I2V problem simply applied I2I techniques for video search. In this case, the video database is treated merely as an image database of video frames. Araujo and Girod [4] proposed an algorithm based on Query as an image and video as an output through the fisher vector and bloom filters technique, as shown in Fig. 2.3. And after feature extraction, the system will match feature vectors with the database to retrieve similar outputs using Euclidean Distances.

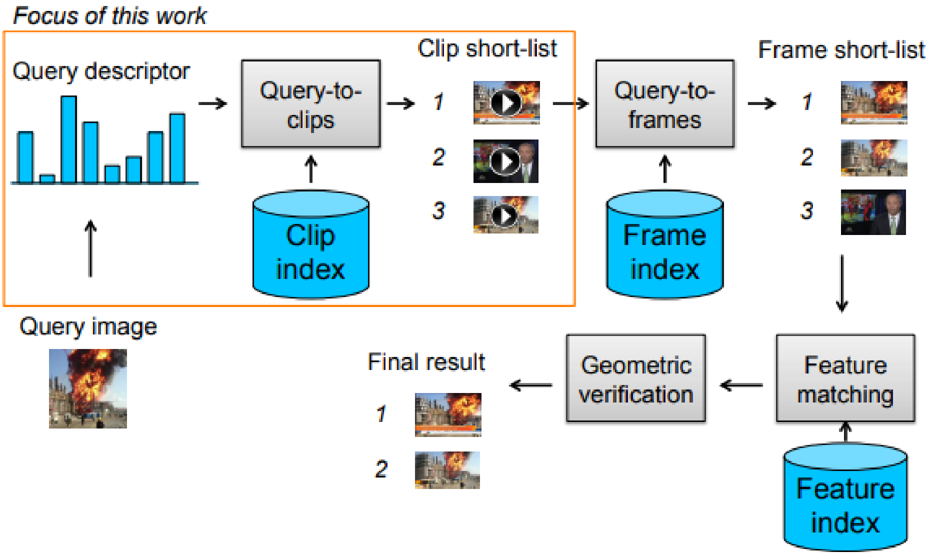


Figure 2.3: Algorithm based on Query to image

Mounika et al. [6] proposed the Integration of Curvelet Transform and Simple Linear Iterative Clustering (ICTSLIC) method using superpixels as features to retrieve similar videos, as shown in Fig. 2.4.

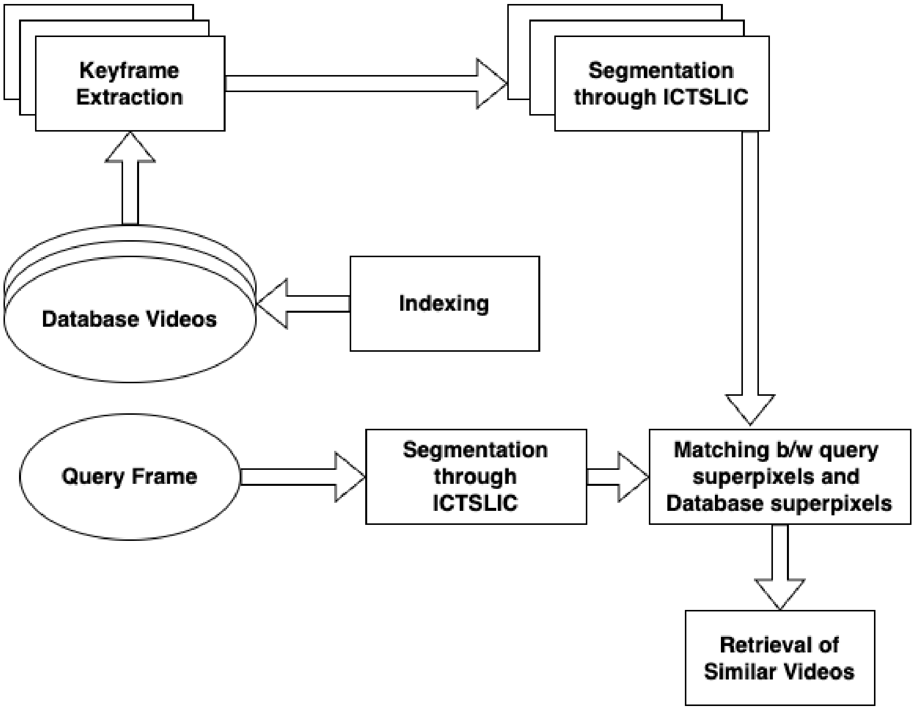

Keyframe Extraction
Segmentation through ICTSLIC
Database Videos
Indexing
Query Frame
Segmentation through ICTSLIC
Matching b/w query superpixels and Database superpixels
Retrieval of Similar Videos

Figure 2.4: ICTSLIC Method

The Local Binary Pattern (LBP) texturing operator labels each pixel in an image by thresholding its immediate surroundings and treating the result as a binary number. The LBP texture operator has gained popularity as a strategy in many applications due to its discriminative power and computational simplicity. The robustness of the LBP operator to monotonic grey-scale changes LBP is a spatiotemporal local texture descriptor as shown in Fig. 2.5. where g_c is the grey level of the centre pixel of the local neighbourhood and g_p is the grey level of the pixels corresponding to P equally spaced pixels on a square of length R. Histograms are used to extract feature descriptor vectors. The features that were taken from the three orthogonal planes must be combined. In order to account for pixels that are not exactly in the middle, interpolation is used. Each pixel in a picture is linked to a specific binary pattern, which is based on a specific arrangement or orientation of disparities between the grey level value of the centre pixel and its circular neighbours with radius R. Thus, the LBP codes are computed as

$$LBP_{P,R} = \sum_{p=0}^{p-1} s\left(g_p - g_c\right) 2^P \tag{2.1}$$

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & otherwise \end{cases} \tag{2.2}$$

The sign function s(x) as in Eq. 2.1 and 2.2 specifies that the LBP code is invariant against any monotonic transformation of the brightness of an image.

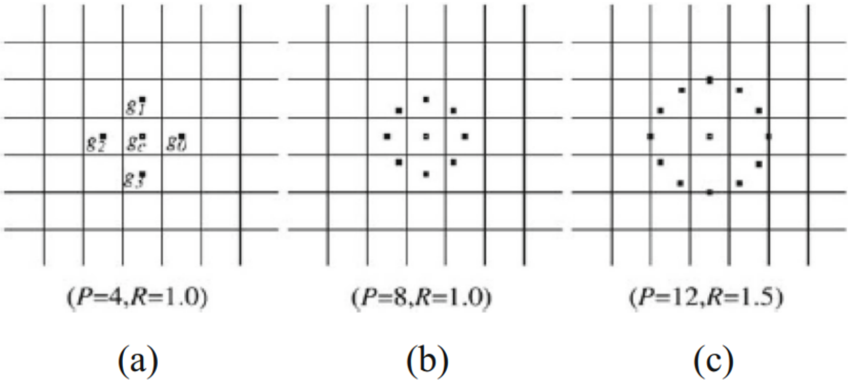


Figure 2.5: The neighbourhood of surrounding pixels shown with different radii in (a), (b) and (c)

Mounika et al. [18] proposed a method that focuses on Linear Binary patterns (LBP), capable of jointly describing motion and appearance features and robust to the problems mentioned earlier. The general algorithm for this algorithm is shown in Fig. 2.6. Mounika et al. [7] proposed a method that uses Laplacian of Gaussian(LOG) and Histogram of Oriented Gaussian(HOG) descriptors for CBVR.

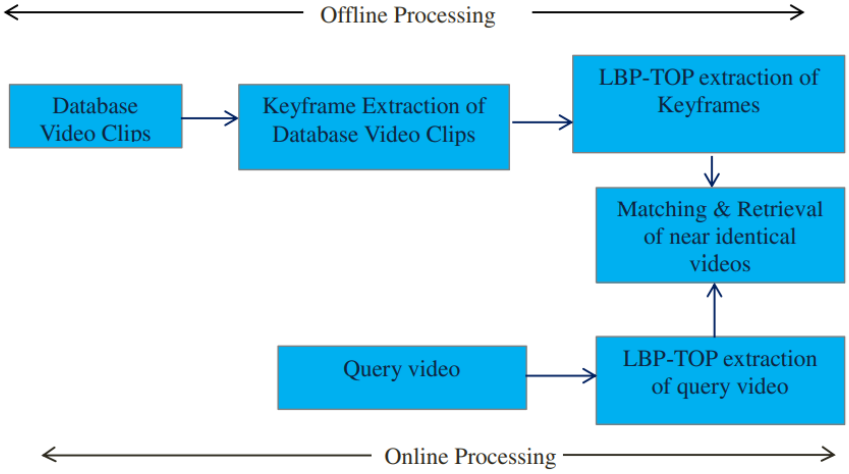

Offline Processing
Database Video Clips
Keyframe Extraction of Database Video Clips
LBP-TOP extraction of Keyframes
Matching & Retrieval of near identical videos
Query video
LBP-TOP extraction of query video
Online Processing

Figure 2.6: General algorithm for LBP

Janarthanan et al. [14] used the edge histogram descriptor (EHD) to capture the spatial distribution of edges for the video search and LBP as texture features based on clustering. Mounika and Khare [17] propose a method of frame fusion with HOG and Euclidean distance as matching retrieval.

2.2.3 Wavelet Transform

The wavelets are sampled at distinct intervals for the DWT wavelet transform. DWT offers simultaneous information about the image's spatial and frequency domains. A combination of the analysis filter bank and decimation operation can be used in DWT operation to analyse an image. Each decomposition level's low and high pass filters are included in the analysis filter bank. The high pass filter collects features like edges, whereas the low pass filter extracts the image's approximation information. Two distinct 1D transforms are used to create the 2D transform. The detail coefficients in 1D DWT carry high-frequency information, whereas the approximation coefficients contain low-frequency information. When 2D DWT is applied, the input image is divided into four distinct subbands: low-frequency components in both the horizontal and vertical directions (cA), low-frequency components in both the horizontal and vertical directions (cV), high-frequency components in both the horizontal and low-frequency directions (cH), and high-frequency components in both the horizontal and vertical directions (cD). Additionally, cA, cV, cH, and cD can be depicted as LL, LH, HL, and HH, respectively. The image after 3-level decomposition is shown in Fig. 2.7.

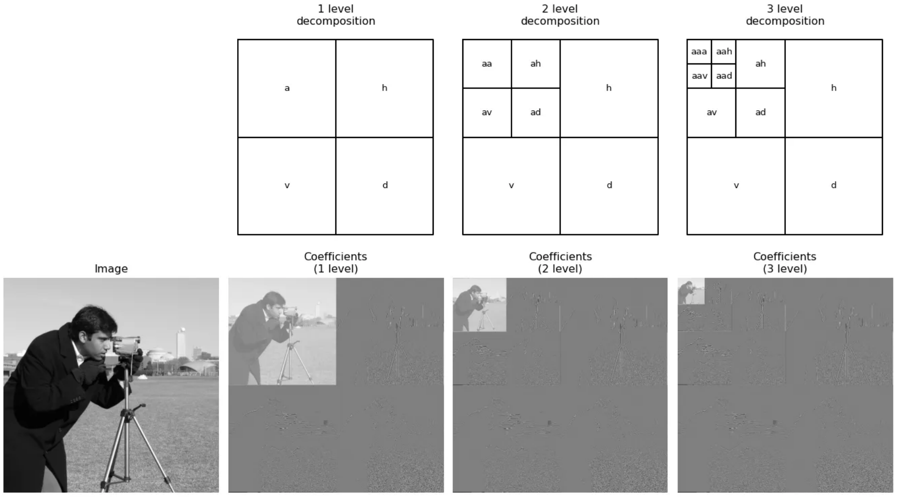


Figure 2.7: 2D DWT with three levels of decomposition

Wavelet transform plays a crucial role in Content-Based Video Retrieval (CBVR) by decomposing video frames into different frequency sub-bands, capturing local and global spatial information. It enables the extraction of multi-resolution features that enhance video representation of texture, edges, and motion. Wavelet transform offers robustness to scale and translation variations, contributing to improved retrieval accuracy and efficiency in CBVR systems. Yadav et al. [24] proposed a technique based on hybrid wavelet transforms with colour spaces. Chivadshetti et al. [8] Proposed a Content-Based Video Retrieval system that emphasizes upon energy contents of the videos. A novel method of generation of orthogonal transform is induced called Self Mutation of Hybrid Wavelet Transform (SMHWT).

To get around the drawbacks of the conventional wavelet transform, Kingsbury [10] proposed using dual-tree complex wavelet transform. The basic DWT is extended with complex values to provide the complex wavelet transform (CWT). In the transform domain, CWT employs complex value filtering to divide the real/complex signal into real and imaginary components. Information about amplitude and phase is computed using real and imaginary coefficients. Positive and negative orientations are separated into separate subbands in DTCWT. Using two distinct, discrete wavelet transform (DWT) decompositions, DTCWT calculates the complex transform of a signal. Two parallel trees are produced by the DWT decomposition. To employ 1D DTCWT and take into account $(h_x + jg_x)$, where hx is the set of filters h0, h1 and gx is the set of filters g0, g1 both sets being in only the x-direction (1-D). The real-valued lowpass and highpass filters for real trees are designated as h0 and h1, respectively. For the hypothetical tree, the same holds true for g0 and g1. DTCWT has a higher degree of directional selectivity than other techniques. Additionally, [11] it has a reduced shift variant characteristic.
The following are features of DTCWT:

- Approximate shift variant
- Good directional selectivity
- Perfect reconstruction using short linear filters
- Limited redundancy
- Efficient order n computations

Jameel et al. [13] proposed a content-based video retrieval framework utilizing Dual-Tree Complex Wavelet Transform to improve the accuracy and efficiency of video retrieval. By leveraging the multi-resolution analysis provided by DT-CWT, the framework captures spatial and temporal information, resulting in more comprehensive representation and more accurate retrieval results. Yuk et al. [9] done performance analysis of using wavelet transform in a content-based video retrieval system reveals its positive impact on video retrieval's accuracy, efficiency, and robustness. The findings emphasize the potential of wavelet transform in improving the overall performance of video retrieval systems. The results of this analysis contribute to the understanding and advancement of content-based video retrieval techniques, aiding in the development of more effective and efficient retrieval systems.

2.3 Limitations of Existing Work

We have found that colour Layout Descriptor as a colour feature, Linear Binary Pattern, and Edge Histogram Descriptor both as a texture feature, which is low-level and middle-level descriptors, respectively, are used for image retrieval, not for video retrieval. So, we are using the fusion of these feature descriptors along with Discrete Wavelet Transform (Multi-level Resolution also included) and Dual-Tree Complex Wavelet Transform to increase the accuracy of retrieved videos as output.

The fusion of colour and texture features in CBVR methods offers a more comprehensive and discriminative representation of video content, enhances variations' robustness, and improves retrieved videos' relevance. By leveraging the complementary nature of these features, the fusion approach aims to enhance the overall performance of the CBVR system.

CHAPTER 3
Dataset

This chapter will discuss the dataset used to perform our proposed method. The Stanford I2V dataset[3], presented by Andre Araujo, is a sizable dataset relevant to studying the issue of querying a video database using pictures. More than 200 ground-truth questions have been annotated, and more than 3,800 hours of newscast films have been collected. The largest dataset of human behaviours, UCF101[22], is presented by Khurram Soomro. Over 13k movies, 101 action lessons, and 27 hours of video content are included. Realistic user-uploaded films with camera movements and crowded backdrops are included in the database. The baseline action recognition scores they present on this new dataset using the conventional bag of words method have an overall performance of 44.5%.

To the best of my knowledge, the UCF101[22] dataset of actions is the most difficult one at the moment because of the sheer volume of classes, clips, and unconstrained clips it contains. With 51 action categories and 7,000 personally annotated clips pulled from various sources, including digitized movies and YouTube, H. Kuehne amassed the most extensive action video library to date. This database is named HMDB[15]. ActivityNet[11], a brand-new large-scale video benchmark for comprehending human activity, is introduced by Fabian Caba Heilbron. The benchmark seeks to encompass a wide variety of intricate human activities that individuals are interested in on a daily basis. ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours.

UCF YouTube Action Data Set[10] is a widely used dataset in content-based video retrieval. It comprises a sizable selection of YouTube films depicting different human behaviours. The dataset offers various action categories, making it appropriate for testing and comparing video analysis systems. It has 11 different action genres, including basketball shooting, cycling, diving, golf swinging, horseback riding, juggling soccer, trampoline jumping, tennis swinging, volleyball spiking, and walking a dog. Due to the wide variations in camera motion, item look and posture, object scale, viewpoint, cluttered background, illumination conditions, etc., this data set is quite difficult to work with. The videos are divided into 25 groups, with at least four action clips in each group for each category. The videos in the same collection all have certain things in common, including the same actor, similar backgrounds, similar viewpoints, and so on.

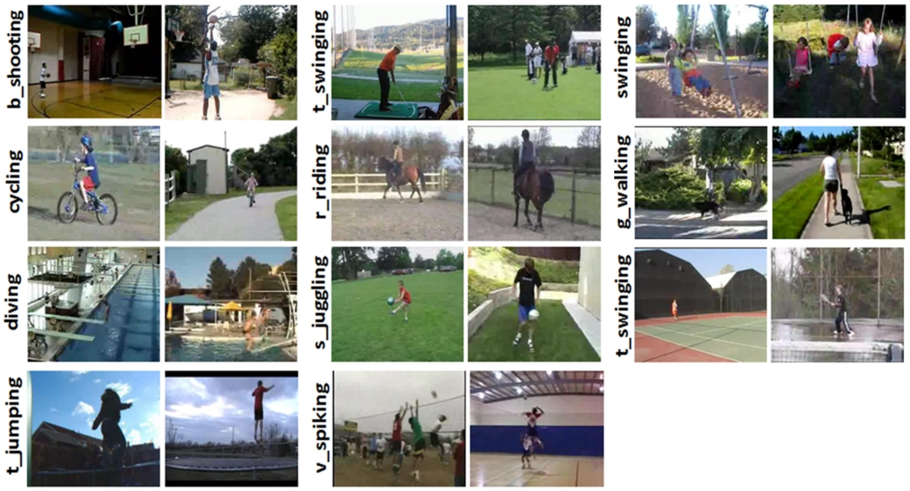


Figure 3.1: UCF YouTube Action Dataset[10]

Another dataset that was employed was UCF101[22], which includes 101 different action categories, from commonplace ones like "brushing teeth" and "playing the guitar" to sports-related ones like "basketball" and "soccer juggling." Each action category includes several videos showing various examples of the designated activity. The dataset contains 13320 video clips, offering a sizable amount of data for developing, testing, and assessing video retrieval algorithms. 101 different action categories are included in the dataset. The action categories can be divided into five types:

1. Human-Object Interaction
2. Body-Motion Only
3. Human-Human Interaction
4. Playing Musical Instruments
5. Sports

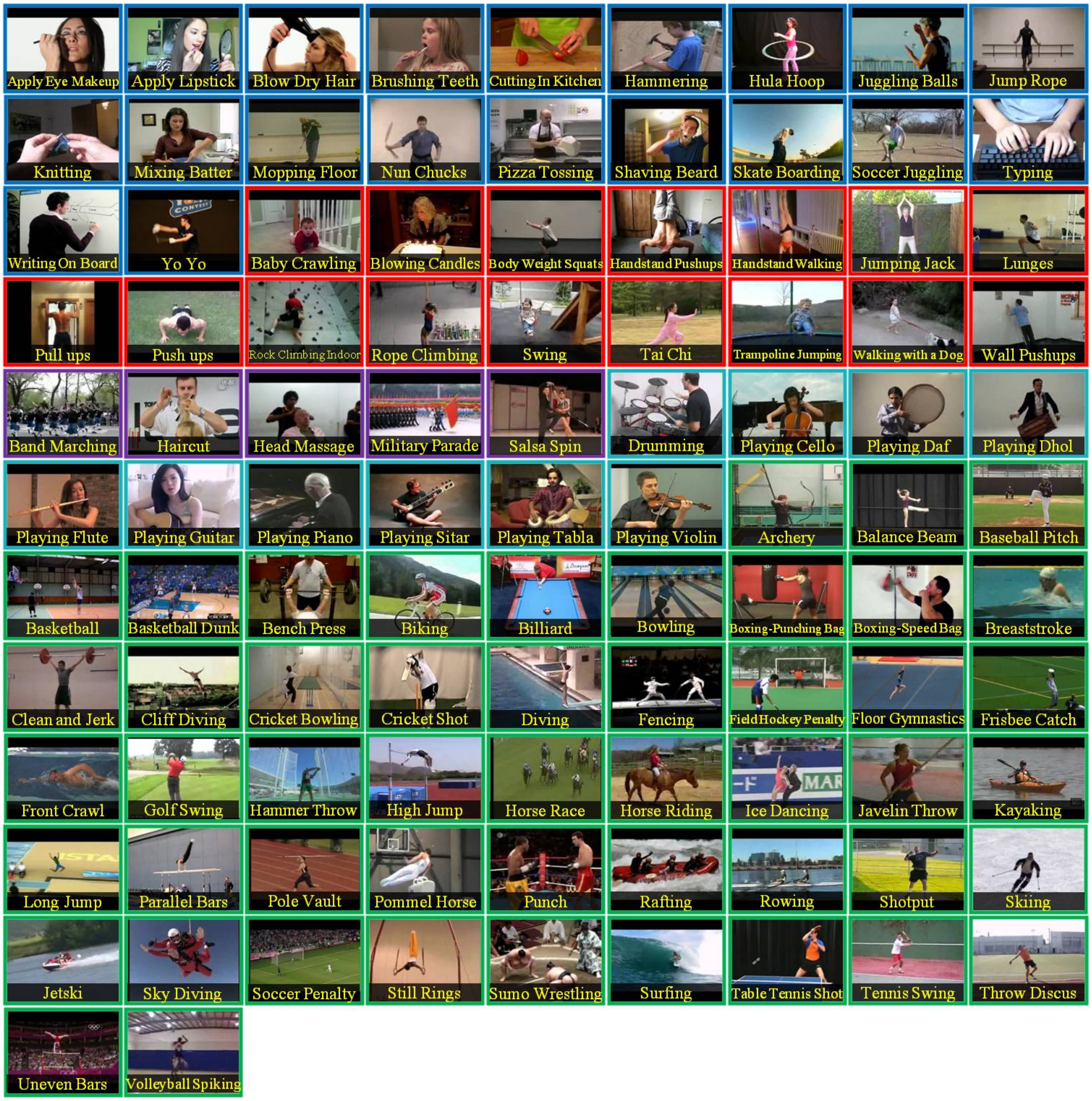


Figure 3.2: Various Categories in UCF101[22] Dataset

CHAPTER 4

Proposed Method

This chapter contains a detailed discussion of the proposed method for Content-Based Video Retrieval. The CBVR system collects the visual characteristics (colour, shape, texture, and spatial information) from each video frame in the database and stores them in a separate feature vector database. As seen in Fig. 4.1, the generalized CBVR system. The user provides the system with a video clip for a query. The system will decompose these video clips into frames, extract the features from those frames as it does for each database video frame, and then search the database for frames with feature vectors that match the query image. It then sorts the most similar frames according to how closely they reflect each other. So, it mainly involves two processes, first, the **feature extraction process**, which is the training process, and the second is the **feature matching process**, which is the testing process.

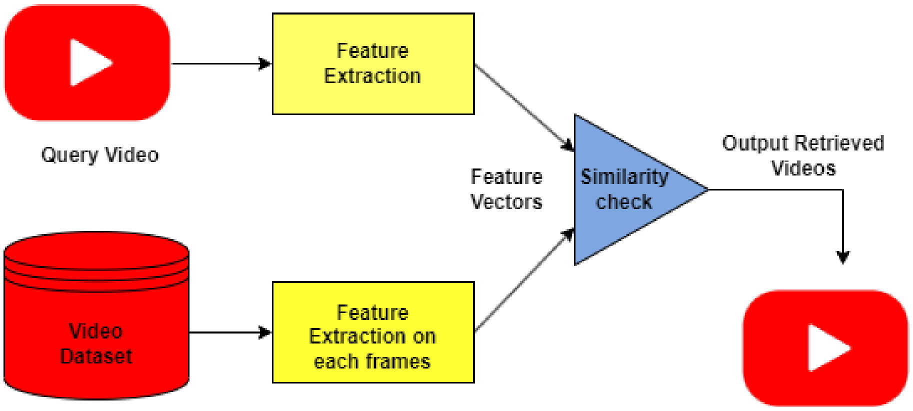


Figure 4.1: A generalized CBVR System

4.1 Training Process

For the training process, we have tried different combinations of features(CLD, EHD, LBP) along with two types of wavelet transforms (DWT, DTCWT), which are mentioned below:

1. CLD + DWT
2. CLD + DTCWT
3. EHD + DWT
4. EHD (Multilevel)
5. LBP + DWT
6. CLD + DTCWT + EHD
7. CLD + DWT + EHD
8. CLD + DWT + LBP

By using only CLD, it can capture only colour distribution information and may not effectively capture the structural or textural characteristics of the videos. This can result in difficulties in distinguishing visually similar videos with different content, leading to lower retrieval accuracy. Similarly, EHD will give only structural information, while LBP will give only Textural information about the video. Different frames of the video contain varying complex and irregular textures as well as colour features. For extracting texture features, LBP or EHD can be used, owing to its ability to acquire texture features effectively and for colour information, we can use CLD. Hence a combination of multiresolution techniques such as Discrete wavelet transform, which have three directional (horizontal, vertical, and diagonal) coefficients, with LBP, EHD or CLD, produces a better retrieval result.

The Training process converts videos of the dataset into frames and then extracts the image features to a different extent and prepares a database of feature vectors. These feature vectors are obtained using wavelet transforms (DWT or DTCWT), colour layout descriptors, Linear Binary Patterns, and edge histogram descriptors on selected wavelet coefficients (Approximation Coefficients). We have two proposed algorithms. One uses CLD as a colour descriptor, EHD as a texture descriptor, and another uses CLD as a colour descriptor while LBP is a texture descriptor. The overall training process is shown in Fig 4.2.

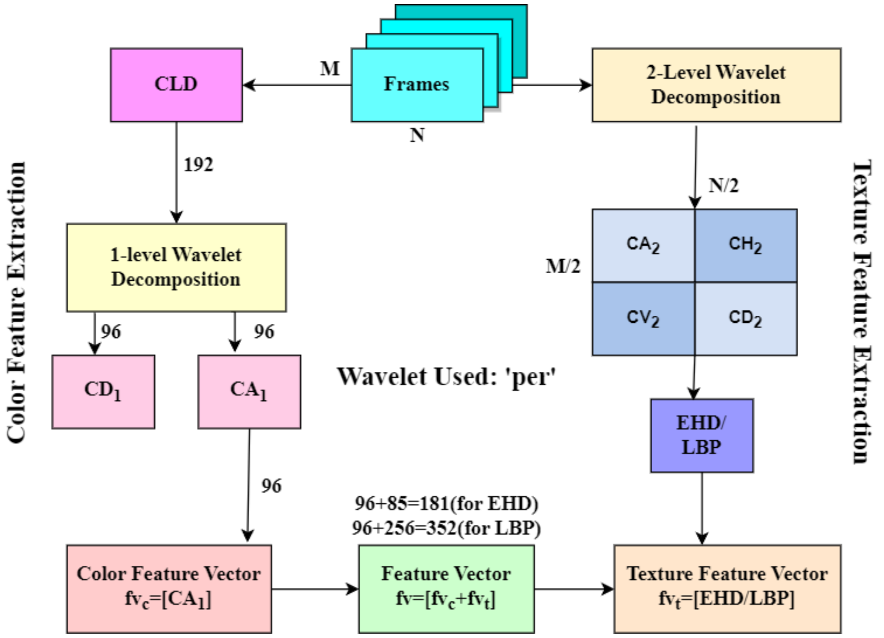

Color Feature Extraction
CLD
M
Frames
N
2-Level Wavelet Decomposition
Texture Feature Extraction
192
N/2
1-level Wavelet Decomposition
M/2
CA_2
CH_2
CV_2
CD_2
96
96
CD_1
CA_1
Wavelet Used: 'per'
EHD/ LBP
96
96+85=181(for EHD)
96+256=352(for LBP)
Color Feature Vector fv_c=[CA_1]
Feature Vector fv=[fv_c+fv_t]
Texture Feature Vector fv_t=[EHD/LBP]

Figure 4.2: Training process of proposed Algorithm

The main steps of the training algorithm are as follows:

Step 1: Take the video dataset and convert it into its respective frames.

Step 2: Due to EHD requirements, resize each frame so it is divisible by 4.

Step 3: Take 2 Level Discrete Wavelet Transform of the input frame.

Step 4: The two-level DWT gives the four matrices cA2, cH2, cV2, and cD2 of wavelet coefficients at level 2.

Step 5: Calculate the edge histogram using the approximation and detailed coefficients, but ignore the detailed coefficients because they primarily include noisy details rather than valuable information. The EHD provides 85 information points, 80 of which come from standard bins and the remaining five from the global bin. As a result, for a single wavelet coefficient matrix, the texture feature vector (fvt) has a length of 85.

OR

Step 5: Apply 2-level wavelet decomposition on each frame so that it will generate a matrix that includes cA, cH, cV, and cD coefficients. Then it will calculate the LBP of this matrix with a distance of 1. As a result, a texture feature vector (fvt) is generated, which has a length of 256.

Step 6: Take a frame now and apply CLD. To create two matrices, we must first convert each frame from an RGB image to a grayscale image using the 1-level wavelet transform approach (Approximation). Consider the approximation coefficients to be 96-length colour feature vectors.

Step 7: Calculate the feature vector (fv = fvc + fvt (96+85=181) or (96+256=352) for each frame in the database and arrange all these feature vectors in a database.

In the end, this training process will generate a feature vector database, which stores all frames with its feature vectors of length 181(for EHD as a texture feature) or 352(for LBP as a texture feature).

4.2 Testing Process

The testing procedure is necessary to confirm the effectiveness of the suggested algorithm. To produce visually comparable frames, this technique requires comparing the feature vector retrieved from the query video clip with feature vectors stored in the database. The following are the steps in this testing process:

Step 1: Input the video clip and convert it into frames.

Step 2: Apply the same method used in training to determine its feature vector, as shown in Fig. 4.2.

Step 3: Make a similarity comparison between the feature vector(fv) of the requested video frame and each feature vector found in the database. By sorting the feature vector database according to the query feature vector(x), the similarity is determined. This distance can be found as given in equation 4.1. Here *i* are the number of trained frames

$$distarray(i) = sum(abs(x(i,:) - fv)) \tag{4.1}$$

And 181 or 352 is the length of the feature vector. The distance of a frame from itself is zero. The distances are then stored in increasing order, and the first n-matches are retrieved as output video frames. Here we are taking the first nine matches to represent our results.

4.3 Color Layout Descriptor

The steps for computing CLD are:

1. The input frame is divided into 8 x 8 blocks to achieve scale invariance in Color Layout Descriptor (CLD).

2. A representative color is computed for each block, often using the average color of the block. The image is also reduced to 64 x 64, resulting in each block being 8 x 8 in size.

3. The image, now consisting of average color blocks, is converted from RGB space to YCbCr space.

4. Each color channel (Y, Cb, Cr) undergoes an 8 x 8 Discrete Cosine Transform (DCT), transforming the blocks into DCT coefficients. Each block contains 64 DCT coefficients.

5. Zigzag scanning is performed to select and quantize a few low-frequency coefficients. 6 coefficients are selected from Y-DCT, while only 3 coefficients are selected and quantized from both Cb and Cr.

6. The Color Layout Descriptor (CLD) feature vector is formed by taking only the DC coefficient from each block. This reduces the number of coefficients to 1 for Y, 1 for Cb, and 1 for Cr. Therefore, the total number of coefficients for one block is 3, and for all 64 blocks, the length of the CLD feature is 192.

4.4 Edge Histogram Descriptor

The steps to compute EHD are:

1. The input image is divided into 4 x 4 non-overlapping blocks. If the image dimensions are not divisible by 4, the image is resized accordingly. Each extracted block is further divided into 2 x 2 blocks for capturing local edge orientation.

2. For each block, a five-point bin is initialized to capture vertical, horizontal, diagonal (45°), diagonal (135°), and non-edge orientations. This results in a total of 16 bins for all 16 blocks, creating an Edge Histogram Descriptor (EHD) vector of length 80.

3. The local edge orientations from each 2 x 2 sub-block are captured using 2 x 2 operators, resulting in five values: EOv, EOh, EOd45, EOd135, and EOnoe.

4. The maximum of these five values is compared to a threshold (T) to determine the dominant edge orientation. The count of the corresponding bin point is increased by one. This process is repeated for all 2 x 2 sub-blocks within one image block.

5. For each image block, a complete bin (Bin[1] = [bo, b, b, b, b]) is obtained by performing the above operations. This is repeated for all 16 image blocks, resulting in 16 bins.

6. The 16 bins are arranged in a matrix called AllBins. The global bin is obtained by calculating the mean (column-wise) of the AllBins matrix. The EHD vector is formed by combining the global bin with all the calculated bins. The length of the EHD vector becomes 85 and is represented as EHD = [Bin[1], Bin[2], Bin[3], ..., Bin[16], GlobalBin].

Other mentioned techniques LBP and wavelet transform, are described above.

4.5 Performance Measurements

Retrieving videos relevant to the submitted query is the aim of content-based video retrieval. We have selected four quantitative performance measures to evaluate the suggested algorithm's potential to retrieve related videos of the user-inputted queries. All four selected performance measures are created using ground truth as a reference. We manually developed ground truth for our research. Four parameters are used to assess the performance of the proposed method: Precision, Recall, Mean Average Precision (MAP), and Accuracy. The following gives them mathematically.

Precision (P) denotes the fraction of retrieved videos relevant to the query and mathematically given in equation 4.2.

$$P = \frac{TP}{TP + FP} \tag{4.2}$$

Recall (R) denotes the fraction of relevant videos retrieved and mathematically given as in equation 4.3.

$$R = \frac{TP}{TP + FN} \tag{4.3}$$

MeanAveragePrecision (mAP) Mean Average Precision (mAP) is the average of the average precision (AP) scores for all relevant images. The equation for mAP can be written as:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} APi \tag{4.4}$$

Where N is the total number of frames and AP is a vector containing the average precision for each relevant image.

Accuracy denotes the degree of unity between the retrieved result and ground truth, mathematically given as in equation 4.5.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.5}$$

where
True Positive (TP) is the Number of videos that are identified as relevant to query by both the ground truth and the algorithm.
True Negative (TN) is the Number of videos that are identified as irrelevant to query by both the method and ground truth.
False Positive (FP) is the Number of videos that are incorrectly retrieved as relevant by the method but not present in the ground truth.
False Negative (FN) is the Number of videos that are relevant to the query as per the ground truth but not present in the retrieved result of an algorithm.

Here, all value ranges between *0(min) to 1(max)*.

CHAPTER 5

Experiment and Results

This chapter presents the analysis and evaluation of the proposed algorithm. Youtube's annotated Database[10] is used for the experiment. This database contains a total of 1080 video clips of a different category, which includes BasketBall(116 video clips), Cycling(145 video clips), Swimming (156 video clips), Golf (142 Video Clips), Horse Rididng(198 video clips), FootBall(163 video Clips), Tennis(160 video clips). From each category, 1 video clip is selected for testing purposes. The user provides a query video clip. Then similar video frames from the database are selected and displayed. These testing video clips are shown in Fig 5.1.

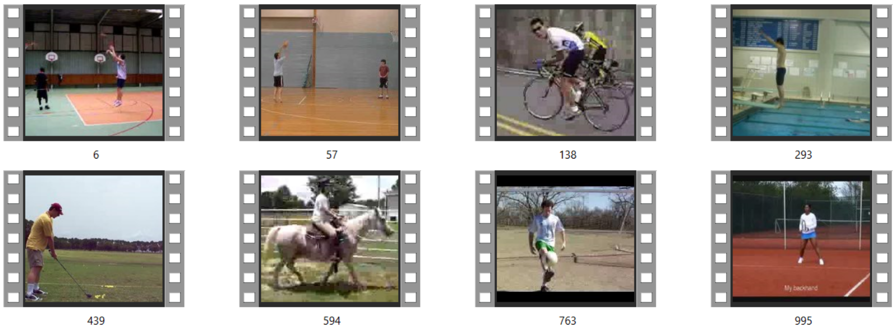
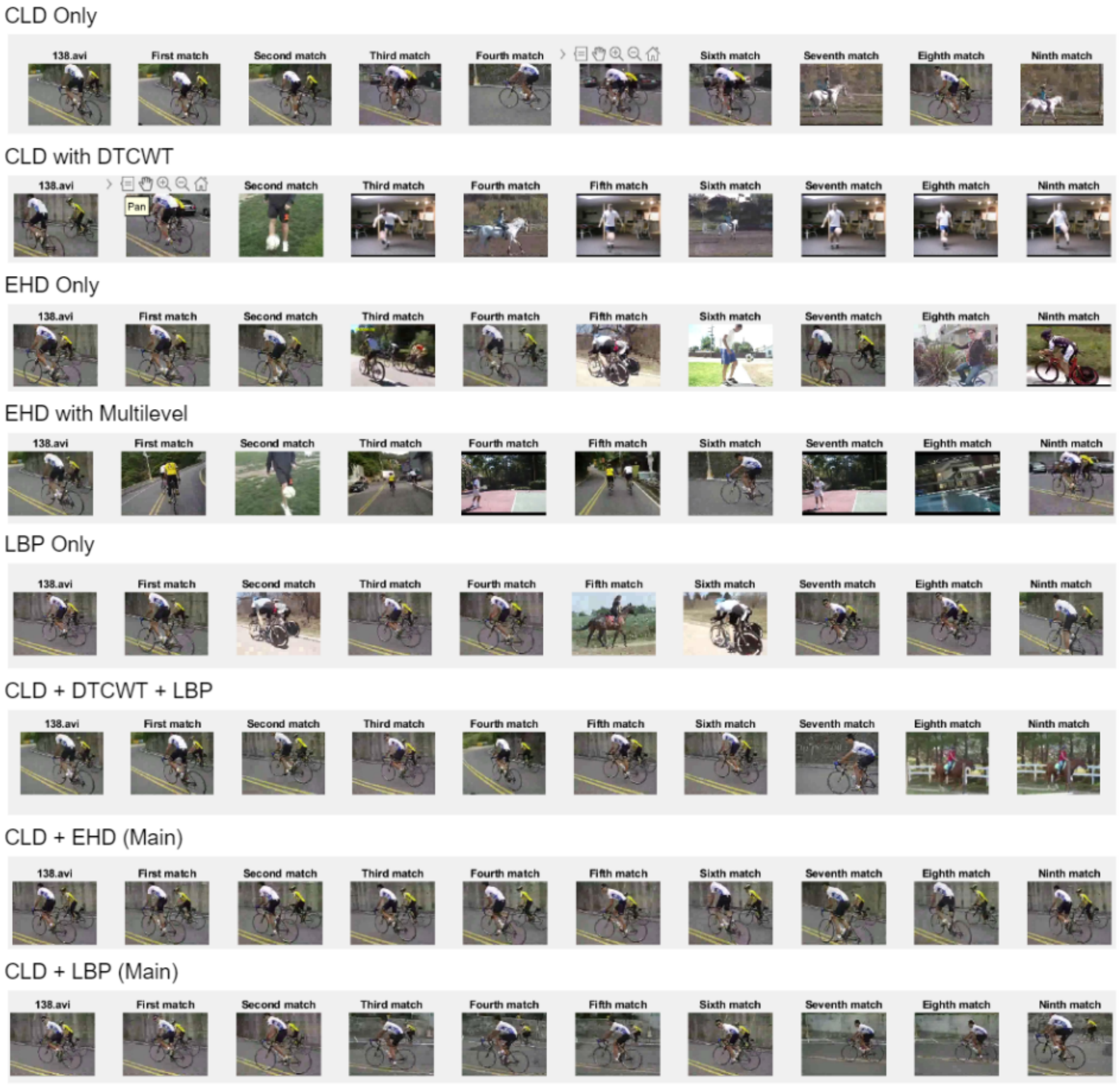


Figure 5.1: Testing Video Clips of Youtube Dataset[10]

Now, when the user gives input, it will take a random frame from the video clip and calculate its feature vector. The system will compare this feature vector with the database of feature vectors and display similar matches of the query video frame. First of all, 1 video is applied as a testing video clip to all the algorithms as mentioned above. The visual results are shown in Fig. 5.2. And comparative results are shown in Table 5.1; from This table, It can be concluded that the fusion of two features would give us better results.

Our Proposed Method, which is a fusion of colour feature (CLD) and texture feature(EHD), is tested on the different categories of video (including in testing video clips), and results are visually represented in Fig. 5.3. And quantitatively checking the performance by using different parameters is shown in Table 5.2.



CLD Only

CLD with DTCWT

EHD Only

EHD with Multilevel

LBP Only

CLD + DTCWT + LBP

CLD + EHD (Main)

CLD + LBP (Main)

Figure 5.2: Retrieved Results for different algorithms tested on YouTube Dataset[10]

Table 5.1: Comparisons of different Algorithms on Youtube Dataset[10]

138.avi	Precision	Recall	F1_Score	MAP	Accuracy
CLD Only[1]	0.68	0.68	0.74	0.60	0.78
CLD + DTCWT	0.67	0.67	0.67	0.38	0.33
EHD Only[2]	0.75	0.75	0.75	0.93	0.44
EHD(Multilevel)[20]	0.22	1	0.36	0.00	0.22
LBP Only	0.66	1	0.80	0.84	0.66
CLD + DTCWT + EHD	0.67	1	0.80	0.59	0.66
CLD + EHD(Proposed)	0.85	0.75	0.80	0.66	0.88
CLD + LBP(Proposed)	0.85	0.75	0.80	0.66	0.88

This is the first proposed method. High precision, Recall, and Accuracy values represent a good performance of the retrieval process.

The proposed method utilizes Local Binary Patterns (LBP) as the texture feature for content-based video retrieval. To evaluate the performance of the method, query-testing video clips from various categories are tested, and top nine matched results are depicted in Fig. 5.4. The retrieval performance of the proposed technique is further assessed using metrics such as Precision, Recall, Accuracy, and Mean Average Precision (MAP), as presented in Table 5.3. These performance measures provide quantitative insights into the effectiveness and reliability of the proposed method in retrieving relevant videos based on their texture features.

Table 5.2: Performance comparison of the proposed method (CLD + EHD)

	Precision	Recall	F1_Score	MAP	Accuracy
6.avi	1	1	1	1	1
57.avi	0.87	0.87	0.87	0.94	0.89
138.avi	0.85	0.75	0.80	0.66	0.88
293.avi	1	1	1	0.96	0.77
439.avi	1	1	1	1	1
594.avi	1	1	1	1	1
763.avi	1	1	1	1	1
995.avi	1	1	1	1	1

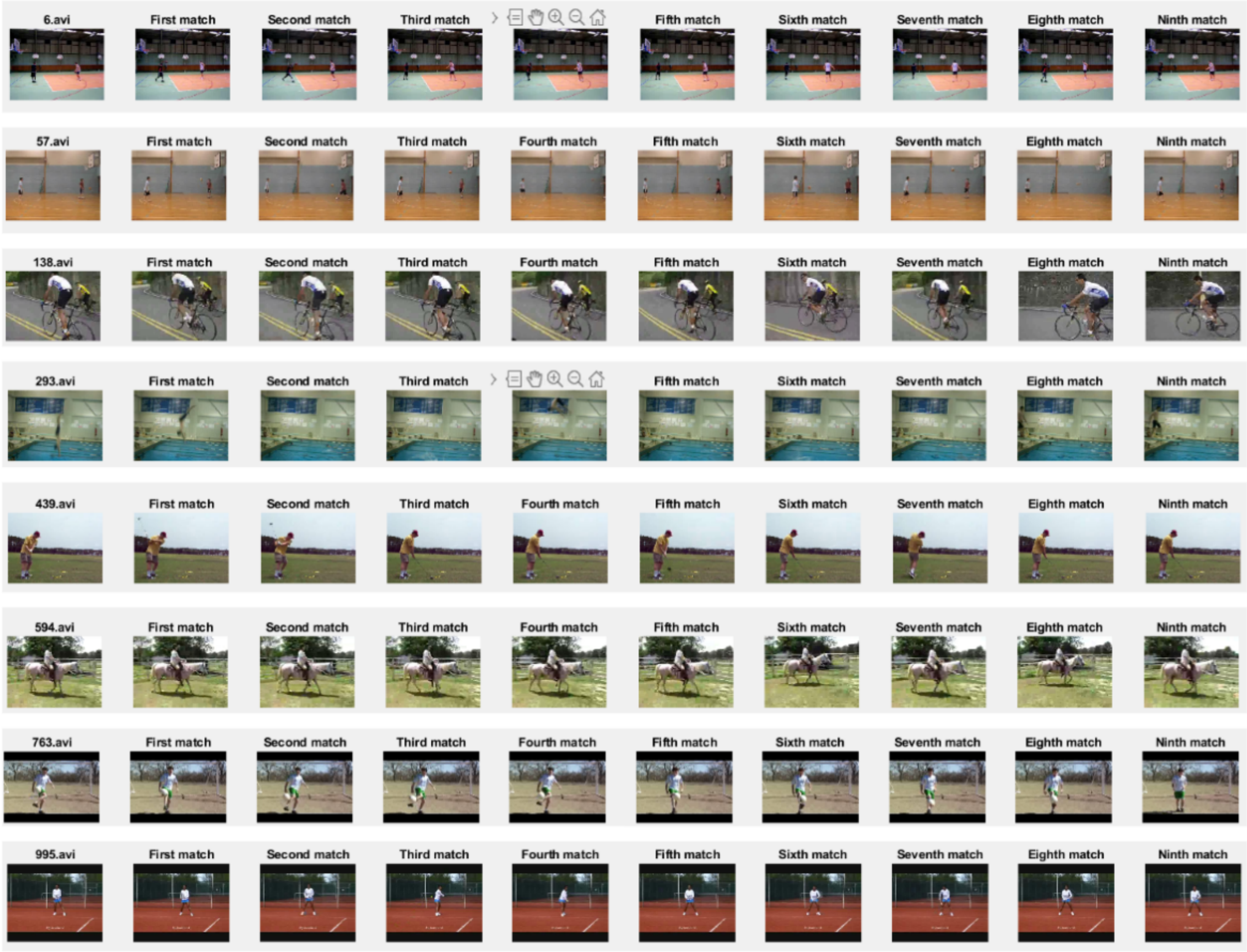


Figure 5.3: Retrieved Results for CLD + EHD algorithm tested on Youtube Dataset[10]

Table 5.3: Performance comparison of the proposed method(CLD + LBP)

	Precision	Recall	F1_Score	MAP	Accuracy
6.avi	1	1	1	1	1
57.avi	0.71	0.71	0.71	0.87	0.78
138.avi	0.85	0.75	0.80	0.66	0.88
293.avi	0.80	1	0.88	1	0.44
439.avi	1	1	1	1	1
594.avi	1	1	1	1	1
763.avi	1	1	1	1	1
995.avi	1	1	1	1	1

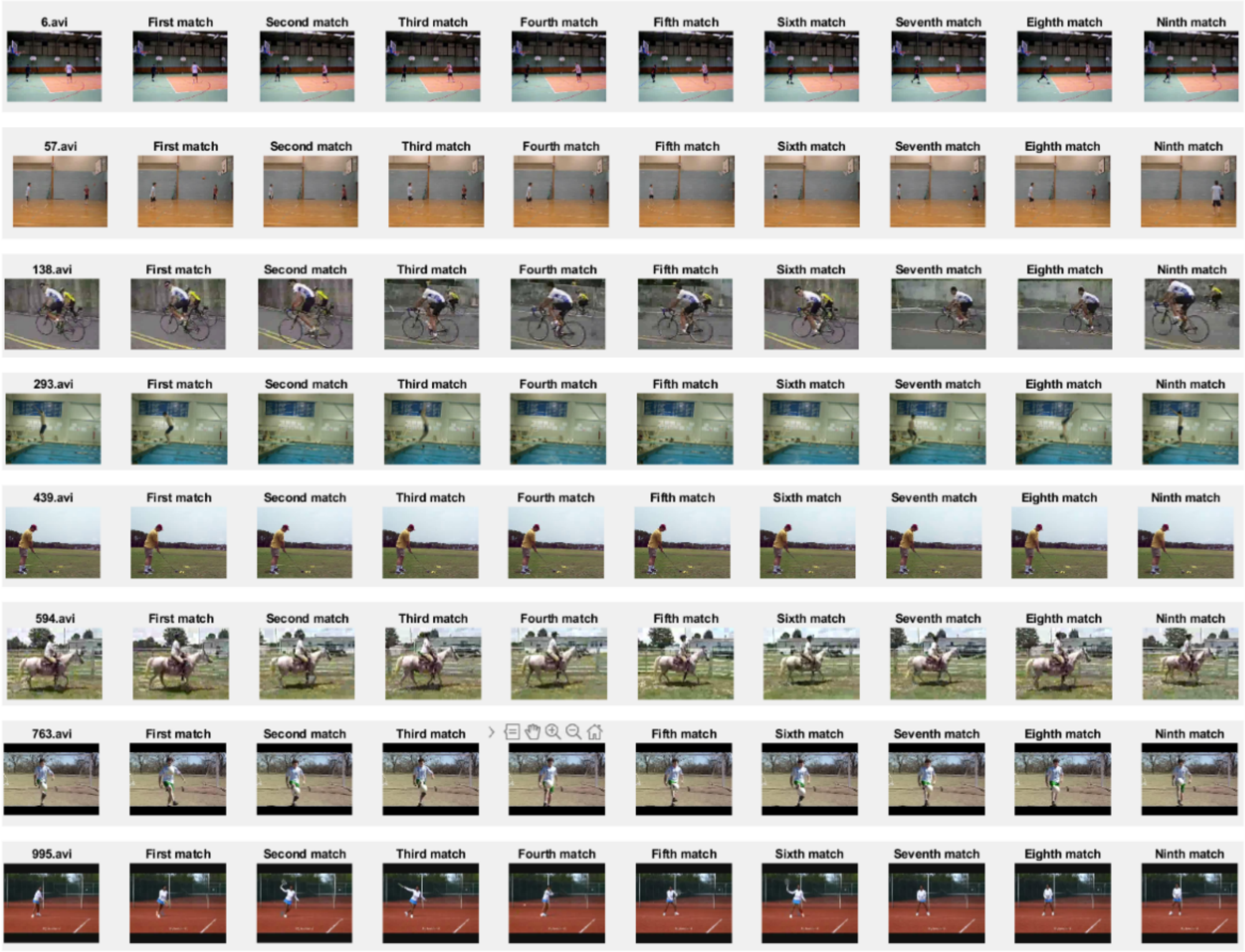


Figure 5.4: Retrieved Results for CLD + LBP algorithm tested on Youtube Dataset[10]

Another dataset used is UCF101[22]. In this dataset, there are 101 classes, but for experimental purposes, we have selected ten random videos to evaluate our proposed methods. It has a total of 13320 video clips. Randomly selected testing images are shown in Fig. 5.5.

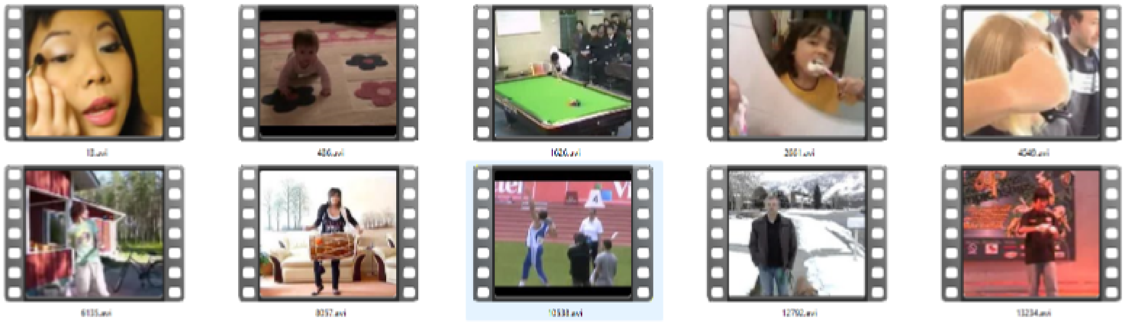


Figure 5.5: Testing video clips of UCF101[22] Dataset

The proposed fusion method, combining CLD and EHD features, demonstrates promising results when tested on randomly selected video clips from UCF101[22] dataset. The visually represented results in Fig. 5.6 showcase the effectiveness of the proposed method. Moreover, the quantitative evaluation, as shown in Table 5.4, further confirms the method's performance by assessing parameters. The high values of these parameters indicate the success of the retrieval process.

The proposed fusion method, combining CLD and LBP features, demonstrates promising results when tested on randomly selected video clips from UCF101[22] dataset. The visually represented results in Fig. 5.7 showcase the effectiveness of the proposed method in retrieving visually relevant videos. Moreover, the quantitative evaluation, as shown in Table 5.5, further confirms the method's performance by assessing parameters. The high values of these parameters indicate the success of the retrieval process.

Based on the visually and qualitatively observed results, it can be concluded that the fusion of colour Layout Descriptor (CLD) and Edge Histogram Descriptor (EHD) yields better performance compared to the fusion of CLD and Local Binary Patterns (LBP) but also the other eight combinations described in the study.

The fusion of CLD and EHD combines the colour distribution information captured by CLD with the edge-based structural information provided by EHD. This fusion approach enables a more comprehensive representation of the video content by incorporating both colour and edge features. The visual results demonstrate that the fused CLD and EHD features are capable of capturing the distinct colour patterns and structural characteristics of the videos, leading to more accurate and meaningful retrieval results.

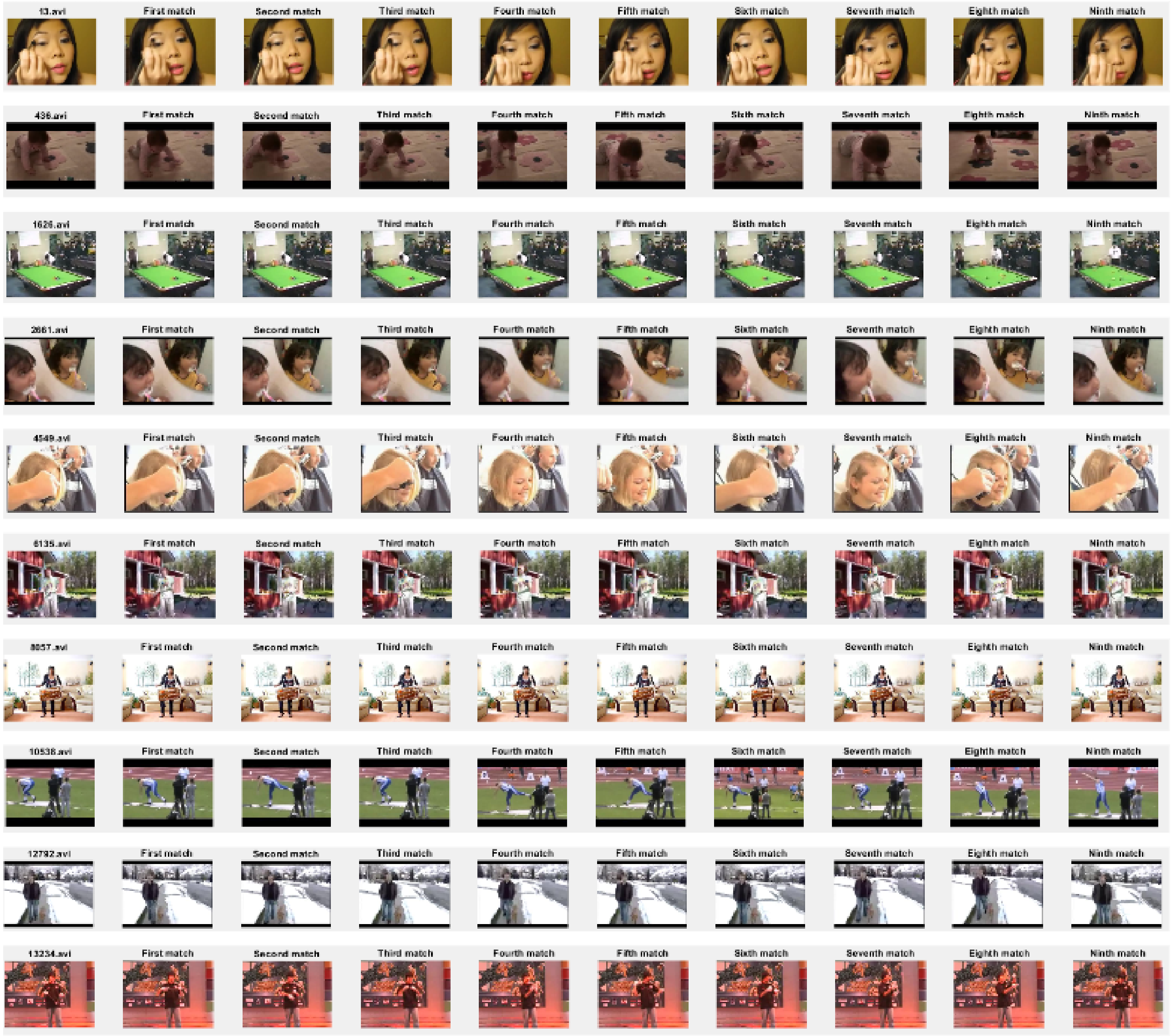


Figure 5.6: Retrieved Results for CLD + EHD algorithm for UCF101[22]

Table 5.4: Performance comparison of the proposed method (CLD + EHD) for UCF101[22]

	Precision	Recall	F1_Score	MAP	Accuracy
13.avi	1	1	1	1	1
436.avi	1	1	1	1	1
1626.avi	1	1	1	1	1
2661.avi	1	1	1	1	1
4549.avi	1	1	1	1	1
6135.avi	1	1	1	1	1
8057.avi	1	1	1	1	1
10538.avi	1	1	1	1	1
12792.avi	1	1	1	1	1
13234.avi	1	1	1	1	1

Table 5.5: Performance comparison of the proposed method (CLD + LBP) for UCF101[22]

	Precision	Recall	F1_Score	MAP	Accuracy
13.avi	0.87	0.87	0.87	0.97	0.88
436.avi	1	1	1	1	1
1626.avi	1	1	1	1	1
2661.avi	1	1	1	1	1
4549.avi	1	1	1	1	1
6135.avi	1	1	1	1	1
8057.avi	1	1	1	0.98	0.88
10538.avi	1	1	1	1	1
12792.avi	0.71	0.71	0.71	0.82	0.78
13234.avi	0.70	0.59	0.72	0.74	0.77

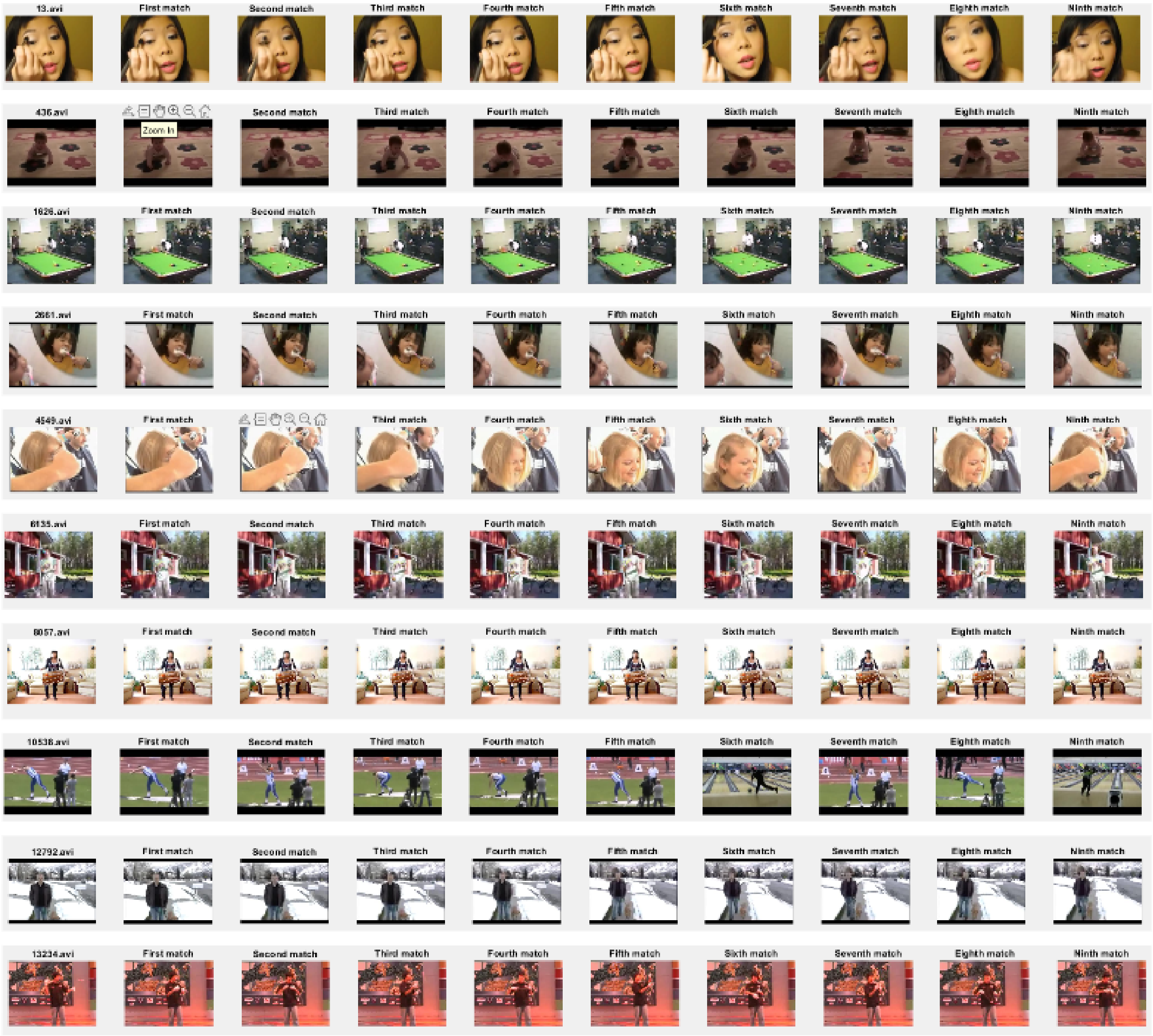


Figure 5.7: Retrieved Results for CLD + LBP algorithm for UCF101[22]

In contrast, the fusion of CLD and LBP combines the colour distribution information of CLD with the texture features extracted by LBP. While LBP is effective in capturing local texture variations, it may not be as effective in representing the global colour properties of the videos. This fusion approach might result in a less discriminative representation, as the colour information is not fully utilized or integrated with the texture features.

CHAPTER 6

Conclusion and Future Scope

In this Thesis, a content-based video retrieval algorithm is presented mainly for texture and colour-based features. After surveying papers related to image retrieval and video retrieval, that have concluded that the proposed method, which is based on the colour feature(CLD) and texture feature(EHD or LBP) along with the discrete wavelet transform(DWT), would improve the effectiveness of video retrieval. Because experimental findings validate the suggested technique over either EHD with CLD or LBP with CLD for extracting features.

In conclusion, based on the visually and qualitatively observed results, it can be stated that the fusion of CLD and EHD outperforms the fusion of CLD and LBP in content-based video retrieval. The fusion of CLD and EHD provides a more comprehensive representation of video content by integrating colour and edge information, resulting in improved retrieval accuracy and relevance.

In this Thesis, We have tried different combinations of texture(EHD or LBP) and colour(CLD) feature extraction techniques on videos, along with two different types of wavelet transform (DWT and DTCWT). In the future, to improve the effectiveness of these algorithms, one can do the multilevel resolution of different types of Wavelet Transform up to 1x1 image matrix.

References

[1] S. Agarwal, A. Verma, and N. Dixit. Content based image retrieval using color edge detection and discrete wavelet transform. In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, pages 368–372. IEEE, 2014.

[2] S. Agarwal, A. K. Verma, and P. Singh. Content based image retrieval using discrete wavelet transform and edge histogram descriptor. In *2013 International Conference on Information Systems and Computer Networks*, pages 19–23. IEEE, 2013.

[3] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod. Stanford i2v: a news video dataset for query-by-image experiments. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 237–242, 2015.

[4] A. Araujo and B. Girod. Large-scale video retrieval using image queries. *IEEE Transactions on circuits and systems for video technology*, pages 1406–1420, 2017.

[5] H. Bhaumik, S. Bhattacharyya, M. D. Nath, and S. Chakraborty. Hybrid soft computing approaches to content-based video retrieval: A brief review. *Applied Soft Computing*, pages 1008–1029, 2016.

[6] R. M. Bommisetty, A. Khare, M. Khare, and P. Palanisamy. Content-based video retrieval using an integration of curvelet transform and simple linear iterative clustering. *International Journal of Image and Graphics*, page 2250018, 2022.

[7] R. M. Bommisetty, P. Palanisamy, and A. Khare. Content based video retrieval—methods, techniques and applications. In *Advanced Soft Computing Techniques in Data Science, IoT and Cloud Computing*, pages 81–99. Springer, 2021.

[8] M. P. Chivadshetti, M. K. Sadafale, and M. K. Thakare. Content based video retrieval using integrated feature extraction. In *Fourth Post Graduate Conference, IEEE*, 2015.

[9] Y. Y. Chung, W. K. J. Chin, X. Chen, D. Y. Shi, E. Choi, and F. Chen. Performance analysis of using wavelet transform in content-based video retrieval system. In *Proceedings of the 2007 Annual Conference on International Conference on Computer Engineering and Applications, CEA*, pages 277–282, 2007.

[10] R. C. Gonzalez. *Digital image processing*. Pearson education india, 2009.

[11] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 961–970. IEEE, 2015.

[12] P. Hiremath and J. Pujari. Content-based image retrieval using color, texture and shape features. In *15th International conference on advanced computing and communications (ADCOM 2007)*, pages 780–784. IEEE, 2007.

[13] T. Jameel, S. Gilani, and A. Mumtaz. Content-based video retrieval framework using dual-tree complex wavelet transform. In *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*, pages 449–454. Springer, 2007.

[14] Y. Janarthanan, B. Jeyakumar, and S. Raghavan. Content-based video retrieval and analysis using image processing: A review. *International Journal of Pharmacy and Technology*, pages 5042–5048, 2016.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[16] A. K. Mallick and S. Mukhopadhyay. Video retrieval framework based on color co-occurrence feature of adaptive low rank extracted keyframes and graph pattern matching. *Information Processing & Management*, page 102870, 2022.

[17] B. R. Mounika and A. Khare. Content-based video retrieval using histogram of gradients and frame fusion. In *twelfth international conference on machine vision (ICMV 2019)*, pages 688–695. SPIE, 2020.

[18] B. R. Mounika, P. Palanisamy, H. H. Sekhar, and A. Khare. Content-based video retrieval using dynamic textures. *Multimedia Tools and Applications*, pages 59–90, 2023.

[19] S. Padmakala, G. S. Anandhamala, and M. A. Shalini. An effective content-based video retrieval utilizing texture, color and optimal key frame features. *2011 International Conference on Image Information Processing*, pages 1–6, 2011.

[20] C. Patvardhan, A. Verma, and C. V. Lakshmi. Robust content-based image retrieval based on multi-resolution wavelet features and edge histogram. In *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, pages 447–452. IEEE, 2013.

[21] T. Potluri, T. Sravani, B. Ramakrishna, and G. R. Nitta. Content-based video retrieval using dominant color and shape feature. In *Proceedings of the First International Conference on Computational Intelligence and Informatics: ICCII 2016*, pages 373–380. Springer, 2016.

[22] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] N. Spolaôr, H. D. Lee, W. S. R. Takaki, L. A. Ensina, C. S. R. Coy, and F. C. Wu. A systematic review on content-based video retrieval. *Engineering Applications of Artificial Intelligence*, page 103557, 2020.

[24] N. Yadav and S. Thepade. Content-based video retrieval using self-mutated hybrid wavelet transforms with assorted colour spaces. *International Conference Workshop on Electronics Telecommunication Engineering*, page 6, 2016.