ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 12 N. 1 (2023), e28824
eISSN: 2255-2863
DOI: https://doi.org/10.14201/adcaij.28824

An Efficient Video Frames Retrieval System Using Speeded Up Robust Features Based Bag of Visual Words

Altaf Hussain

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
Correspondence: altafkfm74@gmail.com

KEYWORDS

CBVR, CBIR; K-Mean, SURF; Color Histogram; Accuracy; Loss.

ABSTRACT

Most studies in content-based image retrieval (CBIR) systems use database images of multiple classes. There is a lack of an automatic video frame retrieval system based on the query image. Low-level features i.e., the shape and colors of most of the objects are almost the same e.g., the sun and an orange are both round and red in color. Features such as speeded up robust features (SURF) used in most of the content-based video retrieval (CBVR) & CBIR research work are non-invariant features which may affect the overall accuracy of the CBIR system. The use of a simple and weak classifier or matching technique may also affect the accuracy of the CBIR system on high scale. The unavailability of datasets for content-based video frames retrieval is also a research gap to be explored in this paper.

1. Introduction

In the last two decades, evolutional advancement took place in computer, communication and multimedia technologies which led to massive video/image data production and big image databases/repositories (Anzid et al., 2023). The collections of image data are increasing day by day i.e., medical imaging, road traffic videos, surveillance videos etc. To control the growth of such a large amount of data, it is essential to develop new image retrieval systems that run on a very large scale. The main objective of content-based image retrieval (CBIR) & content-based video retrieval (CBVR) is to develop an accurate and efficient system for creating, managing, and querying image databases. CBIR is the mechanism/procedure of indexing images automatically using the information (features)

hidden in the images e.g., shape (Awasthi& Srivastava 2022), texture (Bhagat& Kumar 2023), statistics (Fan et al., 2023) and e.g., object recognition, 3-D reconstruction (Huang et al., 2023), image registration etc. As discussed above, features are the low-level representation of an image, and they are used to compare whether two or more images are similar or not and also calculate the percentage of matching. The similarity between images can be found using supervised machine learning algorithms e.g. Support Vector Machine (Huo et al., 2023), Naïve Bayes (Hussain et al., 2022), K-Nearest Neighbor (KA et al., 2023), Random Forest or just by using a similarity measure based matching technique (e.g. Euclidean Distance, cosine Similarity, Manhattan distance) for calculating the distance transform of the extracted features from the images (Kakizaki et al., 2023).

Robust feature extraction and calculation of similarity between image data is very crucial and time consuming for the retrieval of optimal images from the video or image database (Kavitha et al., 2023). Due to big datasets and complexity in image data, CBIR is still a developing area with multidisciplinary research. The CBIR system can further be divided into two categories (image database and video file), CBIR using image database consist of images of different classes, a query image is passed to the CBIR system and all the similar images from the database are retrieved using the CBIR system. Machine learning is a branch of artificial intelligence (AI) which has the ability to learn from the data and make predication about a future event that is yet to occur. In the current era, many advancements have been made in computer vision system. Bag of visual words based on speeded up robust features (SURF) (Kovač, I., & Marák, 2023) is a robust method for video summarization, CBIR and bio-metric systems. SURF features, according to Megala et al. (2023), are invariant features in term of scale, rotation, translation, and illumination. The objective of our work is to develop a method which extracts similar frames to the query image, so that the retrieved images and query images are almost the same. We have used the popular computer vision algorithm by modifying it with SURF features. The conventional CBIR system works on image datasets and uses non-invariant features. Scale invariant features transform is also a popular method but the main problem with this method is the extraction of a large number of features from the image which may cause memory issues. The second problem of this method is that it is time consuming. SURF is an alternative method which extracts less features and consumes significantly less time. Once the SURF features are extracted from the video frames, an unsupervised machine learning algorithm is used to quantize the extracted SURF features which are in double format. For quantization, the k-means clustering algorithm has been used. The centroid or mean is called visual word and by the combination of all these visual words a vocabulary is created. The visual words are encoded by a histogram which calculate the number of occurrences of each word in the vocabulary.

For the similarity measure between the features extracted from the video frames and the features from the query images, a distance function has been used. The distance function has calculated the distance between the query and video frame features. The distance is between 0-1, image features with shortest distances are considered to be most similar and are retrieved.

1.1. Content-Based Image Retrieval Techniques

The purpose of this work is to use the popular SURF feature extraction method in computer vision for the development of a novel framework for the retrieval of frames from a video. Different feature detector and descriptor are tested for the recognition of content-based features in a video. The conventional methods use scale-invariant feature transform (SIFT) for interest point selection and feature extraction from the point of interest. In this work, the SIFT features are replaced by SURF features for interest point selection and feature extraction. Unsupervised and supervised classification methods

are compared to show the difference in performance. The use of similarity measure between the image database features and the query image features make it efficient and accurate. In general, CBIR system can be sub-divided into 3 parts that are given below:

- Input image database and query image are both images and the output is/are also images.
- Input video database and query video are both videos and the output is/are also videos.
- Input single/multiple video files and query are images whereas the outputs are video frames.

1.1.1. Images Retrieval System

In an image retrieval system, a database of images that consists of thousands of images is used. The input and output of the system are both images. Features are extracted from the database images and are stored in a feature vector. For the matching and retrieval of similar images, a query image is used. The features from the query image are extracted and, using a classifier or similarity measure method, the most similar images from the database are retrieved.

1.1.2. Video Retrieval System

In a video retrieval system, a database of videos that consists of thousands of videos is used. The input and output of the system are both video data. The videos are first converted to video frames. Features are extracted from the video frames and are stored in a feature vector. For the matching and retrieval of similar videos, a query video is used. The features from the query videos are extracted and using a classifier or similarity measure method the most similar videos from the database are retrieved.

1.1.3. Image Based Video Frame Retrieval System

In a video frame retrieval system from a video file, videos and query images are used. The inputs to the system are a video image and a query image, while the output of the system are similar video frames. The video is first converted to video frames. Features are extracted from the video frames and are stored in a feature vector. For the matching and retrieval of similar videos, a query image is used. The features from the query image are extracted and using a classifier or similarity measure method, the most similar video frames from the video are retrieved.

1.2. Cosine Similarity

The cosine similarity between two vectors (or two documents on the vector space) is a measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because not only is the magnitude of each word count of each document taken into consideration, but also the angle between the documents. To build the cosine similarity equation it is necessary to solve the equation of the dot product for the $cos\theta$ as shown in Equations (1) & (2):

$$a \,.\, b = \left|\left| a \right|\right| \left|\left| b \right|\right| cos\theta \tag{1}$$

$$\cos\theta = \frac{a.b}{\|a\|\|b\|} \tag{2}$$

The terms a, b, theta, and angle all denote the variations in the image and video. The image and video quality can be retrieved using these terms with different angles. This is the cosine similarity formula. Cosine similarity generates a metric that determines how related two documents are by looking at the angle instead of the magnitude, such as the examples shown below in Figure 1.

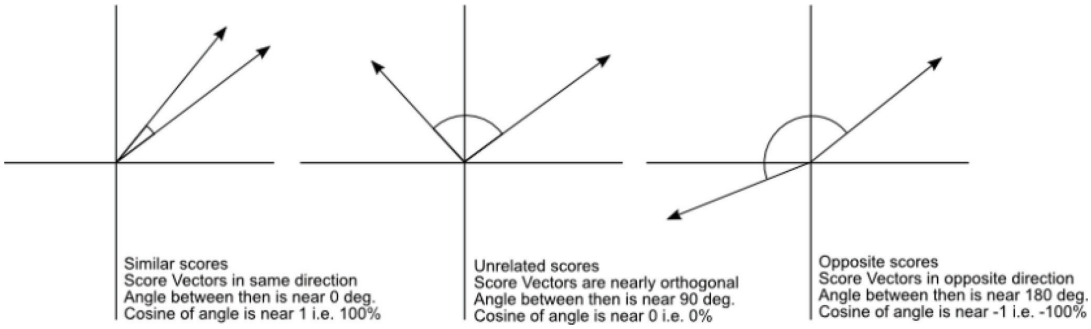

Figure 1. The cosine similarity values for a match between video frame and query image is 1 (same direction), and 0 for no match (opposite directions)

Most of the work done in the CBIR system is done using image databases of different classes. There is a lack of an automatic video frame retrieval system based on the query image. Low-level features i.e., the shape and the colour of most of the objects are almost the same e.g., sun and orange both are round and red in colour. Features used in most of the CBIR research work are non-invariant features which may affect the overall accuracy of the CBIR system. The use of simple and weak classifiers or matching techniques may also affect the accuracy of the CBIR system on a high scale. The unavailability of datasets for the content-based video frame retrieval system is also a research issue. In this paper these issues are investigated.

1.3. Contributions

- We propose a method that removes redundant frames from videos using HOG (histogram of oriented gradients) features and Euclidean distance similarity measure.
- We propose a bag of visual words-based features extractor method using SURF features, which is invariant in term of scale, rotation, translation, and illumination.
- To simulate the proposed model in Matrix Laboratory (MATLAB).

2. Literature Review

In the research of Victoria Priscilla, C., & Rajeshwari (2023), an image retrieval system was developed for the retrieval of bio-medical images that was based on Hausdorff similarity/distance with the combination of texture and wavelet features. In the first phase, image data was enhanced using

nonlinear transformation (Vieira et al., 2023a) & (Vieira et al., 2023b). For the classification/matching Hausdorff distance/similarity is used to retrieve the most similar images from the database. MRI brain and CT lung images were used in the experimental work. The experimental results show a higher accuracy when combining both feature extraction methods instead of using them separately. Walter et al. (2023), Wickstrøm et al. (2023), Usher, (2023) presented novel techniques which increase precision in the CBIR system and are based on electromagnetism optimization method. The electromagnetism optimization technique considers each single image/frame and electrical charge. The technique has two major phases: fitness function measurement and electromagnetism optimization. This technique is implemented on 8,000 images consisting of 80 classes. The experimental work on CBIR shows improvement in term of precision. Veselý, P., & Peška (2023) presented a new technique for the retrieval of similar images. This CBIR system uses encrypted cloud data. Firstly, a binary clustering technique is an integrated technique for classification. The hue, saturation, validation (HSV) histogram and directional connectivity terminology (DCT) histogram are used to extract features. The distance/matching of image database features and query image features is classified using the binary clustering technique. The analysis and experimental work show higher accuracy. Sowmyayani, S., & Rani (2023) proposed a novel image representation for CBIR. The core idea of the proposed method was to do deep learning for the local features of an image and to melt semantic component into the representation through a hierarchical architecture which was built to simulate the human visual perception system, and then a new image descriptor of features conduction neural response (FCNR) was constructed. Compared with the classical neural response (NR), FCNR has lower computational complexity and is more suitable for CBIR tasks. The results of experiments on a commonly used image database demonstrate that, compared with the performance of NR methods and other image descriptors that were originally developed for CBIR, the proposed method has superior performance in terms of retrieval efficiency and effectiveness. Salih, F. A. A., & Abdulla (2023) & Salih, S. F., & Abdulla (2023) proposed different feature extraction methods, tested with the k-nearest neighbor (K-NN) classifier. The aim of the proposed work was to choose the most appropriate feature descriptor method. Color features based on RGB & HSL (hue, saturation, lightness), shape/geometrical and texture features were extracted from a large image database and then using a supervised machine learning technique i.e., K-NN classifier, the images with greatest similarity between the database image and host image were retrieved with high accuracy. Sikandar et al. (2023) investigated a parallel processing based novel graphics processing unit (GPU) adaptive index structure. A comparative study of the central processing unit (CPU) and the graphical processing unit (GPU) based on experiments, was conducted. The GPU-based brute force approach demonstrated that the proposed approach could achieve high speed up with little information loss. Datasets used in the proposed work were Corel 10K and graphical imaging sequence terminology (GIST) 1. The experimental work showed that for the CBIR system, the GPU based retrieval system is more efficient in terms of time. Prathiba et al. (2023) CBIR is nowadays one of the most promising solutions for the management of image databases with high accuracy and efficiency, as well as high performance in image retrieval. In the authors' study, a new optimization technique was proposed involving vocabulary building, image retrieval and matching. More precisely, scale invariant feature transform (SIFT) was extracted. The built similarity indexing method removed the error caused by large-scaled SIFT. The proposed method achieved higher accuracy and efficiency than most of the state-of-the-art methods in the domain of CBIR. Rastega et al. (2023) suggested the two different image descriptor texture and color are used to make a CBIR system. As low block processes (LBP) (texture features) are not invariant features in terms of scale, rotation, and translation, it is important to use other features along the LBP descriptor. To end the robustness issue, color features

are fused with the LPB, and using Euclidean distance similarity measure, the similar images from the database and host image are retrieved with a higher accuracy. The author claims that this is a robust, efficient, and accurate method for the CBIR system. Mounika et al. (2023) proposed the advantages of a content-based image retrieval system, as well as key technologies. In comparison to the shortcoming of the traditional system in which a single feature is used, this paper introduces a method that combines color, texture, and shape for image retrieval. This is a considerable advantage. Then, the authors focused on feature extraction and representation, several commonly used algorithms and image matching methods. Prathiba, T., & Kumari (2023) proposed the content-based image retrieval from large image/video databases. This area of research has received attention in the scientific community for the last two decades. The CBIR system is an important application of computer vision, in this work Ranklet Transform and red, green & blue (RGB) color features were extracted. In the preprocessing phase, Ranklet Transform is used to make the image invariant to rotation. For classification, the k-means unsupervised learning algorithm is used.

2.1. Our Core Contributions in Contrast with the Existing Literature Review

- We propose a novel and hybrid approach for video retrieval.
- We propose a CNN based approch with AlexNet implementation for image retrieval.
- We propose the SURF based bag of visual words technique to identify and calssify image processing.
- We propose a dual image processing approach, namely CBIR & CBVR.

The proposed work is different from the previous approaches.After reviewing the related research work, it can be realized that a lot of work on the topic of content-based image retrival system has been carried out in past and that it is still a intensively researched area by researchers from different background. Various standard techiques were applied for the extraction of robust features from videos frames. Similarly, not only the feature extraction and feature selection (e.g. features dimensionality reduction such as image and video contents) techniques were studied but also a lot work was done on recognition/prediction, which is achieved by using similarity measure technique for the automatic retrival of similar video frames from the video.

3. Research Methodology

Our proposed work consists of two parts i.e., redundant frames removal and similar video frames retrieval. In video processing, removing the redundant frames is a challenging task; most videos have a 30 frame-per-second rate. This frames per second (FPS) rate of a video may be high or low depending on the situation and application. Obtaining 30 frames in one second means that most frames have redundant information. Researchers have proposed methods to pick every 5^{th} frame, picking only the odd or even frames, which is not a professional approach. In our method, we reduced the number of video frames by removing the redundant video frames. For this purpose, we used a feature descriptor and similarity measure score. Features from each single frame are extracted from the video, HOG features work on image edges. Edges in an image are extracted using a high pass filter which enables the high frequency components in an image and stops the low frequency component; the higher frequencies in an image can be found in the region where colour change occurs. These edge angle information is stored and represented on a histogram, the orientation of edge in this work is an important information in the images/frames.

The similarity measure is the measure of how much alike two video frames are. Similarity measure in data mining and computer vision context is a distance with dimensions representing the features of the objects. If this distance is small, then the degree of similarity is high; if the distance is large, then the degree of similarity is small. The similarity is subjective and is highly dependent on the domain and application. For example, two frames are similar because of colors or edge information. Care should be taken when calculating distance across dimensions/features that are unrelated. The relative values of each element must be normalized, or one feature could end up dominating the distance calculation. Similarity is measured in the range 0 to 1 [0, 1].

3.1. Research Flow

In the research flow, we define the steps that have been carried out as part of our research. Figure 2 shows the flow diagram of the conducted research.

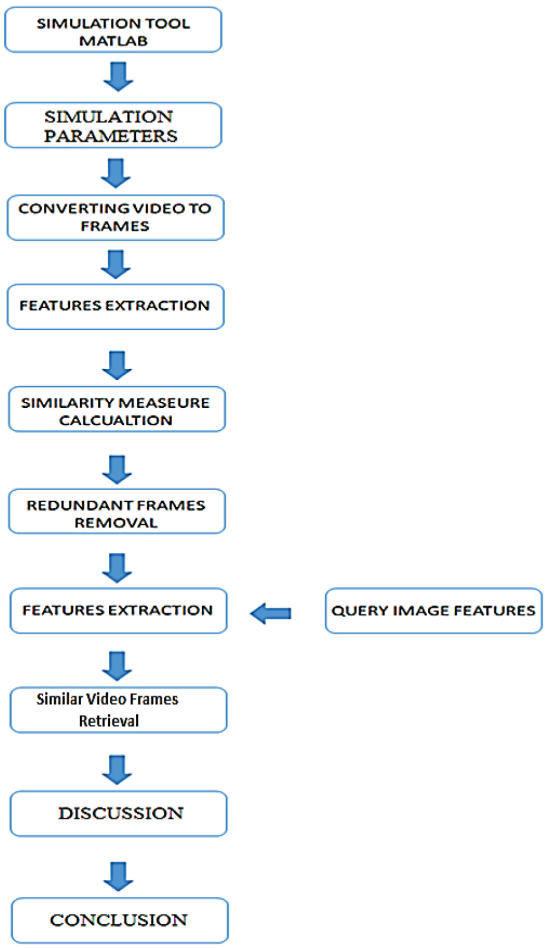

Figure 2. Paper Flow Illustrations

3.2. Simulation Tool

Matrix Laboratory (MATLAB) is a multi-paradigm numerical computing programming language and fourth generation programming language. MATLAB allows for matrix manipulations, plotting of functions and data, and implementation of algorithms. MATLAB is widely used for signal processing, digital image processing, computer vision system, data mining and machine learning algorithm development and simulation. In the development of our algorithm, we have used an image processing toolbox, a machine learning toolbox and a computer vision system toolbox.

3.2.1. Simulation Parameters

The simulation parameters are given in Table 1. These values denote that the overall simulations were carried out by using these parameters and values. Multiple data values are presented here with different values depending on the task required for them.

Table 1. Simulation parameters

Parameter Name	**Value**
Programming Evironment	MATLAB R2015a
Toolboxes	DIP, CV and ML
Datasets	N/A
No. of Clusters	500,1000
Feature Extraction Time	N/A
Matching Training Time	N/A
Average Time of Testing Frame	N/A
Similarity/Distance	Cosine Similarity
No. of Supported Classes	N/A
Experimental Evaluation Parameters	Confussion Matrix

3.2.2. Converting Video into Frames

A video is a sequence of RGB, grayscale or binary images. A video file consist of metadata which contains all the necessary information about the video, such as size of video file, resolution, colour infromation, compression type, date of creation etc. A loop is used to load each frame of the video onto MATLAB and save each frame as an image.

3.2.3. Feature Extraction

Features are numerical values which are computed by applying different mathematical and statistical methods. Images consist of different contents (information) such as EDGE HISTOGRAM DESCRIPTOR, SURF, SIFT etc. Features enable the development of an object recognition system as classifiers do not understand images nor can compare two images directly. Image features are computed and can then be used by the classifiers for learning and decision making.

3.2.4. Similarity Measure

In machine learning, classifier and similarity measure methods are used to assign a class to the testing data. As discussed above, the features of different images are extracted, and a distance function is used to classify the images into similar and non-similar images. Some state-of-the-art similarity measure methods are Euclidean distance, hamming, cosine similarity etc.

3.2.4.1. Main Consideration about similarity

$$\text{Similar} = \text{if distance of X, Y} <= 25 \tag{3}$$

Where, X, Y are two different video frames denotes the similar distance of these.

The Euclidean distance or Euclidean metric is the ordinary straight-line distance between two points in Euclidean space dented as in Equation (3). With this distance, Euclidean space becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as the Pythagorean metric. A generalized term for the Euclidean norm is the L^2 norm or L^2 distance. The given features, such as features 1 & features 2, are illustrated in the equations that represents the constant values for both features and then both features are subtracted from each other and so on.

To calculate the distance between two video frames, HOG features are denoted in Equations (4), (5), (6), (7), & (8) as:

$$\text{Features\_1} = [13,24,35,16,57,32,53,31,25,26,37,13,22] \tag{4}$$

$$\text{Features\_2} = [26,12,17,5,25,16,11,22,36,21,17,15,12] \tag{5}$$

$$V = \text{Features\_1} - \text{Features\_2} \tag{6}$$

$$V\_2 = V*V^T \tag{7}$$

$$\text{Distance} = \sqrt{V\_2} \tag{8}$$

3.2.5. Redundant Frames Removal

In a database system, redundant data is the repetition of data that increase the size of the file. Similarly, a video consists of redundant information in the form of video frames as shown in Figure 3. In a case where there is no change occurring between two or more video frames then we consider them as redundant data. This redundant data increases the size of the feature vector and may cause an overfitting problem in matching and reduce the speed of video frame matching with the provided query image.

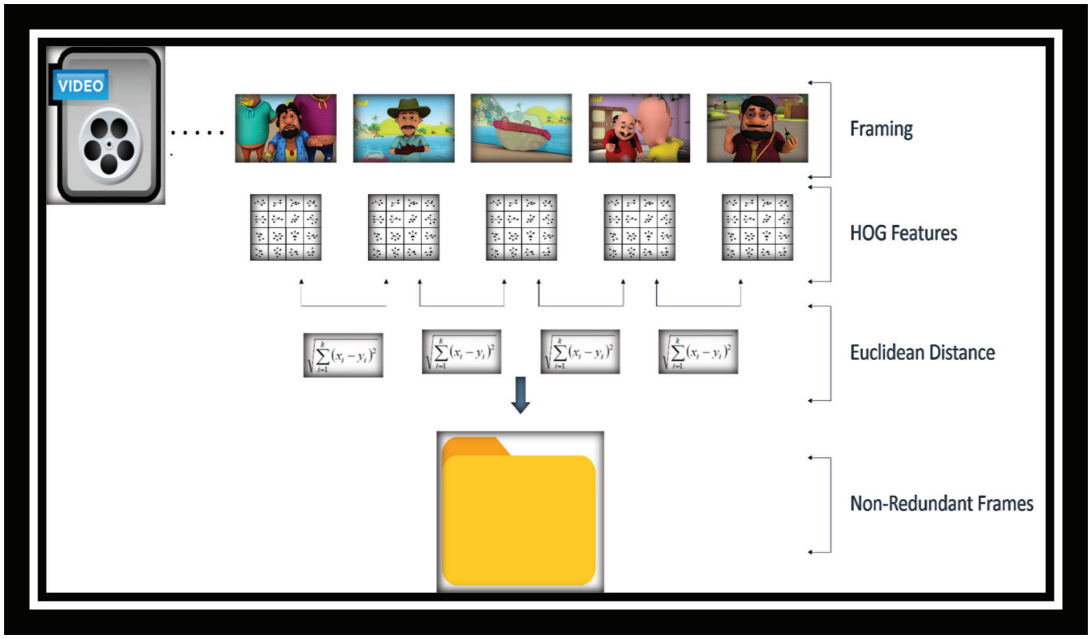

Figure 3. Redundant video frames removal using HOG features and Euclidean similarity measure technique

3.2.6. Similar Video Frames Retrieval

In the CBIR system, robust image descriptor and matching/classification techniques play an important role in the accurate and efficient retrieval of images. Image features are a type of special representation/information that we extract from the image/frame. Images features are numerical values and the vector used to store these features is called the feature vector. The mechanism of the CBIR system is similar to the classification problem in machine learning, but features are not labeled in the CBIR system. Our proposed video frames retrieval system method composed of two major steps i.e., SURF based bag of features for feature extraction and cosine similarity for matching between video frames and query image.

Our proposed approach is summarized into the following steps:

1) SURF based features are extracted from video frames using the bag of features (BoF) extraction technique.
2) These features are quantized using k-means clustering technique for reducing the size of feature vector.
3) An encoding method is used for visual word occurrences, and a histogram of these visual words is constructed.
4) For the matching/retriveal cosine similarity, the distance calculation method is used.

Figure 4 illustrates our proposed CBIR & CBVR on the basis of the SURF model.

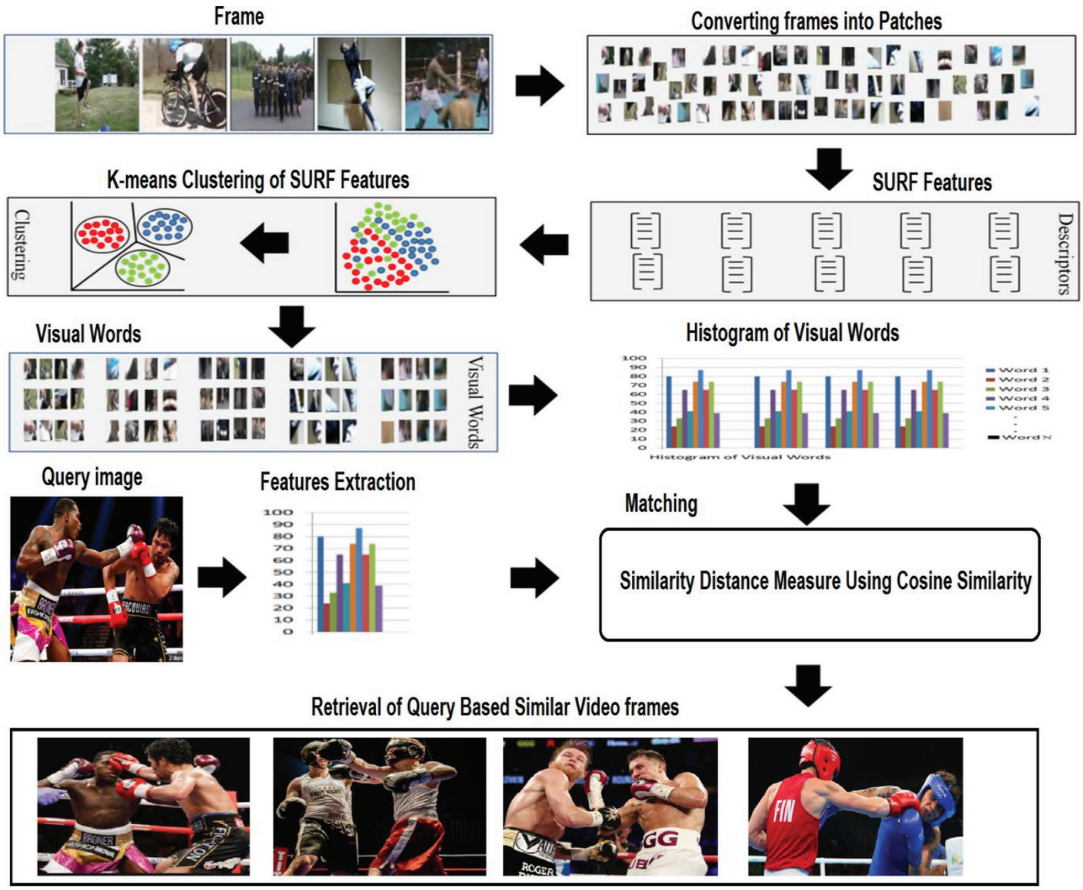

Figure 4. Our proposed framework for the retrieval of similar video frames from a video

4. Results and Discussion

In this section, a detailed discussion of the simulation results is discussed. This section consists of simulation-based results in which each scenario has been explained with proper justification. These have been portrayed in the form of graphs and tables for better and clear understanding. Also, the evaluations have been debated with the other feature extraction models of the video retrieval system.

4.1. Dataset

We are using UCF101; in which data is organized in 101 groups, and 13320 videos from 101 categories, which is as of now the biggest dataset of human activities. It comprises 101 activity classes, this informational index comprises over 13k clasps and 27 hours of video information. The database contains sensible client transferred recordings containing camera movement and a jumbled foundation.

Moreover, this informational index gives gauge activity acknowledgment results on this new dataset utilizing the standard bag of words approach with the general execution of 44.5%. Supposedly, UCF101 is currently the most used testing dataset of activities because of its enormous number of classes, countless clasps, and the unconstrained nature of such clasps.

Figure 5. Visual results of the UCF101 dataset.

Figure 5 shows the 101 actions of the UCF101 dataset in a single frame. The border color of the frame specifies which action belongs to which category. The label on the frame specifies which class the video belongs to, according to the interaction of the human with the videos in which humans can perform different tasks such as running, fighting, playing games etc., or to whom that video belongs.

The average length of the clips for each action is depicted in green.

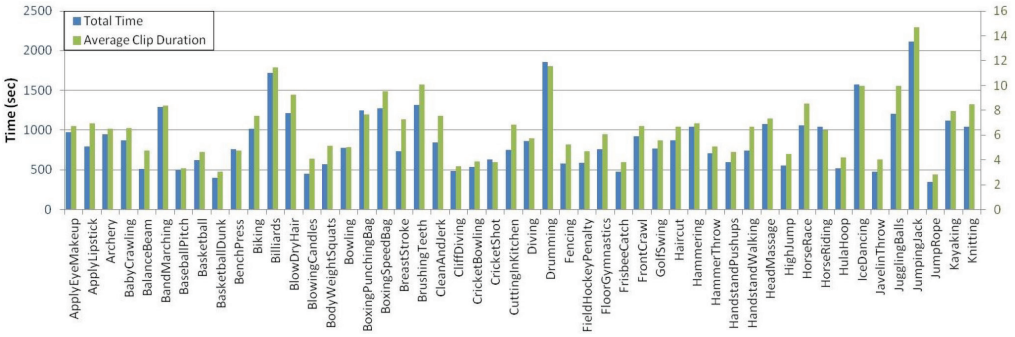

Figure 6. Total time of videos for each class is illustrated using the blue bars

Figure 6 indicates the total time of videos for each class with blue bars and the average length for each class in green color. All clips have the same frame rate and resolution 320*240. And all the files are in the avi format.

The distribution of clip durations is illustrated by the colors.

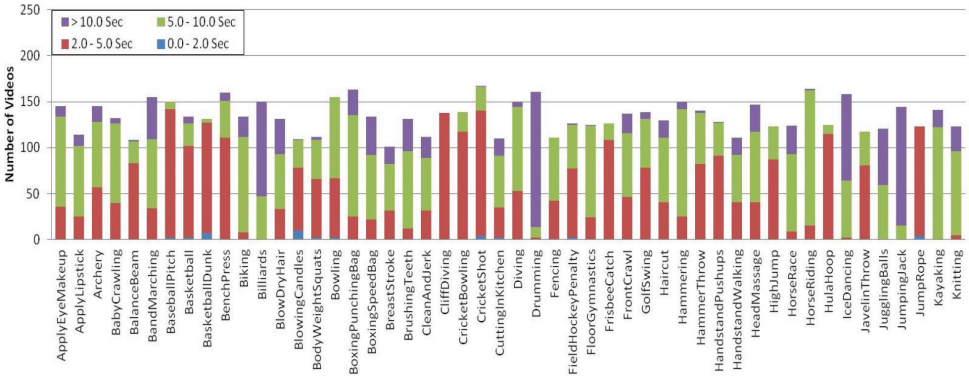

Figure 7. Number of clips per action class

The graph in Figure 7 shows the total number of clips in each class. The clips are distributed in colors. The colors in each bar show the duration of different clips in each class.

4.2. Evaluation for the Proposed System

In the case of classification, only one accuracy classification might not offer insight as it does not give a full classification for a single dataset, instead multiple datasets and objects need to be used. Thus, accuracy or loss are used to summarize algorithm performance. Through the calculation of accuracy and loss, we find out where the system makes errors and which type of error coming into the system. Accuracy and loss are used to check the performance of a classification model on a set of test data for which the true values are known.

4.3. Results on UCF-101 Dataset

The UCF-101 dataset has been used in testing because of the large volume of videos and because of the broad range of perspectives it gathers. This is not merely an observation dataset as the recordings were gathered to assess normal movement acknowledgment errands. Given a query video, the goal is to recover a video of a similar movement independent of the changes in perspective, hues, or surfaces, etc. The recordings in this dataset were taken with shifting foundations, which makes them suitable for the assessment of the viability of our technique. We extricated convolutional neural network (CNN) highlights utilizing AlexNet from the images of the recordings and utilized them to recover the top-positioned video. The goal was to recover as many pertinent recordings as possible. Figure 8 contains the results of the selected question recordings, where the video furthest to the left is the query and the remaining videos are top-positioned videos dependent on the separation between the query and the dataset video. Essentially, in query 2, the best 6 pictures have been recovered accurately. In the third query, applicable pictures were recovered at positions 2, 4, 5, 6, and 8, although there is a tremendous divergence in their perspectives. In the remainder of the queries, pertinent pictures were recovered at top positions which shows the ability of proposal. Despite the fact that the outcomes on this dataset are not solid, it shows that an improved CNN model would significantly improve the outcomes.

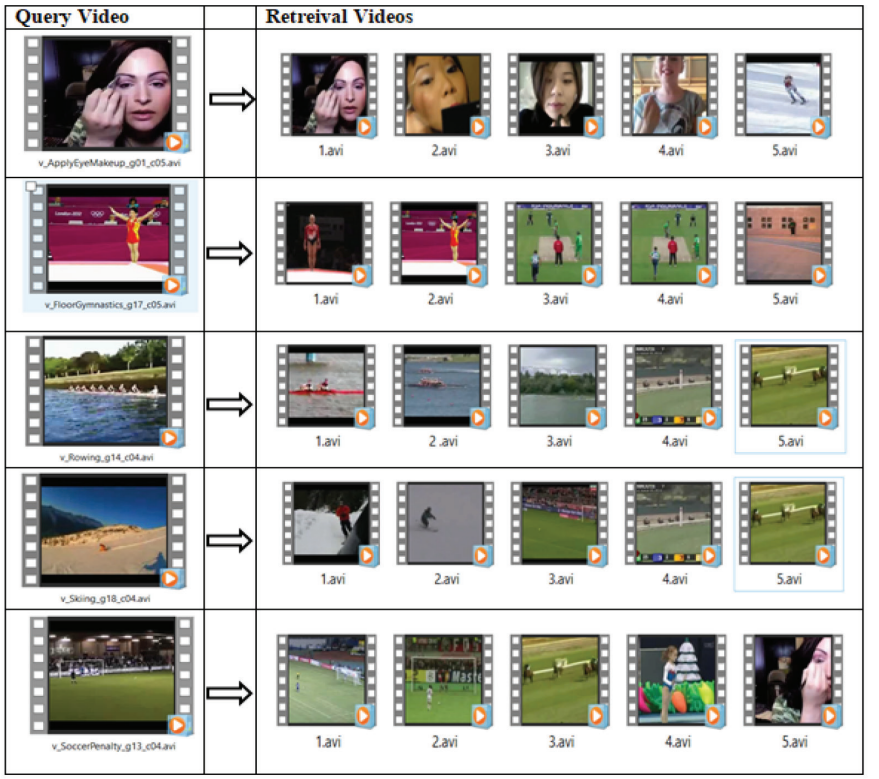

Figure 8. Retrieval result of FC6 on UFC101 dataset

Figure 8 shows the experimental result on dataset UCF101for videos 1, 2, 3, 4, and 5 which illustrates fully connected (FC) layers 1, 2, 3, 4, 5, & 6. It made 6 numbers of clusters for all videos. From the FC6 simulation result, we obtained an accuracy of 44% and a loss of 56%. This layer of the AlexNet model has very poor performance.

Figure 9. Retrieval result of FC7 on UFC101 dataset

Figure 9 shows the experimental results for videos 1, 2, 3, 4, and 5. It made clusters of 6 for all videos. From the FC7 simulation result, we obtained an accuracy of 60% and a loss of 40%. This layer of the AlexNet model has poor performance. Nevertheless, it is 20% better than FC6.

Figure 10 shows the experimental result on videos 1, 2, 3, 4, and 5. It made clusters of 6 for all videos. From the FC8 simulation result, we obtained 90% accuracy and 10% loss. This layer of the AlexNet model gives poor performance. The proposed method test on the UCF101 data set and give very good results on layer FC8.

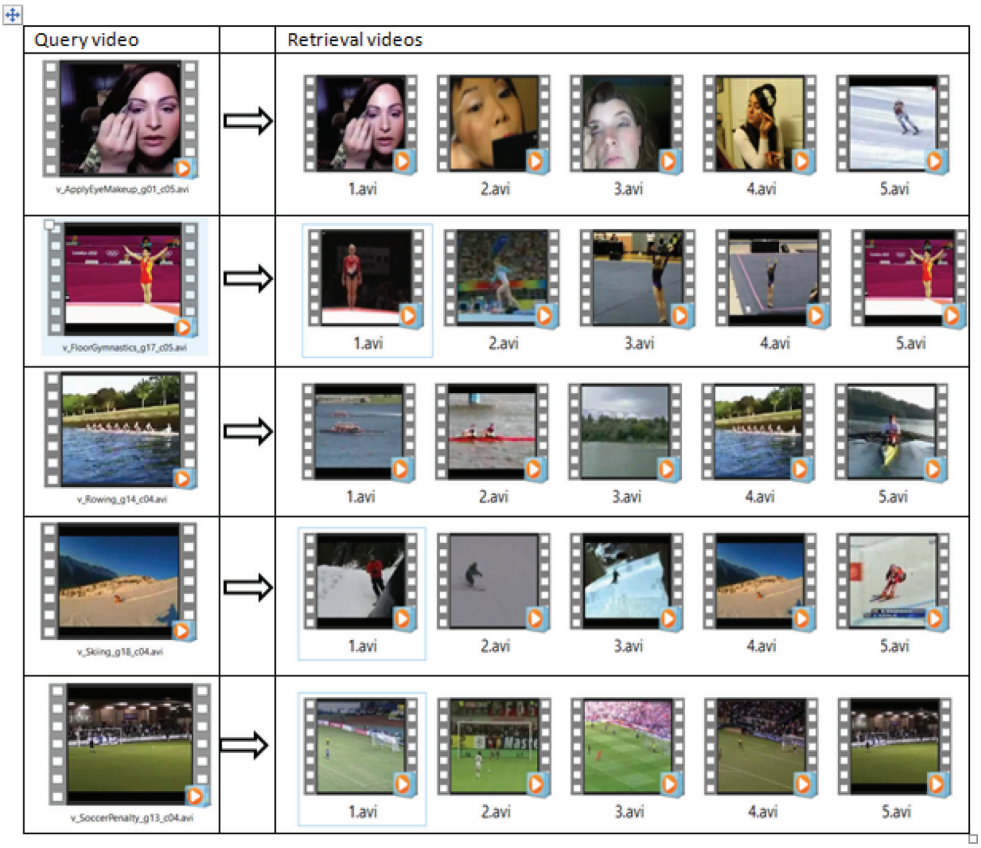

Figure 10. Retrieval result of FC8 on UFC101 Dataset

4.4. Retrieval Performance of Videos

The system performance on matching the input query and the database query is evaluated by subtracting the input query from the database query. The difference between the two videos is found through the Euclidean distance formula. The keyframes are the head of the clusters. These keyframes apply the AlexNet model to extract features. For the performance evaluation, different types of videos are tested. For presentation, a different type of videos is acquired from UCF101 including short movies, military documentaries, music videos, and cartoon videos. The video length is 1-2 minutes. Different keyframes are selected from the videos and the features are saved in the feature vector. The redundant frames are skipped during the reading of the videos. The same videos are searched on the selected keyframes. The experimental result on videos 1, 2, 3, 4, and 5 are divided by the total number of frames 164, 151, 300, 188, and 159, respectively. 6 clusters have been made for all the videos. From the FC6 simulation result we obtained 60% accuracy and 20% loss from video 1, 60% accuracy and 40% loss form video 2, 60 % accuracy and 40% loss form video 3, 40% accuracy and 10% loss form video 4, 40% accuracy and 60% loss form video 5. The experimental result on videos 1, 2, 3, 4, and 5 are divided by the total number of frames 164, 151, 300, 188, and 159, respectively. It made 6 clusters for all the videos. From the FC7 simulation result we obtained 60% accuracy and 40% loss for video

1, 80% accuracy and 20% loss for video 2, 80% accuracy and 20% loss for video 3, 60% accuracy and 40% loss for video 4, 80% accuracy and 20% loss for video 5. The experimental results on videos 1, 2, 3, 4, and 5 are divided by the total number of frames 164, 151, 300, 188, and 159, respectively. It made 6 clusters for each video. From the FC8 simulation results, we have obtained 80% accuracy and 20% loss for video 1. Similarly, 90% accuracy and 10% loss for video 2. 90% accuracy, and 10% loss for video 3. 90% accuracy, and 10% loss for video 4. And 90% accuracy and 10% loss for video 5.

Table 2 illustrates the accuracy of all the fully connected layers in terms of comparison.

Table 2. Accuracy Comparison of FC8 with FC7 and FC6

Model Name	Video_1 Accuracy in %age	Video_2 Accuracy in %age	Video_3 Accuracy in %age	Video_4 Accuracy in %age	Video_5 Accuracy in %age	Average Accuracy
FC6	40	40	60	40	40	44%
FC7	60	80	80	60	80	72%
FC8	80	90	90	90	90	90%

Note: The value of 44% in Table 2 is the model accuracy which is FC6 and 72% is the accuracy of FC7, whereas 90% is the average accuracy of the proposed model, called FC8. In this sense, the FC8 is the proposed model which has given the highest accuracy as compared to FC6 and FC7.

The results of the proposed method on dataset UCF101 are shown in Figure 11. The same videos were retrieved despite having different colors and the same structure. Different videos were also collected from the database. This is due to the nature of the videos which are similar to the input query.

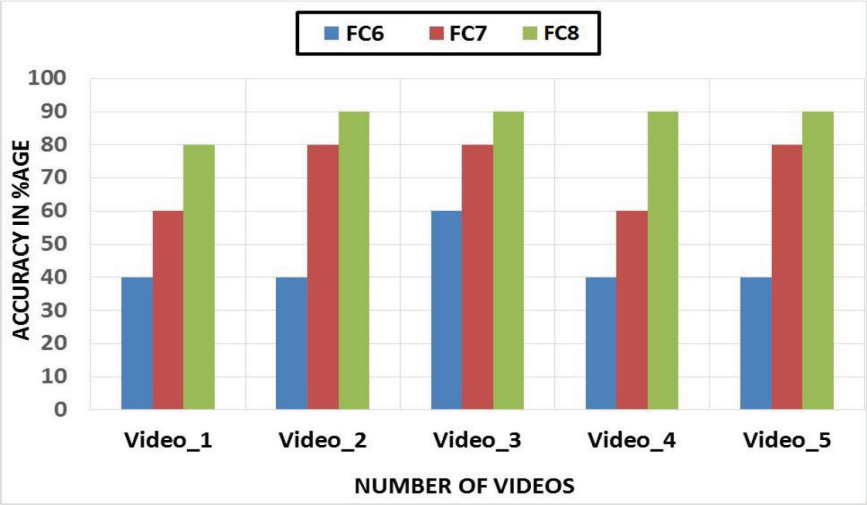

Figure 11. Comparison of FC8 (the proposed model) with FC7 and FC6

Note: Figure 11 illustrates, in percentages, the overall accuracy of the proposed FC8 model in comparison with the FC6 and FC7 models. The accuracy has been measured for 5 videos different videos and for each one proposed FC8 model has given highest accuracy.

In Figure 11, the results of all three layers are shown in the term of accuracy. We have evaluated the precession and recall. The measured accuracy was 40% accuracy on FC6, and 60% on FC7, and 90% on FC8. The accuracy graph in Figure 11, shows that the retrieval performance of our proposed FC8 method is better than the other state-of-the-art methods. Hence, it is clear from these results that the proposed algorithm outclasses existing techniques by a significant margin.

4.5. Loss of Model Layers in Terms of MSE

Mean square error (MSE) is a metric that finds the difference between the predicted values from the observed values between 0 and 1 in the analysis. The symbol of sigma in mathematics, the character that looks like E is called summation. That is the summation of all values, from start i=1 till n. For each point, y are the predicted points for the i^{th} observations, and y' are the correct observed values for the i^{th} observation. We subtract the correct observed values from the y predicted values and calculate the square of the result. Finally, to obtain the mean square error, the summation of all the $(y-y')^2$ values is divided by n, where n is the total number of videos in the database. Table 3 shows the results of the mean square error for layers FC6, FC7, and FC8.

Table 3. MSE of Layers FC6, FC7, and FC8

Model Layer	MSE
FC6	0.7
FC7	0.6
FC8	0.2

Note: The value of MSE is the error in which the lowest error value denotes that the model has performed well. In Table 3, the values of MSE are illustrated in which the FC6 has 0.7, FC7 has 0.6 and FC8 has the MSE of 0.2. This means that the lowest error ratio is shown by FC8 which is better as compared to other FC6 and FC7.

A mean square error close to 1 indicates a high level of error in the model while 0 indicates a low level of error. In the given table, FC6 has an error of 0.7, which means high error. FC7 has an error of 0.6, which is one point low than FC6. Finally, FC8 has an error of 0.2, which is very low error. Among the state-of-the-art techniques, the FC8 results achieve good performance.

4.6. Precision and Recall

Our system has been evaluated on two metrics precision and recall, which are very popular evaluation methods. They are used for the content-based video retrieval system. Precision P is calculated for the number of multiple elements that are extracted from the given datasets and also the number of elements that are inserted to the datasets; it calculates the accuracy of the retrieval system. Recall R is calculated with the two values. The ratio between the number of correct elements retrieved and the total number of related elements that are presented in the database.

4.7. F-Measure

Through the f-score, we calculate precision and recall measuring scores. It calculates the results of precision and recall more accurately. It defines the ratio between precision into recall and precision plus recall into two. F-score is a more accurate and balanced value of recall and precision. After the calculation of the f-score, we get the values which are given in Table 4.

Table 4. Precision, Recall, and F score evaluation comparison of three layers

	FC6			FC7			FC8		
Video	**Precision**	**Recall**	**F Score**	**Precision**	**Recall**	**F Score**	**Precision**	**Recall**	**F Score**
1	0.40	0.20	0.16	0.60	0.30	0.40	0.80	0.40	0.53
2	0.40	0.20	0.16	0.60	0.30	0.40	1.00	0.50	0.76
3	0.60	0.30	0.40	0.60	0.30	0.40	1.00	0.50	0.76
4	0.60	0.30	0.40	0.80	0.40	0.53	1.00	0.50	0.76
5	0.60	0.30	0.40	0.80	0.40	0.53	1.00	0.50	0.76

Table 4 shows the simulation results of three layers in terms of precision, recall, and F-score. In the above table, every layer corresponds to the video 1, 2, 3, 4, and 5, and the achieved precision, recall, and f-score values. For the five videos FC6, FC7, and FC8 achieved the values listed in the above table; the FC6 layer obtained a precision value of 0.60, recall of 0.30, and f-score of 0.40. Similarly, the FC7 layer obtained a precision value of 0.80; recall value of 0.40 and f-score value of 0.53, respectively. Finally, FC8 obtained a precision value of 1.00, recall value of 0.50, f-score value of 0.76. Furthermore, the proposed Alex Net FC8 layer achieved the highest values of precision, recall, and f-score. It can be observed that the precision value of the proposed AlexNet FC8 model is better than that of the FC6 and FC7 models.

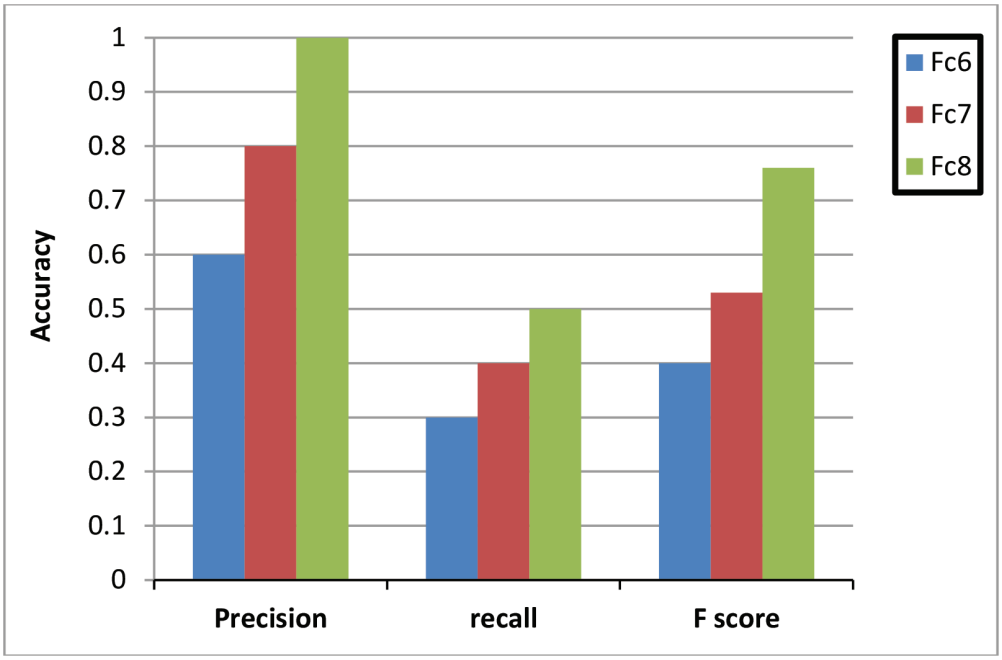

Figure 12. Performance of proposed Alex net layer-based Confusion matrix

Furthermore, in Figure 12 the proposed Alex Net FC8 layer achieved the highest values of precision, recall, and f-score.it is observed that the precision value of the proposed AlexNet FC8 layer is better than other FC6 and FC7 layers. Because FC6 and FC7 give low precision and recall values and FC8 gets a high value closer to 1. When the precision and recall values are near 1 it shows that the system performance is better.

4.8. Time Complexity Performance

The proposed system has been evaluated and tested on GeForce NVidia 8 GB dedicated GPU with Windows 10 operating system. MATLAB has been used as a simulation and programming tool which is best suitable for rapid prototyping. Time complexity model and performance is illustrated in Table 5.

Table 5. Time Complexity

Video No.	Total Number of Frames	Number of Cluster Frames	Time on GPU
1	164	6	1186 sec
2	151	6	1120 sec
3	300	6	1471 sec
4	188	6	1232 sec
5	159	6	1594 sec

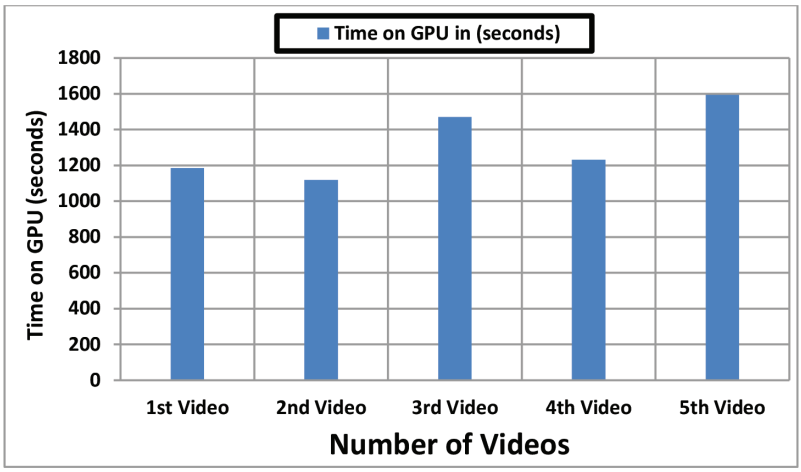

Figure 13. Performance of Proposed Method on GPU

Figure 13 shows the GPU testing times for each video. In the first video, there were 164 frames in total and six frames were chosen as key frames and the testing took a total of 1186 seconds. In the second video, there was a total of 151 frames and six frames were selected as key frames and the testing time took 120 seconds in total. In the third video, there were 188 frames in total, frames were selected

as key frames, and the testing took 1471 seconds in total. In the fourth video, there were 188 frames in total and six frames were selected as key frames and the training took1232 seconds. In the fifth video, there were 159 frames in total and six frames were selected as key frames and the training took1594 seconds in total.

4.9. Training and Testing of the Proposed System

Training and testing are key aspects of the CBIR & CBVR system in which the real-time concept is used for recently obtained datasets, so that the required output may be generated. This system has been utilized, over the last two decades, by platforms such as YouTube and Dailymotion in which many related videos are displayed because of the contents in the queried videos. Thus, when a video is searched, the results display several related videos in which similar but not the same activities are taking place. First of all, the data are obtained and trained for testing, however, this is a very generic explanation. In some systems, the data are stored in their databases i.e., pre-stored data in which the feature models do exist. In the new FV (feature vector) system, the features of the dataset are saved and then are used for testing against the mentioned dataset. Now, if we take a dataset that needs training and then that new dataset is taken accordingly. These two datasets are tested against each other in the current scenario for further processing. In other words, the new data is taken and then tested in an efficient and effective manner with respect to the desired task. The proposed CBVR is a real-time system which needs no pre-data storage or taking data from the database. This system takes the dataset as input which is trained and then is tested. It can take new queries for that data to be processed. The data is tested against the recently taken data due to the real-time nature of the proposed system. The new queries are used to search for the data that is taken as input. The queries are used to search for the data required for training/testing in the existing database.

5. Conclusion

In this work, we have proposed a new technique through which we extract the features of the contents of videos. For video retrieval, the color histogram of the video was stored in the feature vector and the color histogram of the selected frames was calculated. Key frames were selected through the SURF algorithm. From the key frames, we extracted color histogram features through the AlexNet model of CNN. The proposed approach is fundamentally very powerful because of low execution times. The results of the system were calculated through the Euclidean distance equation. The value of the input query was subtracted from the database query. The lowest values become the topmost result. The performance was evaluated on the equation of accuracy and loss. While the recent methods have the ability to process millions of videos in a very short time, the contribution of this paper is the implementation of the color histogram and bag of words of the SURF, as they had not been implemented in the existing literature. To the best of our knowledge, this is the first attempt that has been conducted on video retrieval systems for the extraction of features from video files rather than text files. By applying color histogram and SURF, the proposed model has achieved an accuracy of over 90%, which is considered outstanding in contrast with the existing state of the art solutions. Another main and key contribution is the evaluation based on the performance parameters, which has allowed us to improve the proposed model by modifying it and generating alternative results. Finally, these results were compared and properly justified.

Conflict of Interest: The author declares no conflict of interest.

6. References

Anzid, H., le Goic, G., Bekkari, A., Mansouri, A., & Mammass, D. (2023). A new SURF-based algorithm for robust registration of multimodal images data. The Visual Computer, 39(4), 1667-1681. https://doi.org/10.1007/s00371-022-02435-z

Awasthi, D., & Srivastava, V. K. (2022). Robust, imperceptible and optimized watermarking of DICOM image using Schur decomposition, LWT-DCT-SVD and its authentication using SURF. *Multimedia Tools And Applications*, *82*(11), 16555-16589. https://doi.org/10.1007/s11042-022-14002-8

Bhagat, M., & Kumar, D. (2023). Efficient feature selection using BoWs and SURF method for leaf disease identification. Multimedia Tools and Applications, 1-25.

Fan, J., Yang, X., Lu, R., Li, W., & Huang, Y. (2023). Long-term visual tracking algorithm for UAVs based on kernel correlation filtering and SURF features. *The Visual Computer*, *39*(1), 319-333. https://doi.org/10.1007/s00371-021-02331-y

Huang, C., Vasudevan, V., Pastor-Serrano, O., Islam, M. T., Nomura, Y., Dubrowski, P., et al. (2023). Learning image representations for content-based image retrieval of radiotherapy treatment plans. *Physics in Medicine & Biology*, *68*(9), 095025. https://doi.org/10.1088/1361-6560/accdb0

Huo, S., Zhou, Y., Xiang, W., & Kung, S. Y. (2023). Weakly-supervised content-based video moment retrieval using low-rank video representation. *Knowledge-Based Systems*, *277*, 110776. https://doi.org/10.1016/j.knosys.2023.110776

Hussain, A., Ahmad, M., Hussain, T., & Ullah, I. (2022). Efficient content based video retrieval system by applying AlexNet on key frames. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *11*(2), 207-235. https://doi.org/10.14201/adcaij.27430

KA, R., Simon, M. D., & Sumathy, G. (2023). Novel Fuzzy Entropy Based Leaky Shufflenet Content Based Video Retrival System.

Kakizaki, K., Fukuchi, K., & Sakuma, J. (2023). Certified Defense for Content Based Image Retrieval. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 4561-4570). https://doi.org/10.1109/WACV56688.2023.00454

Kavitha, A. R., Simon, M. D., & Sumathy, G. (2023). Novel Fuzzy Entropy Based Leaky Shufflenet Content Based Video Retrival System.

Kovač, I., & Marák, P. (2023). Finger vein recognition: utilization of adaptive gabor filters in the enhancement stage combined with sift/surf-based feature extraction. Signal, *Image and Video Processing*, *17*(3), 635-641. https://doi.org/10.1007/s11760-022-02270-8

Megala, G., Swarnalatha, P., Prabu, S., Venkatesan, R., & Kaneswaran, A. (2023). Content-Based Video Retrieval With Temporal Localization Using a Deep Bimodal Fusion Approach. In P. Swarnalatha & S. Prabu (Eds.), *Handbook of Research on Deep Learning Techniques for Cloud-Based Industrial IoT* (pp. 18-28). IGI Global. https://doi.org/10.4018/978-1-6684-8098-4.ch002

Mounika, B. R., Palanisamy, P., Sekhar, H. H., & Khare, A. (2023). Content based video retrieval using dynamic textures. *Multimedia Tools and Applications*, *82*(1), 59-90. https://doi.org/10.1007/s11042-022-13086-6

Prathiba, T., & Kumari, R. S. S. (2023). Retraction Note to: Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer

interaction. *J Ambient Intell Human Comput, 14* (Suppl 1), 315. https://doi.org/10.1007/s12652-022-04085-4

Prathiba, T., Shantha Selva Kumari, R., & Chengathir Selvi, M. (2023). ALMEGA-VIR: face video retrieval system. The Imaging Science Journal, 1-11.

Rastegar, H., & Giveki, D. (2023). Designing a new deep convolutional neural network for content-based image retrieval with relevance feedback. Computers and Electrical Engineering, 106, 108593.

Salih, F. A. A., & Abdulla, A. A. (2023). Two-layer content-based image retrieval technique for improving effectiveness. Multimedia Tools and Applications, 1-22.

Salih, S. F., & Abdulla, A. A. (2023). An effective bi-layer content-based image retrieval technique. The Journal of Supercomputing, 79(2), 2308-2331.

Sikandar, S., Mahum, R., & Alsalman, A. (2023). A Novel Hybrid Approach for a Content-Based Image Retrieval Using Feature Fusion. Applied Sciences, 13(7), 4581.

Sowmyayani, S., & Rani, P. A. J. (2023). Content based video retrieval system using two stream convolutional neural network. Multimedia Tools and Applications, 1-19.

Usher, L. E. (2023). The case for reflexivity in quantitative survey research in leisure studies: lessons from surf research. Annals of Leisure Research, 26(2), 269-284.

Veselý, P., & Peška, L. (2023, January). Less Is More: Similarity Models for Content-Based Video Retrieval. In International Conference on Multimedia Modeling (pp. 54-65). Cham: Springer Nature Switzerland.

Victoria Priscilla, C., & Rajeshwari, D. (2023). Performance Analysis of Spatio-temporal Human Detected Keyframe Extraction. Journal of Survey in Fisheries Sciences, 10(2S), 233-243.

Vieira, G. S., Fonseca, A. U., & Soares, F. (2023a). CBIR-ANR: A content-based image retrieval with accuracy noise reduction. Software Impacts, 15, 100486.

Vieira, G., Fonseca, A., Sousa, N., Felix, J., & Soares, F. (2023b). A novel content-based image retrieval system with feature descriptor integration and accuracy noise reduction. Expert Systems with Applications, 120774.

Walter, K. H., Otis, N. P., Miggantz, E. L., Ray, T. N., Glassman, L. H., Beltran, J. L., ... & Michalewicz-Kragh, B. (2023). Psychological and functional outcomes following a randomized controlled trial of surf and hike therapy for US service members. Frontiers in Psychology, 14, 1185774.

Wickstrøm, K. K., Østmo, E. A., Radiya, K., Mikalsen, K. Ø., Kampffmeyer, M. C., & Jenssen, R. (2023). A clinically motivated self-supervised approach for content-based image retrieval of CT liver images. Computerized Medical Imaging and Graphics, 107, 102239.

Author's Biography

Altaf Hussain received his MS and BS Degrees in Computer Science from The University of Agriculture Peshawar, Pakistan (2017) and University of Peshawar, Pakistan (2013), respectively. He worked at The University of Agriculture as a Student Research Scholar from 2017-2019. During his MS Degree he has completed his Research in Computer Networks especially in Routing Protocols in Drone Networks. He is a PhD Scholar in School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. He has served as a Lecturer in Computer Science Department in Govt Degree College Lal Qilla Dir L, KPK Pakistan from 2020-2021. He has published many Paper papers including survey/review and conference papers. He was Research Scholar in (Career Dynamics Paper Academy) Peshawar, Pakistan for one and a half year. He worked as a Research assistant with the department of Business and Economics, Qatar University, Doha, Qatar. He also worked as IT clerk in the court of district and session judge Timergara Dir Lower. Currently, he is a PhD scholar in School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His Research specialties/interests include Wireless Networks, Sensor Networks, Unmanned Aerial Vehicular Networks, Deep Learning, and Image Processing.