Course code and name:	**F21RP -Research Methods and Project Planning**
Type of assessment:	**Individual**
Coursework Title:	**An End-to-End Framework for Content Based Video Retrieval Using Image Query**
Student Name:	MOHAMMED SHAIKATHUL ZIYAD
Student ID Number:	H00354944

Declaration of authorship. **By signing this form:**

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *MOHAMMED SHAIKATHUL ZIYAD*

Date: *20/04/2022*

Heriot-Watt University

Masters Thesis

An End-to-End Framework for Content Based Video Retrieval Using Image Query

Author:
Mohammed Shaikathul Ziyad (*H00354944*)

Supervisor:
Dr. Md. Azher Uddin

A thesis submitted in fulfillment of the requirements for the degree of MSc. Artificial Intelligence

in the

School of Mathematical and Computer Sciences

April 2022

Declaration of Authorship

I, *Mohammed Shaikathul Ziyad* confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: MOHAMMED SHAIKATHUL ZIYAD

Date: 20/04/2022

Abstract

Video retrieval refers to efficient retrieval of similar-looking videos from the video database based on the query. The query can be text, image, or video. Acknowledging that inexpensive storage, brand monitoring, content linking, and nimble video editing tools would result in a flood of unorganized video content; researchers have been developing video retrieval approaches.

In this work, we develop an end-to-end framework for content Based video retrieval using image queries. We will use video processing, Key frame extraction, feature extraction and similarity measure techniques. We will use the UCF50 dataset [34] to evaluate the performance of the proposed method. The feature extraction methods used are LBP (Local Binary Pattern), LTP (Local Ternary Pattern), LDN (Local Directional Number),LDSN(Local directional-structural pattern),HOG (Histogram of oriented gradients), VGG and ResNet. We will compare all the feature extraction algorithms to find the best feature extraction technique used for the dataset. The similarity measures used to find the similarity are Euclidian distance, Manhattan distance, Chi-Square and Bhattacharyya . We will also compare each similarity measures to find the best similarity measure for the system .The evaluation metrics will be precision and recall.

Acknowledgements

I would like to extend my gratitude to my thesis supervisor Dr. Md Azher Uddin for his support, patience time and guidance in helping me to conduct the thesis.

Table of Contents

List of Figures

Abbreviations

LBP – Local Binary Pattern

LTP – Local Ternary Pattern

LDN – Local Directional Number

LDSN – Local directional-structural pattern

HOG – Histogram of oriented gradients

1. Introduction

Video retrieval in general is to retrieve a video based on an image query or a text query. Recent years have witnessed a spike in huge amount of video data. This is mainly due to the emergence of smart devices embedded with cameras. On addition to this, video sharing on social media has been increased exponentially. Videos have the much richer content than individual images and huge amount of raw data. This increase in the video data has been challenging to correctly retrieve and index the videos which are needed for advertisement [3], video recommendation [4]and near-duplicate video detection [5][2].

Videos *are* decomposed to *key-frames*. which are frames sub-sampled from videos. Temporally consecutive and content similar key-frames are clustered into *shots*.. A 1-hour news video can be sub-sampled into 3600 key-frames with a 1 fps sampling rate. These key-frames form different shots based on their content and temporal relationship. The three-level of video, i.e., clip, shot and key-frame, determines the task grain of video retrieval[2] .

Feature extraction is a process of extracting features from the image into vector form .This is done by use of descriptors. **Visual descriptors** or **Image descriptors** give out descriptions of the visual features. They extract features such as **the texture**, the color, the shape or the motion, among others.

Visual descriptors are divided in two main groups:

- General information descriptors: Information with respect to color, shape, regions, textures and motion.
- Specific domain information descriptors: give description about events and objects in the scene like face recognition.

Local Binary Patterns (LBP) is a texture descriptor which were introduced as early as 1993 but made popular by the work of Timo Ojala and his colleagues [6]. Local binary patterns extracts the local representation of the image texture. LBP implementation captures fine-grained details in the image. It measure the spatial structure of local image texture, discards the other very important properties, i.e., contrast, by definition since it depends on the gray scale . The biggest draw back of LBP implementation is it cannot capture the image texture details on varying scale of pixels i.e., the window size of pixels to compute the LBP value. An extended version to the original LBP implementation was proposed by Ojala et al. and his colleagues to handle variable neighborhood sizes [6].

Local Ternary Pattern (LTP) are basically a texture descriptor which are the extended version of Local Binary Pattern (LBP). LTP uses a constant to threshold the pixels. Unlike LBP, LTP uses three pixel values i.e., 1,0,-1. Xiaoyang et al and his colleagues proposed a LTP to resolve the sensitivity in noise in uniform pattern region[11]. LTP has better performance in face recognition application than LBP[12]

Local Directional Number Pattern is an edge based descriptor which uses the sign of the directional numbers to encode the structural information of the image [26]. Consequently, a pattern is created by computing the edge response of the neighborhood using the compass mask, and by taking the top directional numbers, that is, the most positive and negative directions of those edge responses that provide valuable information of the structure of the neighborhood [26]. LDN provides a representation of texture which is more consistent in the non-monotonic illumination variation and in the presence of random noise [26]. LDN is suitable for real time applications but more expensive than LBP [25].

Local directional-structural pattern also known as LDSP uses the positional relationship of the top edge responses of the target pixel to extract more detailed structural information of the local texture. LDSP is developed by Makhmudkhujaev and his colleagues to eliminate problems in edge-based descriptors [27]. Experimental study shows that LDSP surpasses existing descriptors and other state-of-the-art methods in several datasets for person-independent expression recognition [27].

The histogram of oriented gradients (HOG) is a feature descriptor. It is mostly used for object detection. The concepts of HOG descriptor was introduced in the year 1986 but was not termed as Histogram of oriented gradients. In 2005, the HOG descriptor was proposed by Dalal and Triggs [29]. Recently, HOG descriptor with SVM classifier is commonly used in various applications for robust object detection and recognition [29]. HOG is one of the well-recognized features due to its superior performance and relatively simple computation [30]. There are many derivatives of the original HOG descriptor. The basic steps of these derivatives are similar. A set of HOG classifiers is trained to recognize different orientation of a vehicle [31]. As a results, the HOG features are a good descriptor to identify a car in the outdoor conditions [31].

Similarity matchings are the method which calculates the similarity are measured in the feature vectors, which are coordinated with images extracted through query, it is also known as the distance method. There are various distance processes like Canberra Distance, City Block Distance, Euclidean Distance [22].

In this work, we develop an end-to-end framework for content Based video retrieval using image queries. We will use video processing, Key frame extraction, feature extraction and similarity measure techniques. We will retrieve videos from the local database with respect to the image query passed as the input. We process the videos from the database into frames. These frames are used to select the key frames which are the important frames in the sequence. These key frames are further used as an input for numerous feature extraction algorithm used. Similarly, the input query, the image will be processed with the same feature extraction method. We will use the UCF50 dataset [34] to evaluate the performance of the proposed method. The feature extraction methods used are LBP (Local Binary Pattern), LTP (Local Ternary Pattern), LDN (Local Directional Number), LDSN (Local directional-structural pattern),HOG (Histogram of oriented gradients), VGG and ResNet.
The aim of this research is to compare all the feature extraction algorithms with one another and to find the best feature extraction technique used for this dataset. The similarity measures used to find the similarity are Euclidian distance, Manhattan distance, Chi-Square and Bhattacharyya . We will also compare each similarity measures to find the best similarity measure for the system. The evaluation metrics for comparison will be precision and recall.

1.2 Motivation

The use of YouTube and google search engines have been common nowadays. Drastic increase of videos due to social media made the web comprise huge amount of data. To find a particular video on the net, text query is mostly used. Google images supports the search of data with respect to image query. Video search using a video is exceedingly rare.

The video retrieval system using image query can be installed as an automated security system looking for specific image targets. This highlights the importance of video retrieval using image query and the motivation to further improve the performance.

2. Literature Review

In this section, we give a brief review of previous literatures that closely relate to our work in either aspect of problem. Video retrieval follows the same methods of image retrieval but with extra set of computation in video processing and retrieving. This section will talk about retrieval of images and videos from text query, Image query and video query.

2.1 Query by text

Retrieving image based on text is also known as TBIR (Text based Image retrieval). The basic idea of using text as a query is to retrieve the images from the existing database based on the textual properties of the image. Some of the properties used are mentioned below.

- Text as keyword index
- Text to describe the properties of the image
- Text as tag

Many techniques have been developed for text based image retrieval [19][17][7] and they are proved to be highly successful for indexing and querying web sites [17] . Some of the system use tags as a query for retrieving the images. This system very efficient and effective. The performance of this system proportional to the availability and quality of manual tags [18]. Another system uses the text query to match the text in the images to retrieve the image [20].

As mentioned above TBIR uses only text information to retrieve the images. This system has many complications. The problems with this system are listed below.

The image retrieved cannot be always accurate as it might not conform to the theme of the image [14]. There can be an issue where same image is retrieved using different words. This is termed as *synonymy* [7]. This has an impact on the recall of most information retrieval system [7]. The problem of annotation of the image can rise laboring cost [15]. The performance of this system can be degraded when the contextual information of the image is noisy and incomplete [16]. If the images are well annotated the text-based image retrieval is fast and reliable, they are incapable of searching in unannotated image collections [17].

Let us look into some of the previous TBIR application.

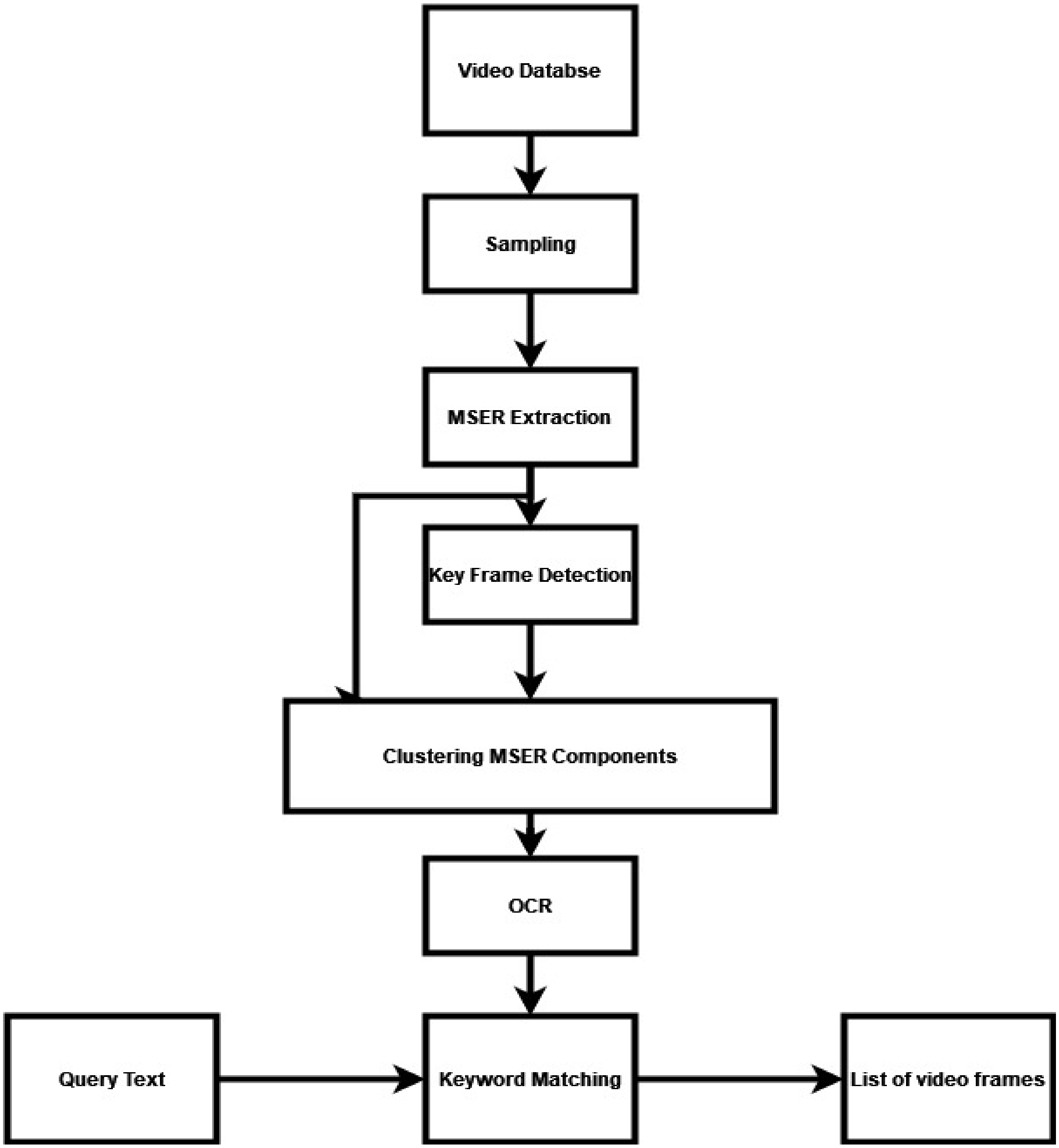


Figure 1 TBIR using MSER Components

Werachard Wattanarachothai and his colleague proposed a new text-based video content retrieval system [37].

The proposed system consists of three main processes. They are

- **key frame extraction**: Maximally Stable Extremal Region (MSER) based feature is introduced for key-frame extraction, which is oriented to segment shots of the video with different text contents. It is also used to reduce the redundancy among consecutive frames.

- **text localization:** a text localization scheme based on Maximally Stable Extremal Regions is developed. The MSERs in each key frame are clustered based on their similarity in position, size, color, and stroke width.
- **keyword matching**: Keyword matching is done by using Tesseract OR engine i.e., recognize the text regions.

Four input images are derived from different pre-processing methods into the Tesseract engine which will improve the recognition accuracy. The text query for querying is matched with OCR in they key matching phase. This returns a set of similar video frames. The videos that are be segmented by using efficient number of shots and provide the better precision and recall in comparison with a sum of absolute difference and edge-based method [37]

Wen Li and his colleagues proposed a new approach to learn a robust classifier for text-based image retrieval (TBIR) using relevant and irrelevant training web images [16].

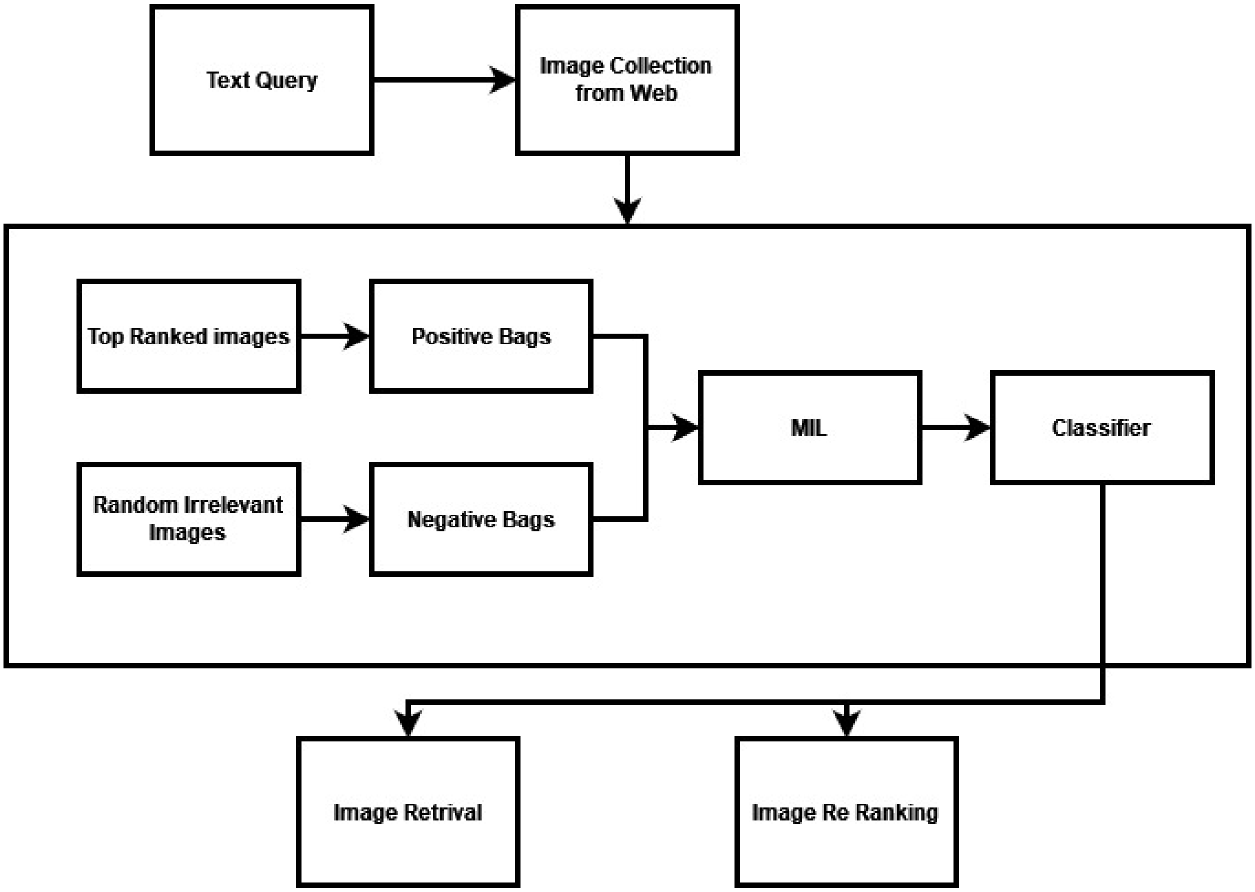


Figure 2 TBIR using MIL

The noise is explicitly handled in the loose labels of training images. On text query, images from the web are automatically collected. These images can be either relevant i.e., surrounding text can contain the query text or irrelevant images.

The images are split into clusters. There are two main type of clusters, the relevant cluster which are the top ranked images and the irrelevant cluster. Each cluster is treated as a bag and the instances in these bags are the images in it to treat the system like Multi -instance learning problem. Firstly, an algorithm called MIL-CPG was introduced to use the constraint on the positive bags to classify the labels of the images. The algorithm was not always giving the best results on the data as the web images are generally associated with noisy textual data.

A new algorithm called Progressive MIL-CPB was introduced to tackle this issue. MIL-CPB proved to be better by improving the overall performance of the system. Below is the performance figure of progressive MIL-CPB .

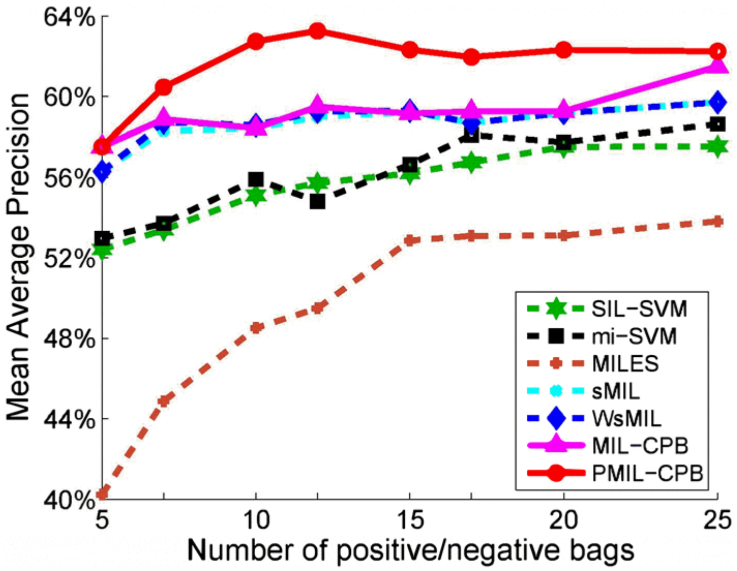


Figure 3 performance measure of classifiers

Progressive MIL-CPB or PMIL-CPB gives greater precision than the rest.

David Grangier and his colleague developed a system to retrieve images from text queries based on neural network[41].

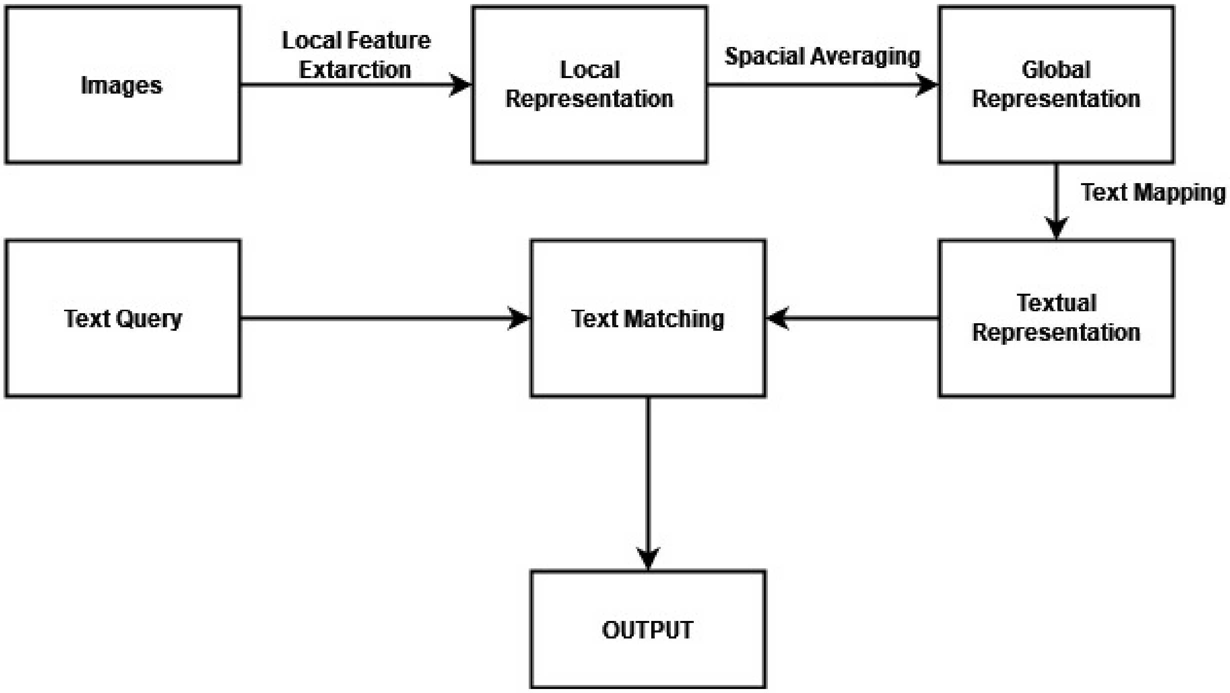


Figure 4 TBIR using discriminative approach

The proposed uses a discriminative approach . The system consists of two main modules. The local representation consists of data from image feature extraction. The global representation aims at extracting global image features from a set of block descriptors. The other module focusses on retrieving task from the features. These two modules are trained as a neural network to reduce the loss for retrieval.

The proposed system uses Block based neural network (BBNN). This consists of

- Local feature extraction layer: extracts local feature descriptors from different positions of the input picture
- Special averaging layer computes the average of the local feature vectors
- Text mapping layer: The text mapping layer then projects the output of Special averaging layer into the text representation
- query matching layer compares the input query with the obtained textual vector an computes the output

2.2 Query by image

Content-Based Image Retrieval (CBIR) was introduced in the early 1980s [21] to overcome the above disadvantages of text-based retrieval system. CBIR is a technique to retrieve images based on the visual content of the image. Visual contents are low level features extracted from the image like color, shape, or texture etc. [15]. The main advantage of CBIB is because it depends inly on metadata, and they are reliant on explanation value and comprehensiveness. In the end the images will be indexed according to their own visual content with respect to the features like color texture, or any other visual feature.

CBIR methods are more capable of searching and retrieving in large database collections.

Query by image content (QBIC) is another form of CBIR. It is an automated technique which takes an image as a query and returns a set of images similar to the query form large set if database [15]. The query image is converted into the internal representation of feature vector using various feature extraction techniques. Some of the known content based image retrieval system are Google Image search, VisualSeek, Blobword etc.[15].

The CBIR methods are still not completely reliable or fast enough to handle huge amount of data. Some of the key problems to be noted in CBIR:

1. Deciding a set of feature extraction technique is a crucial task. As there is no single feature extraction technique can provide all robust feature
2. Selecting the handy features which are most important. This can be done by using dimensionality reduction to reduce the size of the feature and applying feature selection techniques.
3. Selecting the best method to match the similarity between the images
4. Ranking the retrieved images
5. Achieving high precision and recall [15].

Similarity matchings are the method which calculates the similarity are measured in the feature vectors, which are coordinated with images extracted through query, it is also known as the distance method. There are various distance processes like Canberra Distance, City Block Distance, Euclidean Distance [22].

Therefore, to set up a video retrieval by image query CBIR will be a part of the system rather than TBIR. It is termed as CBVR (Content Based Video retrieval). The idea is to match the image from the search query to shots from the vide using CBIR and retrieve the most similar video.

There are numerous video retrieval implementations done. Below listed are some of the interesting approaches.

Vishakha Soni and colleague [7] proposed a system where the data input query (search query) can be either a text or an image query. The system uses Euclidian distance to calculate the similarity between the color features.

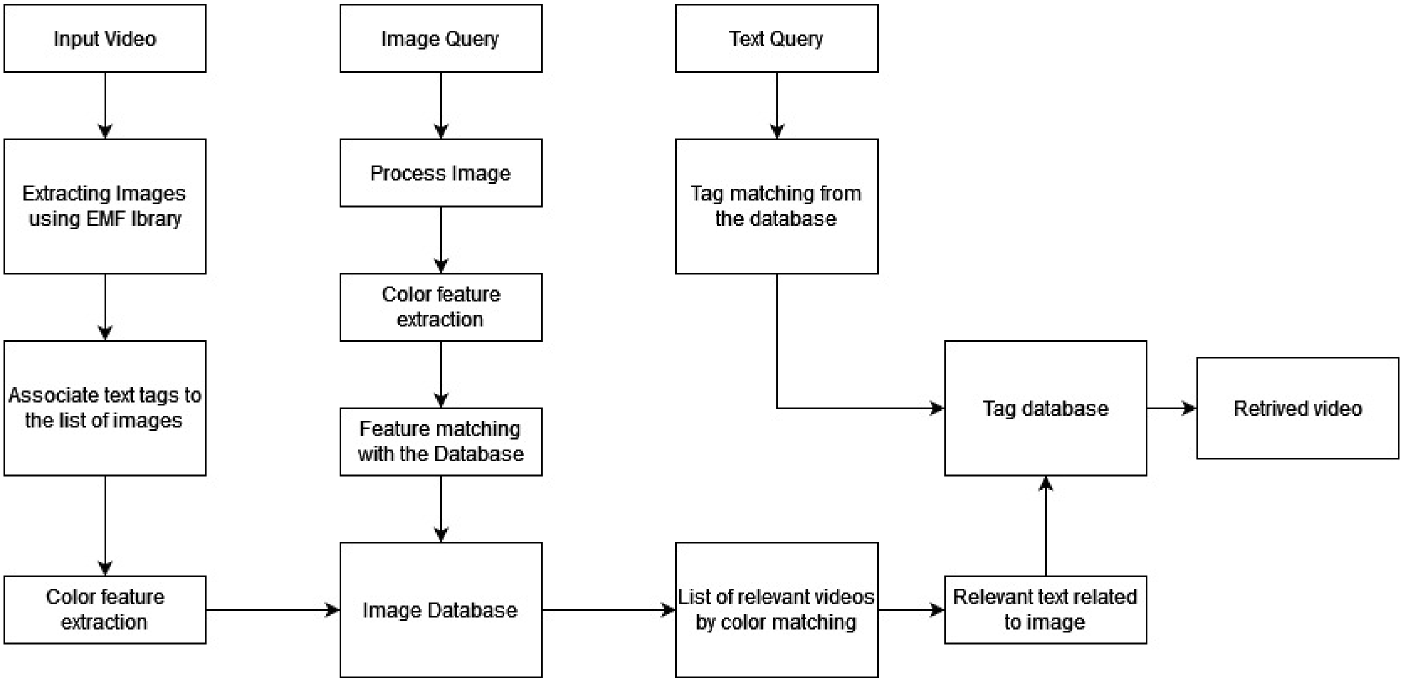

Input Video
Extracting Images using EMF lbrary
Associate text tags to the list of images
Color feature extraction
Image Query
Process Image
Color feature extraction
Feature matching with the Database
Image Database
Text Query
Tag matching from the database
List of relevant videos by color matching
Tag database
Relevant text related to image
Retrived video

Figure 5 CBVR using text or video

The system first consists of set of videos which are stored in the database. The video is broken down into images. The EMF (Enhanced Metafile) library is used to extract image from the video. For each list of images tags are associated to it. Feature extraction is run of each list of images. The color feature extraction is used. The processed image is stored in the image database. The tags are stored int eh tag database.

User can use image or a text as query to search for videos. The input image query I passed through as set of pipelines which process the image. Color feature extraction is again performed on this image. The image is then matched with the image database. to retrieve the video. When a tag I passed as an input query. it checks the tag data base for similar tags and retrieve the videos from the database. Basically It checks the similarity of the search query using Euclidian distance to get the most similar text/color feature of the image.

M. Wang and his colleague Proposed a video retrieval system using deep feature [2]. Th input to this system is an image query. The paper aims to reduce the memory cost for extracting key frame of the videos. The system uses deep neural network to recognize the deep neural network to reduce the memory cost. Deep feature is used to detect shots of similar key frames. These are then represented using aggregation techniques. This helps in removing the redundant key frames of the video.

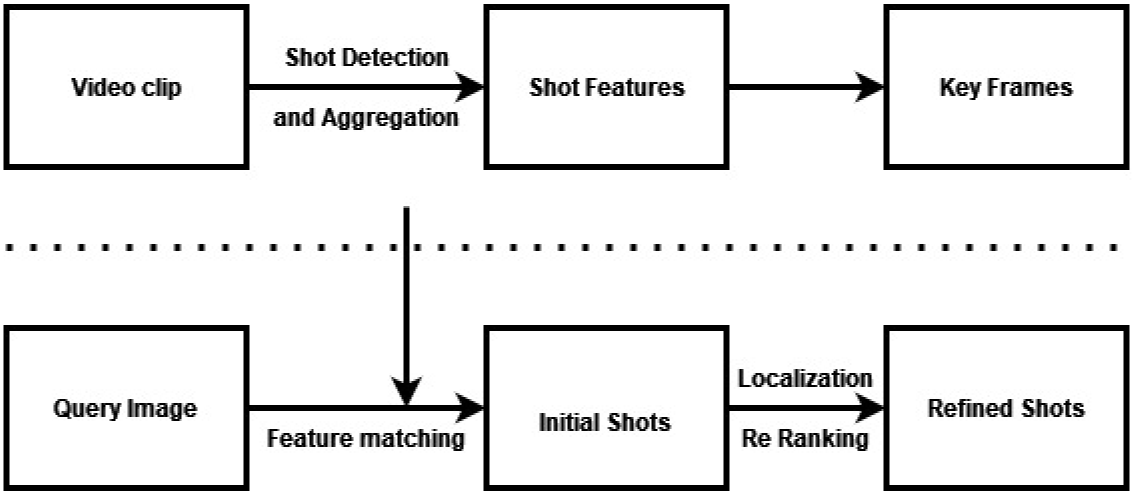


Figure 6 CBVR using deep feature

Firstly the video is converted into shots. This is done by MAC feature. The MAC feature is derived from CNN . It represents the visual contents of key frames which is used to detect shot boundaries. The MAC feature proposed is used to calculate the boundary of the shots. Three different aggregation technique is used to represent each key frame. These are used to reduce the memory cost.

1. **Image Level Aggregation:** For each shot of the video, merge the key-frames to get the average image. This average image is used as an input to extract the MAC feature. This feature is used as the representation for the shot. MAC feature is the final short descriptor in which the shots are with single key frame
2. **Global Feature Aggregation:** For each key frame, the MAC feature is extracted. These features in single shot are aggregated to form a unique representation.
3. **Local Feature Aggregation:** Intermediate feature maps are used to perform the aggregation. The key frame feature maps in a single shot is merged with respect to the convolutional kernel by max pooling. This data is used to extract MAC features.

The input query is matched with the above key frames. Set of videos are retrieved. The system uses a two way localization to re rank the retrieved videos. To automatically locate the

ROI(region of interest) with respect to the input query and video key frame without any annotations, a two way AML approach is used.
Using this approach the matched ROI are to be most similar between their MAC feature.
Firstly query image is used to match the best region in the key-frame. Then recompute the MAC feature to find the top ranked result.

Patil and his colleague proposed an image retrieval system based on image query [42]. The system compares different similarity measures to retrieve the images from the data base.

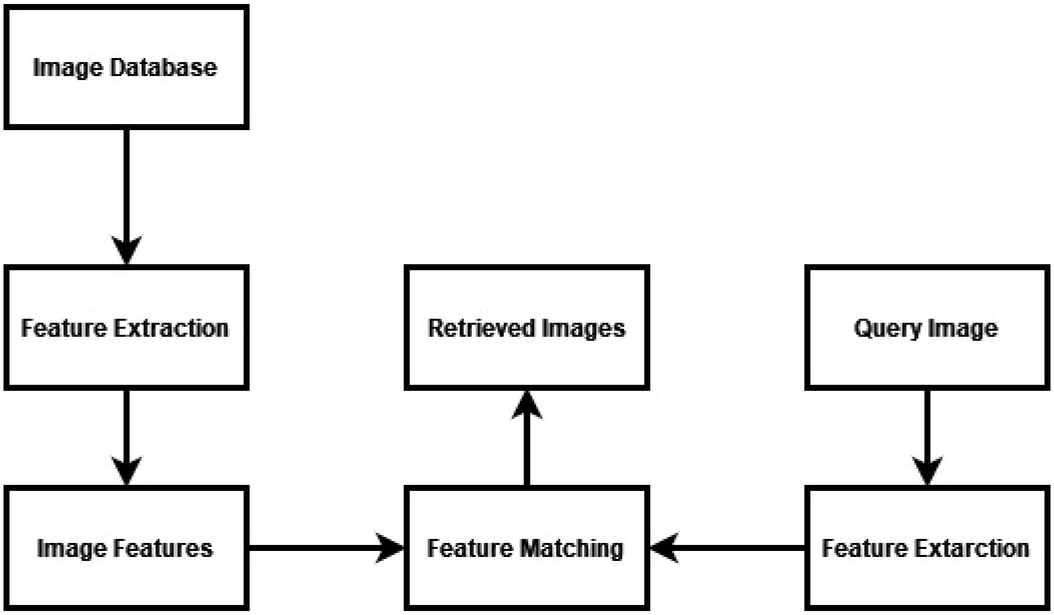


Figure 7 CBIR using similarity measures

The system is comprising of image Database, feature extraction phase, Feature matching phase. The local database stores all the images in it. These images are forwarded to feature extraction algorithms. Texture based feature extraction techniques are used.

Texture representation used are

- Co-occurrence Matrix
- Tamura Texture
- Wavelet Transform

The same procedure is followed by the input query image. These image features from the database and input query is matched in feature matching phase using similarity measures.

Retrieved Images	**Distance between query image and database image for different Distance Metrics**					
	Euclidean	**Manhattan**	**Canberra**	**Bray-Curtis**	**Square Chord**	**Square Chi squared**
IMAGE1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IMAGE2	0.3480	0.7369	0.0317	0.0317	$1.12x10^{-4}$	0.0040
IMAGE3	0.3679	0.6275	0.0273	0.0273	$1.30x10^{-4}$	$8.29x10^{-4}$
IMAGE4	0.4403	0.7746	0.0334	0.0334	$1.79x10^{-4}$	0.0083
IMAGE5	0.4685	0.9669	0.0420	0.0420	$2.07x10^{-4}$	0.0076
IMAGE6	0.4707	0.8385	0.0360	0.0360	$2.09x10^{-4}$	0.0112
IMAGE7	0.4828	0.8350	0.0365	0.0365	$2.24x10^{-4}$	0.0067
IMAGE8	0.9706	1.6366	0.0707	0.0707	$8.80x10^{-4}$	0.0340
IMAGE9	1.0460	2.1135	0.0905	0.0905	0.0011	0.0504
IMAGE10	1.7690	3.0113	0.2341	0.2341	0.0017	0.0927

Interesting results from this system is that Square Chi-squared and Square chord distances performed better than the Manhattan distances and Euclidean.

A. Khare and colleague introduced a content based video retrieval system which uses ICTSLIC (Integration of Curvelet transform and Simple Linear Iterative Clustering) segmentation [43]. The Proposed system uses image as a input query i.e.., query by image frame work

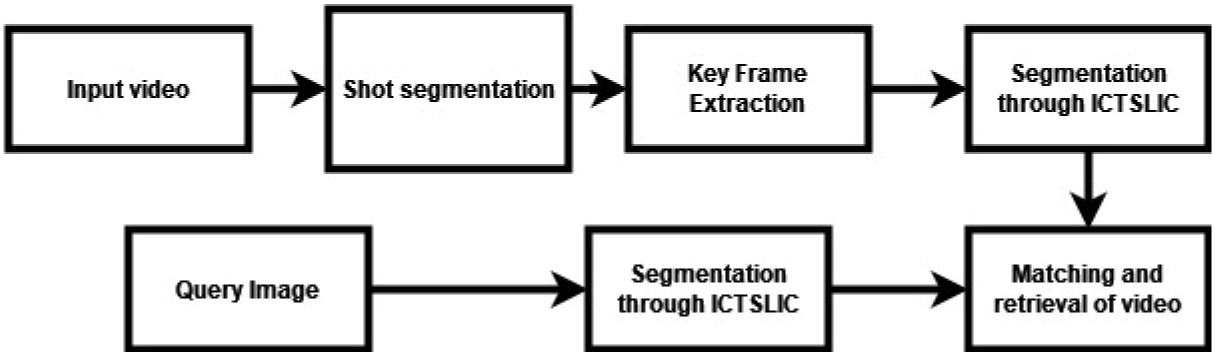

Input video
Shot segmentation
Key Frame Extraction
Segmentation through ICTSLIC
Query Image
Segmentation through ICTSLIC
Matching and retrieval of video

Figure 8 CBIR using ICTSLIC

The system contains three phases. The offline, online and retrieval phase.

In the offline phase shot transition boundaries is established to identify videoclip keyframes in the database. Shot segmentation extracted the key frame in every shot which has the highest standard deviation. In the next step, segmented query frames are extracted from query frame using ICTSLIC algorithm. Incorporation of Curvelet transform and Simple Linear Iterative Clustering algorithm (ICTSLIC) is mainly accustomed to generate super pixels. The final step, the segmentation results of the extracted query frame are searched to obtain matched and retrieve the result from the database. The process of matching is done using similarity measure called Euclidian distance. Where it calculates the distance between the segmentation of query image with respect to all the segmented video images in the database. These are calculated and stored in an array. The top ranked videos are retrieved.

Basically, when processing the input video from the database if offline process. Processing the query image is the online process. Finding the matching videos with respect to the image query is the matching and retrieval phase.

2.3 Query by video

Video retrieval system using video query. The system checks the database and retrieve the similar video. The main difference in image query and video query is the computational cost as image query uses a single frame for computation whereas video is composed of sequence of still frames.

The main issue will be to reduce the computational time to search and comparison of image sequence. Due to the large size of the database, retrieval systems use search functions for retrieving the similar video. Some in practice video search engines such as YouTube, Vimeo etc. retrieve videos based on the metadata mostly created by the user such as title, genre. This often misleads the retrieved video. This is one of the reasons content-based retrieval is acknowledged.

N Poornima and her colleague proposed an automated approach to retrieve lecture videos using context based semantic features and deep learning [39]. The system developed uses feature extraction algorithms in key frame extraction. Precision and recall are used as the evaluation measures.

Some of the challenges in developing the system are

- Recognition of the teaching. Complexity may increase if the search topic is unavailable.
- it is more challenging to retrieve the lecture video than others as it has low level correlation among the features of different videos.
- performance may degrade while using a deep neural network like CNN, for the rotated image query due to the global descriptors.

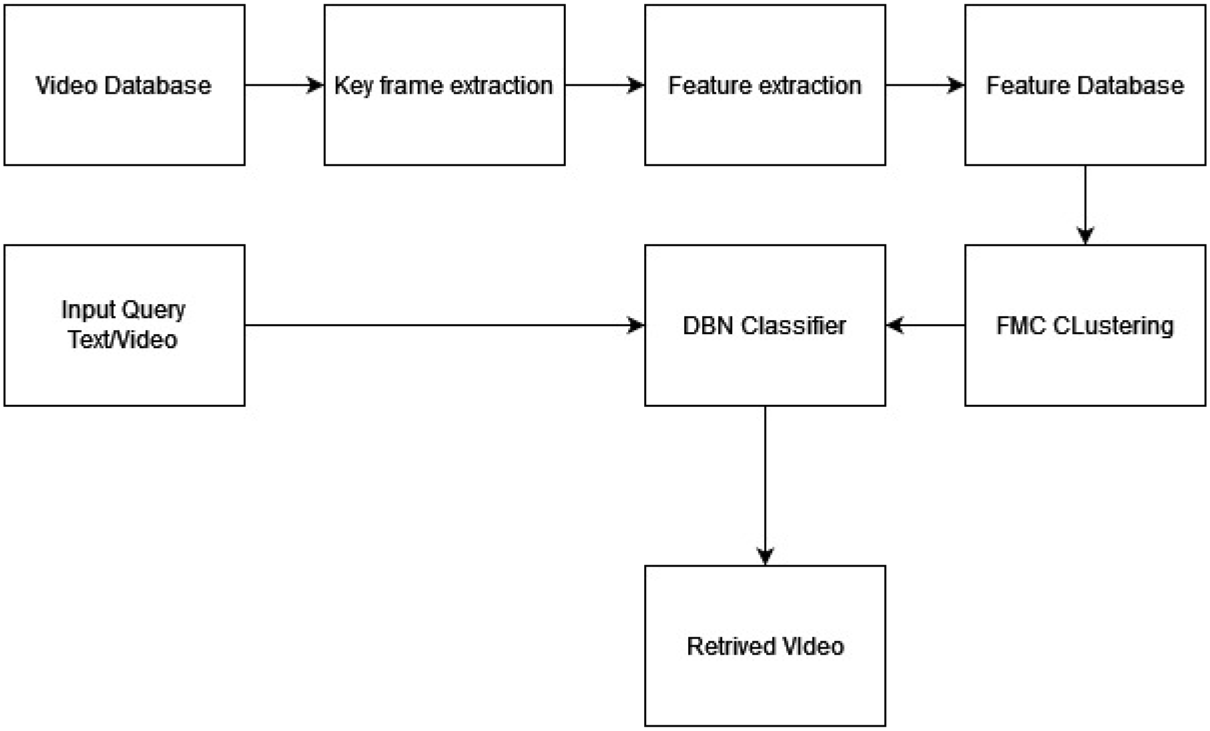


Figure 9 VBVR using automated technique

The proposed architecture for video retrieval using deep learning.

From the database, the key frames are extracted from the videos. This is then then passed through the feature extraction phase where semantic words, context words and LDP features are extracted. These features contribute in making of the feature database. The feature database is subjected to the FCM clustering. The cluster centroids are used to train the DBN classier to retrieve the video. When a user passes a video query all the above mentioned processes are done first and then passes to the DBN classifier as a test data where the DBN classifier identifies the optimal cluster the query belongs to . The video related to the optimal cluster are retrieved.

P. Chivadshetti and her colleague developed a video retrival system which takes text/image or video as input query. The system is divided into three phase that is video segmentation and key frame detection phase, extract textual keyword phase, feature extraction phase. Similarity measures are used. The results can also be personalized to the users interest.

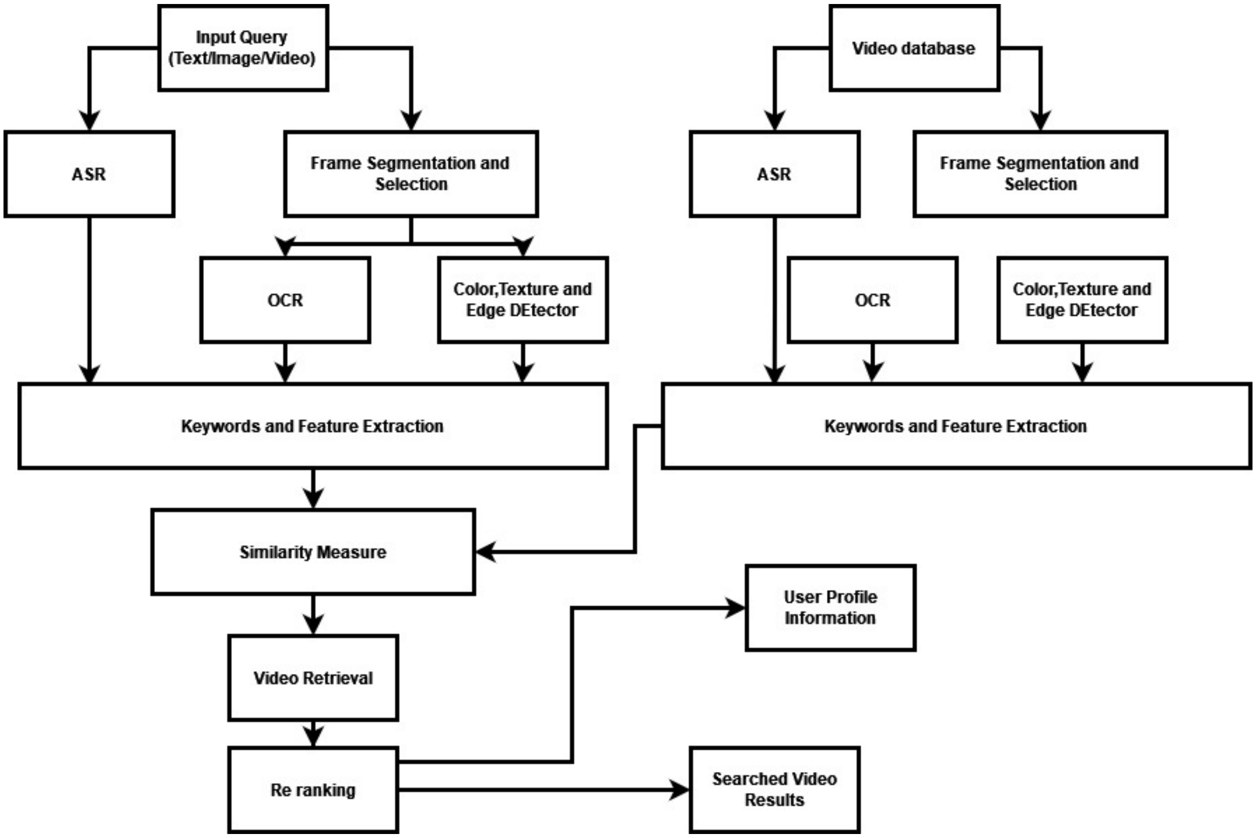

Input Query (Text/Image/Video)
ASR
Frame Segmentation and Selection
OCR
Color,Texture and Edge DEtector
Keywords and Feature Extraction
Video database
ASR
Frame Segmentation and Selection
OCR
Color,Texture and Edge DEtector
Keywords and Feature Extraction
Similarity Measure
Video Retrieval
Re ranking
User Profile Information
Searched Video Results

Figure 10 VBVR using text/image/video as input

The Proposed system is implemented on several modules. When the user inputs a query into the system. This query can be a text, image, or video. If a video query is passed. The video is divided into frames. It undergoes a process of selecting only relevant frames i.e., key frames. Simultaneously Automated Speech recognition (ASR) is processed on the video by the ASR tool and relevant information is extracted. After the key frames are selected , OCR is performed for text recognition. HOG, gabor filter and OCR text are extracted from OCR. Feature extraction is done simultaneously on the key frame . In the end features and keywords are extracted from the this data.

The same process is done to the local video database and stored in the database. Once the user inputs the input query, system will check the similarity in features or keywords of the query metadata with all the video meta data stored in the database. This is done using similarity measures. The matching keywords or features or ASR data or OCR text is retrieved . This is then re ranked and given as the output video.

A separate user interest model will contain user detail i.e., search history, related keywords. The personalized results are given by re ranking the videos based in user profile.

Yang Cai and colleagues developed a VBVR technique which uses word based approach to retrieve the video.

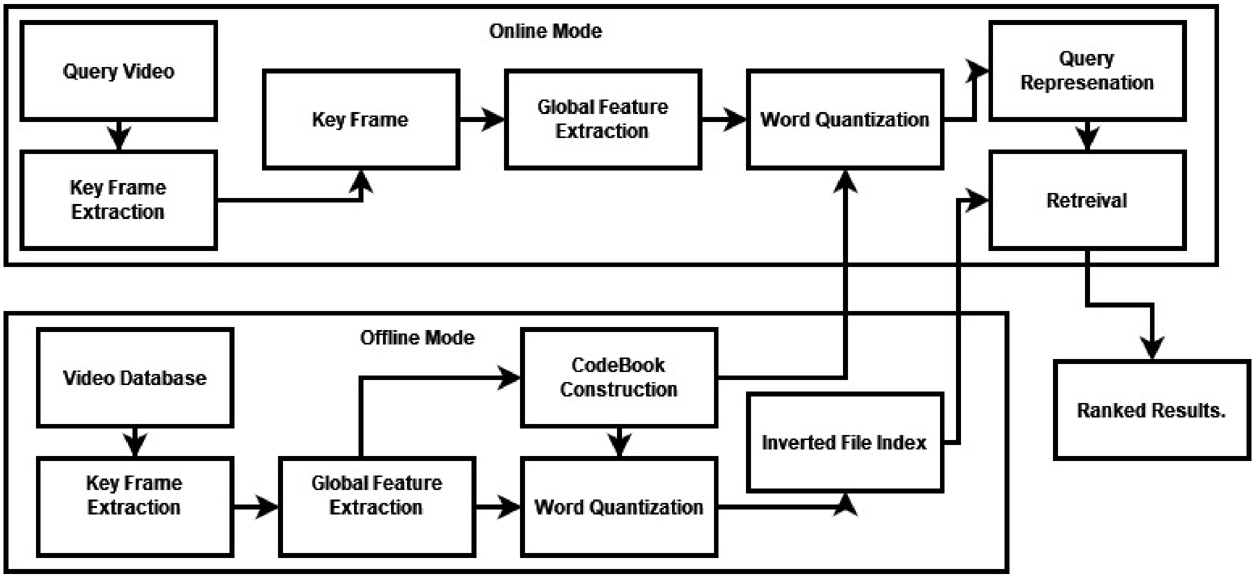


Figure 11 VBVR using word quantization

The word based approach is used to depict each video frame and represents as BOG(ag of words)

From the video , key frames are extracted using key feature extraction. The system does not extract ordinal relations from frames or encode the heuristic ordinal relations. The is extracts global features that are discriminative on image processing techniques like color. These global features are quantized using a clustering algorithm. The clustering algorithm used here is K Means. This is with respect to the online mode. The same process is done in offline mode with the input query. An additional code book is used . K Means tree search algorithm and K Means clustering algorithm is used to construct the code book to increase the speed of quantization and construction of codebook. A visual vocabulary is constructed after the key frames are created based on clustering. The inverted file index consists of indexed visual words and the model used to retrieve the video. The model uses cosine retrieval with tf-idf weight. It computes the relevance score between the input query and the local database videos.

3. Methodology

3.1 Data

UCF50 is an action recognition data set with 50 action categories, consisting of realistic videos taken from YouTube [34]. This data set is an extension of YouTube Action data set (UCF11) which has 11 action categories. For all the 50 categories, the videos are grouped into 25 groups, where each group consists of more than 4 action clips. The video clips in the same group may share some common features, such as the same person, similar background, similar viewpoint, and so on.

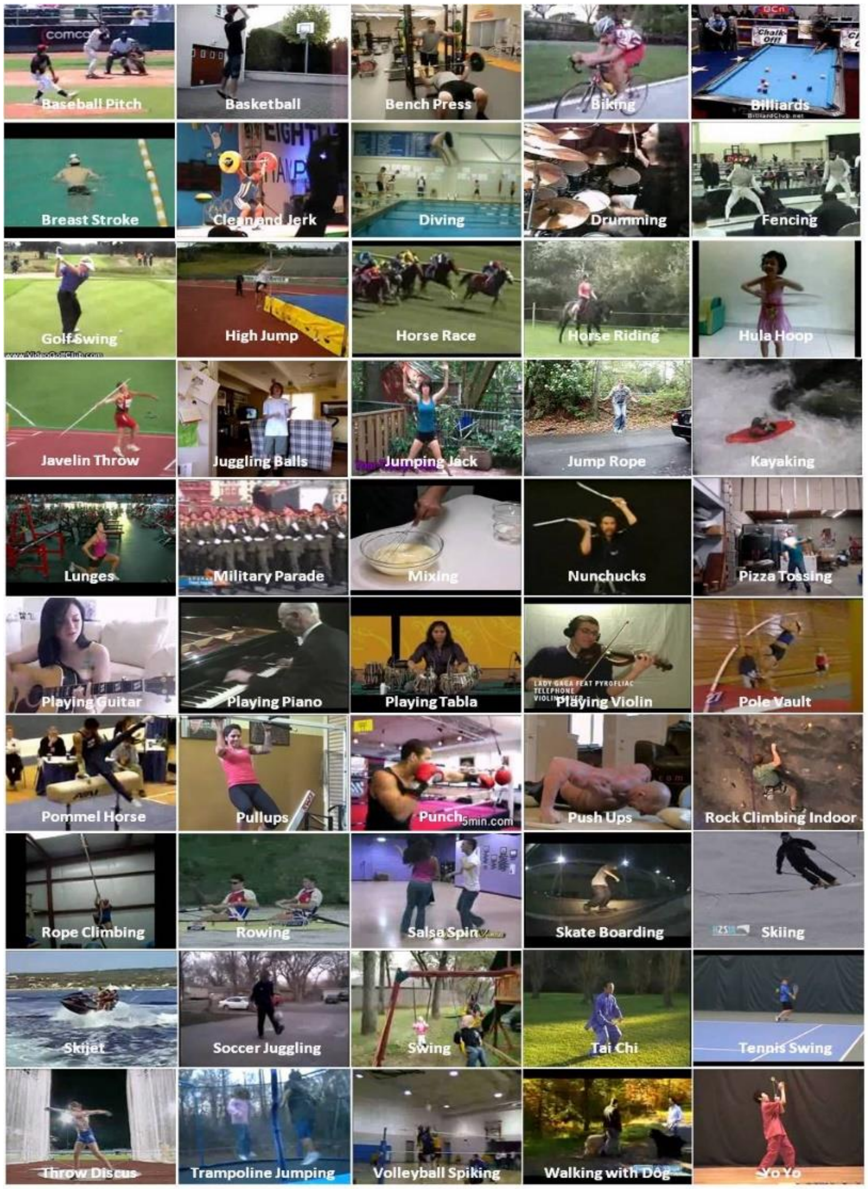


Figure 12 UCF50 dataset

3.2 Framework

The framework will be used as a tool to evaluate or compare different forms of techniques used for video retrieval. It will be feasible and scalable and can also be integrated with new algorithms.

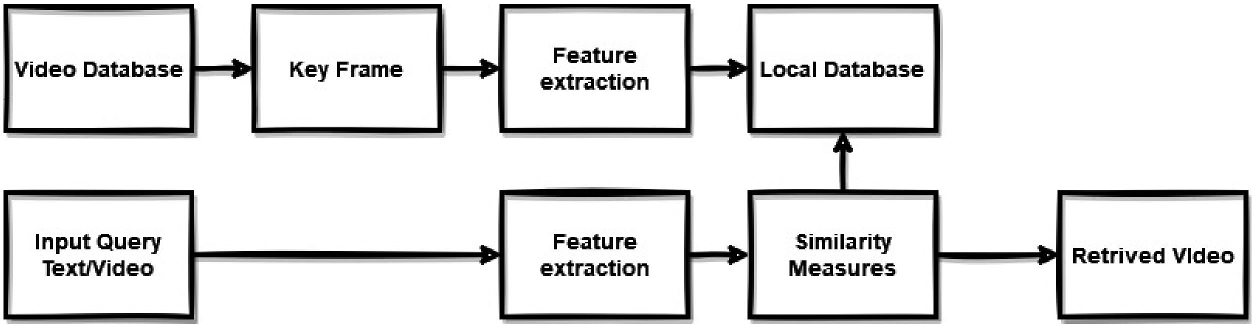


Figure 13 Proposed Framework

First extract Key frames from the video. To do this, the video is broken down into frames. Retrieve only the unique frames i.e., Key frames. Numerous Feature extraction algorithms are run on these key frames like LDP, LTP, LDN, LDSN, VGG and ResNet. All the data is stored in a local database.

When the input query is passed, it goes through feature extraction phase. This data is used to find the similar video from the local database using similarity measures. The similarity measures included are Euclidian distance, Manhattan distance, Chi-Square, and Bhattacharyya. The similarity measure techniques check the similar content image and retrieve the associated video from the database.

3.3 Evaluation Methods

The main evaluation criteria to determine the performance of our proposed system will be the precision and recall. This evaluation methods is used to check the performance of the system with different feature extraction algorithms and similarity measures.

A confusion matrix is a summary of predictions. Below is a representation of a confusion matrix

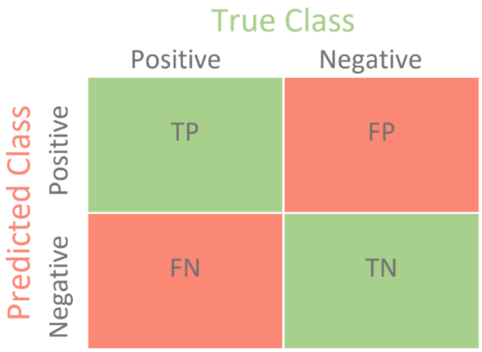


Figure 14 Confusion Matrix

Where ,

- True Positive (TP) – Number of relevant videos that the algorithm correctly retrieved
- True Negative (TN) – Number of irrelevant videos that the algorithm not retrieved
- False Positive (FP) – Number of irrelevant videos that the algorithm wrongly retrieved
- False Negative (FN) – Number of relevant videos that the algorithm not retrieved

3.3.1 Precision

Precision states the percentage of correctly retrieved videos out of all retrieved videos.

$$Precision = \frac{TP}{TP + FP}$$

Precision = Total number of videos retrieved that are relevant/Total number of videos that are retrieved.

The precision value lies between 0 and 1.

3.3.2 Recall

Recall calculates the percentage of correctly retrieved videos out of all similar video.

$$Recall = \frac{TP}{TP + FN}$$

Recall = Total number of videos retrieved that are relevant/Total number of relevant videos in the database.

4.Requirement Analysis

This is the early phase of a software development lifecycle. Requirement analysis also called requirement engineering is one of the lifecycle phases of a software product lifecycle. It is important to spend good time on requirements gathering and identifying any gaps in this stage. It gives a clear idea on user needs and requirements.

Requirements for this project can be classified into one of the below

- **Business requirements**: High level specification of the User requirements
- **Functional requirements**: Mandatory System requirements
- **Non-functional requirements**: Operational capabilities
- **Hardware requirements**: Hardware configurations needed by the system

4.1 Business Requirements

- System must retrieve the most similar video: Out of the whole local data base the system should be able to retrieve the correct or similar video
- System must retrieve video efficiently in a fleeting moment: The execution time to retrieve the video should be minimum

4.2 Functional Requirements

- UCF50 data in the database: The UCF50 is the dataset used in this research for “An End-to-End Framework for Content Based Video Retrieval Using Image Query” . The UCF50 an action recognition data set with 50 action categories, consisting of realistic videos.
- Video Pre processing : The video is first processed into frames. The Key frame is taken from these frames using key frame extraction. This helps in reducing redundancy in the data.
- Different Feature extraction methods: Different feature extraction methods are performed . The feature extraction methods used are LBP (Local Binary Pattern), LTP (Local Ternary Pattern), LDN (Local Directional Number),LDSN(Local directional-structural pattern),HOG (Histogram of oriented gradients), VGG and ResNet. Feature Extraction Algorithms are used to reduce the dimensionality the key frames.
- Different Similarity measures: Different Similarity methods are used. The similarity measures included are Euclidian distance, Manhattan distance, Chi-Square, and Bhattacharyya. Similarity measures are used to measure how similar the two objects are.
- Check the accuracy: Evaluation metrics like precision and recall. These evaluation metrics represents the performance of the system.

- Select the best features for the system. Compare different feature extraction algorithm with one another and also each similarity measure with one another using evaluation metrics
- Computational cost: The execution time for performing the retrieval should be minimum

4.3 Non Functional Requirements

- Data Security: There should not be any data leak. The system must be completely secure. It should follow the information security policy of the municipal authorities (UAE).
- Disk space: There should be enough space in the disk to store all the data and also store the video processed data. It should not have any problems in adding new data to the database.
- High Availability: System should be up and running all the time. There shouldn't be any system failure while retrieving.
- Performance: System should have high performance. There are many previous systems developed to retrieve video using image. The proposed system is a comparative study and will try achieving higher performance than previously developed retrieval system.
- Usability: Easy to configure new feature extraction algorithms or new similarity measures.

4.4 Hardware Requirements

OS: Windows 11 or Windows 10

RAM : Min 8GB

Storage : Min 512 GB HHD or SSD

Processor: i5(INTEL) and above or r5(RYZEN) and above

5.Professional, Legal and Ethical Issues

5.1 Professional Issues

- All the referred papers, code, or libraries will be cited and be used under the terms of the publisher licenses.
- The final code is written and tested to follow British Computing Society codes (BSC) of conduct. Furthermore, it will be commented throughout for clarity with detailed documentation.
- Any code used from authentic third-party sources will be used and cited properly only if permitted by their license
- The dataset will be used under Data Protection Law Policy

5.2 Legal Issues

- Strictly complies with the data security policies of the local government (United Arab Emirates)
- The system will not indulge in privacy breaches or any unauthorized access or any security attacks on any system or make other system to do so.
- No personal information with respect to users sensitive information will be stored.
- No patents will be breached by the application
- No data will be republished.

5.3 Ethical Issues

This is a research-based project that does not involve sensitive datasets or users. All datasets are publicly available in the ***UCF center for Research in Computer Vision*** repository and usually used for benchmarking. Thus, there is no risk of violating any ethics code.

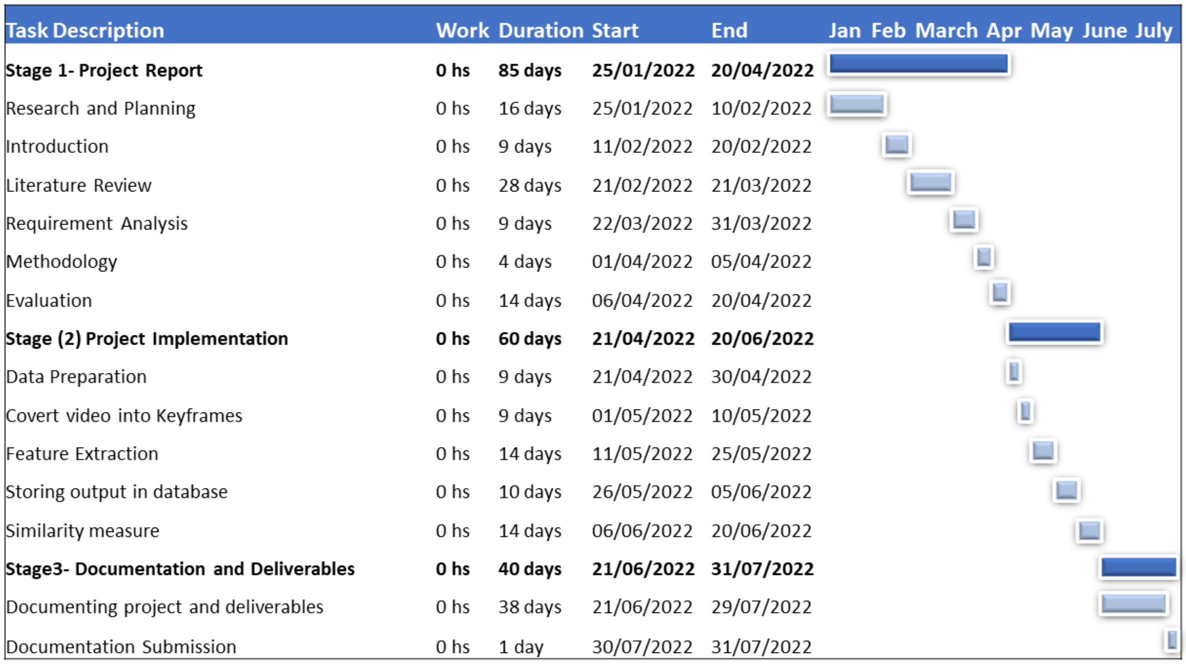
5.4 Social Issue

- As some videos are from YouTube, Respect to UCF and YouTube policies on interacting with the user's content
- The Application will not republish the UCF-50 data.
- The videos will not be posted anywhere and will not post on YouTube on users' behalf

6.Project Plan

A project plan helps in communicating the project status to stake holders, different teams, and organization etc. This chapter discusses the project plan with Gantt chart and risk assessment.

6.1 Ghantt chart



Task Description	Work	Duration	Start	End	Jan	Feb	March	Apr	May	June	July
Stage 1- Project Report	**0 hs**	**85 days**	**25/01/2022**	**20/04/2022**							
Research and Planning	0 hs	16 days	25/01/2022	10/02/2022							
Introduction	0 hs	9 days	11/02/2022	20/02/2022							
Literature Review	0 hs	28 days	21/02/2022	21/03/2022							
Requirement Analysis	0 hs	9 days	22/03/2022	31/03/2022							
Methodology	0 hs	4 days	01/04/2022	05/04/2022							
Evaluation	0 hs	14 days	06/04/2022	20/04/2022							
Stage (2) Project Implementation	**0 hs**	**60 days**	**21/04/2022**	**20/06/2022**							
Data Preparation	0 hs	9 days	21/04/2022	30/04/2022							
Covert video into Keyframes	0 hs	9 days	01/05/2022	10/05/2022							
Feature Extraction	0 hs	14 days	11/05/2022	25/05/2022							
Storing output in database	0 hs	10 days	26/05/2022	05/06/2022							
Similarity measure	0 hs	14 days	06/06/2022	20/06/2022							
Stage3- Documentation and Deliverables	**0 hs**	**40 days**	**21/06/2022**	**31/07/2022**							
Documenting project and deliverables	0 hs	38 days	21/06/2022	29/07/2022							
Documentation Submission	0 hs	1 day	30/07/2022	31/07/2022							

Figure 15 Gantt Chart

6.2 Risk Management

Risk Management is a necessary task to monitor the risks continuously. It is used to mitigate as per the plan and notify the stakeholders accordingly. Below table lists the foreseen risks with the mitigation actions.

Risk	Probability	Impact	Mitigating Action
Author’s medical emergency	Low	Very High	Discuss and ask for extension with supervisor
System failures	Low	Very High	continuously backup the data and Use SCM(Source Control Management) tools
Supervisor’s medical emergency leave	Low	Very High	reschedule the project or Check with university or alternative
Lack of data	Low	High	Can generate more data or take relevant data from the UCI repository
System Incompatibility	Medium	Medium	Check for system compatibility or use tools which the current system supports and avoid such a situation

7. References

[1] T. Yoshida, G. Irie, H. Arai, and Y. Taniguchi, “Towards semantic and affective content-based video recommendation,” in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2013, pp. 1–6.

[2] M. Wang, Y. Ming, Q. Liu and J. Yin, "Image-Based Video Retrieval Using Deep Feature," 2017 IEEE International Conference on Smart Computing (SMARTCOMP), 2017, pp. 1-6, doi: 10.1109/SMARTCOMP.2017.7947017.

[3] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, “Trademark matching and retrieval in sports video databases,” in *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, ser. MIR ’07. New York, NY, USA: ACM, 2007, pp. 79–86. [Online]. Available: http://doi.acm.org/10.1145/1290082.1290096

[4] T. Yoshida, G. Irie, H. Arai, and Y. Taniguchi, “Towards semantic and affective content-based video recommendation,” in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2013, pp. 1–6.

[5] X. Wu, A. G. Hauptmann, and C.-W. Ngo, “Practical elimination of near-duplicates from web video search,” in *Proceedings of the 15th ACM International Conference on Multimedia*, ser. MM ’07. New York, NY, USA: ACM, 2007, pp. 218–227. [Online]. Available: http://doi.acm.org/10.1145/1291233.1291280

[6] Ojala, T., Pietikainen, M. and Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), pp.971-987.

[7] V. Soni, S. K. Mathariya and R. Soni, "User friendly approach for video search technique using text and image as query," 2014 Conference on IT in Business, Industry and Government (CSIBIG), 2014, pp. 1-12, doi: 10.1109/CSIBIG.2014.7056999.

[8] K. Fushikida, Y. Hiwatari and H. Waki, "A content-based video query agent using feature-based image search engine," Proceedings Third International Conference on Computational Intelligence and Multimedia Applications. ICCIMA'99 (Cat. No.PR00300), 1999, pp. 181-185, doi: 10.1109/ICCIMA.1999.798525.

[9] A. Araujo and B. Girod, "Large-Scale Video Retrieval Using Image Queries," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 6, pp. 1406-1420, June 2018, doi: 10.1109/TCSVT.2017.2667710.

[10] R. M. N. Sadat, A. Rakib, M. M. Salehin and N. Afrin, "Efficient design of Local Binary Pattern for image retrieval," 2011 IEEE Symposium on Computers & Informatics, 2011, pp. 510-514, doi: 10.1109/ISCI.2011.5958968.

[11] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions," in IEEE Transactions on Image Processing, vol. 19, no. 6, pp. 1635-1650, June 2010, doi: 10.1109/TIP.2010.2042645.

[12] C. K. Tran, T. F. Lee, L. Chang and P. J. Chao, "Face Description with Local Binary Patterns and Local Ternary Patterns: Improving Face Recognition Performance Using Similarity Feature-Based Selection and Classification Algorithm," 2014 International Symposium on Computer, Consumer and Control, 2014, pp. 520-524, doi: 10.1109/IS3C.2014.141.

[13] J. Majumdhar and S. K. Nayak, "A Novel Method on Summarization of Video Using Local Ternary Pattern and Local Phase Quantization," 2021 2nd International Conference on Range Technology (ICORT), 2021, pp. 1-6, doi: 10.1109/ICORT52730.2021.9581941.

[14] G. Peng, W. Tongming and W. Weina, "Application of the Image Retrieval Technique on the Education Resources Image Database," 2009 Second International Symposium on Computational Intelligence and Design, 2009, pp. 152-154, doi: 10.1109/ISCID.2009.45.

[15] Sagar Chavda, & Mahesh Goyani. (2019). Content-Based Image Retrieval: The State of the Art. *International Journal of Next-Generation Computing*, *10*(3), 193–212. https://doi.org/10.47164/ijngc.v10i3.166

[16] W. Li, L. Duan, D. Xu and I. W. -H. Tsang, "Text-based image retrieval using progressive multi-instance learning," 2011 International Conference on Computer Vision, 2011, pp. 2049-2055, doi: 10.1109/ICCV.2011.6126478.

[17] Dinakaran, Booshnam et al. “Interactive Image Retrieval Using Text and Image Content.” (2010).

[18] L. Wu, R. Jin and A. K. Jain, "Tag Completion for Image Retrieval," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 3, pp. 716-727, March 2013, doi: 10.1109/TPAMI.2012.124.

[19] R. Yanagi, R. Togo, T. Ogawa and M. Haseyama, "Image Retrieval With Lingual And Visual Paraphrasing Via Generative Models," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 2431-2435, doi: 10.1109/ICIP40778.2020.9190966.

[20] Mishra, Anand, et al. “Image Retrieval Using Textual Cues.” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013.

[21] Zhang, Dengsheng & Lu, Guojun & Ma, Wei-Ying. (2007). A Survey of Content-based Image Retrieval With High-level Semantics. Pattern Recognition. 40. 262-282. 10.1016/j.patcog.2006.04.045.

[22] Danish, M., Rawat, R., & Sharma, R. (2013). A Survey: Content Based Image Retrieval Based On Color, Texture, Shape & Neuro Fuzzy. Int. Journal Of Engineering Research And Application, 3(5), 839–844.

[23] R., S. P., & P.V.S.S.R., C. M. (2016). Dimensionality reduced local directional pattern (DR-LDP) for face recognition. Expert Systems with Applications, 63, 66–73. https://doi.org/https://doi.org/10.1016/j.eswa.2016.06.031

[24] T. Jabid, M. H. Kabir and O. Chae, "Local Directional Pattern (LDP) – A Robust Image Descriptor for Object Recognition," 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 482-487, doi: 10.1109/AVSS.2010.17.

[25] S. Arivazhagan, R. Ahila Priyadharshini and S. Sowmiya, "Facial expression recognition based on local directional number pattern and ANFIS classifier," 2014 International Conference on Communication and Network Technologies, 2014, pp. 62-67, doi: 10.1109/CNT.2014.7062726.

[26] A. Ramirez Rivera, J. Rojas Castillo and O. Oksam Chae, "Local Directional Number Pattern for Face Analysis: Face and Expression Recognition," in IEEE Transactions on Image Processing, vol. 22, no. 5, pp. 1740-1752, May 2013, doi: 10.1109/TIP.2012.2235848.

[27] Makhmudkhujaev, Farkhod & Iqbal, Md Tauhid & RYU, Byungyong & CHAE, Oksam. (2019). Local directional-structural pattern for person-independent facial expression recognition. TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES. 27. 516-531. 10.3906/elk-1804-58.

[28] T. Surasak, I. Takahiro, C. -h. Cheng, C. -e. Wang and P. -y. Sheng, "Histogram of oriented gradients for human detection in video," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 172-176, doi: 10.1109/ICBIR.2018.8391187.

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[30] W. Zhou, S. Gao, L. Zhang and X. Lou, "Histogram of Oriented Gradients Feature Extraction From Raw Bayer Pattern Images," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, no. 5, pp. 946-950, May 2020, doi: 10.1109/TCSII.2020.2980557.

[31] S. Guzmán, A. Gómez, G. Diez and D. S. Fernández, "Car detection methodology in outdoor environment based on histogram of oriented gradient (HOG) and support vector machine (SVM)," 6th Latin-American Conference on Networked and Electronic Media (LACNEM 2015), 2015, pp. 1-4, doi: 10.1049/ic.2015.0310.

[32] P. E. Rybski, D. Huber, D. D. Morris and R. Hoffman, "Visual classification of coarse vehicle orientation using Histogram of Oriented Gradients features," 2010 IEEE Intelligent Vehicles Symposium, 2010, pp. 921-928, doi: 10.1109/IVS.2010.5547996.

[33] Simonyan, Karen, and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” *arXiv preprint arXiv:1409.1556* (2014).

[34] Kishore K. Reddy, and Mubarak Shah, Recognizing 50 Human Action Categories of Web Videos, Machine Vision and Applications Journal (MVAP), September, 2012.

[35] Esmaili, I., Dabanloo, N. J., & Vali, M. (2016). Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools. Biomedical Signal Processing and Control, 23, 104–114. https://doi.org/https://doi.org/10.1016/j.bspc.2015.08.006

[36] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[37] W. Wattanarachothai and K. Patanukhom, "Key frame extraction for text based video retrieval using Maximally Stable Extremal Regions," 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom), 2015, pp. 29-37, doi: 10.4108/icst.iniscom.2015.258410.

[38] Bekhet, Saddam & Ahmed, Amr. (2020). Evaluation of Similarity Measures for Video Retrieval. Multimedia Tools and Applications. 79. 10.1007/s11042-019-08539-4.

[39] POORNIMA, N., SALEENA, B. An automated approach to retrieve lecture videos using context based semantic features and deep learning. *Sādhanā* **45,** 254 (2020). https://doi.org/10.1007/s12046-020-01494-z

[40] Bekhet, S., Ahmed, A. Evaluation of similarity measures for video retrieval. *Multimed Tools Appl* **79,** 6265–6278 (2020). https://doi.org/10.1007/s11042-019-08539-4

[41] Grangier, David & Bengio, Samy. (2006). A Neural Network to Retrieve Images from Text Queries. 4132. 24-34. 10.1007/11840930_3.

[42] Patil, Sanjay & Patil, Sanjay. (2010). Content Based Image Retrieval Using Various Distance Metrics. 154-161. 10.1007/978-3-642-27872-3_23.

[43] A. Khare, B. R. Mounika and B. Vasu, "On Retrieval of Nearly Identical Video Clips with Query Frame," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019, pp. 116-121, doi: 10.1109/ICACTM.2019.8776735.

[44] P. Chivadshetti, K. Sadafale and K. Thakare, "Content based video retrieval using integrated feature extraction and personalization of results," 2015 International Conference on Information Processing (ICIP), 2015, pp. 170-175, doi: 10.1109/INFOP.2015.7489372.

[45] Cai, Y., Yang, L., Ping, W., Wang, F., Mei, T., Hua, X.-S., & Li, S. (2011). Million-Scale near-Duplicate Video Retrieval System. Proceedings of the 19th ACM International Conference on Multimedia, 837–838. https://doi.org/10.1145/2072298.2072484