IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

A Survey on Content-Based Video Retrieval Techniques

Nagariya Maitree[1]*, U. K. Jaliya[2], M. S. Holia[3]

[1,2]Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, VVNagar, Anand, Gujarat, India
[3]Department of Electronics Engineering, Birla Vishvakarma Mahavidyalaya, VVNagar, Anand, Gujarat, India

DOI: https://doi.org/10.26438/ijcse/v7i2. 878883 | Available online at: www.ijcseonline.org

Accepted: 18/Feb/2019, Published: 28/Feb/2019

***Abstract*—** In the recent digital world, the amount of processing of videos is increasing rapidly. For this purpose, video retrieval systems are dominating today's world. Video retrieval systems include proper analysis of videos for appropriate retrieval. The retrieval of videos can be done based on the text or annotation attached to it. But retrieval based on the content has become more influencing over text-based retrieval as it describes a video in a much better way than described by text. Content-based video retrieval systems analyze the contents of a video such as colour, texture, shape, etc. This system involves many stages with multiple techniques for each one as per the survey done till now. To analyze the different techniques, multiple datasets have been used containing videos of different categories. The best technique applied at each stage for frame extraction, feature extraction, classification and retrieval of videos makes the system more accurate and efficient.

***Keywords*—** Video retrieval, Key-frame extraction, SURF, SIFT, BRISK, SVM.

I. INTRODUCTION

The increase in amount of data such as images, audio, video, etc. has led to increase in processing of the data [12]. This data must be stored and managed properly which can further be used for searching or analysis purpose. The concept of searching a desired video from a huge database of videos is basically referred to as video retrieval [9]. Such retrieval systems are used for quicker searching of videos, visual e-commerce analysis, digitized galleries, news episodes management, intelligent web videos management, content linking, video surveillance, etc [1][9].

There are basically two ways of retrieving videos from database: (a) Annotation-based retrieval and (b) Content-based retrieval. Annotation-based video retrieval refers to text-based retrieval which uses the metadata attached with it. The metadata involves the caption and keywords of the video. Content-based video retrieval refers to the analysis of the contents of the video such as colour, motion, texture, shape, etc. Text-based retrieval uses text form for query as an input which sometimes becomes irrelevant for the users. Thus, content-based retrieval then becomes useful. It uses an image or object or video clips as an input reducing the burden for user to write the text for any required video [1].

The main objective of content-based video retrieval systems is (a) quicker searching of video and (b) accurate retrieval of a video. The searching of a video should be faster and the video provided as an output by the system must be the most relevant one from the given database [1].

II. Content-Based Video Retrieval:

A video retrieval system takes an input in the form of image, object or video. After processing and analyzing, it gives a video as output. The following figure describes the basic flow of a video retrieval system:

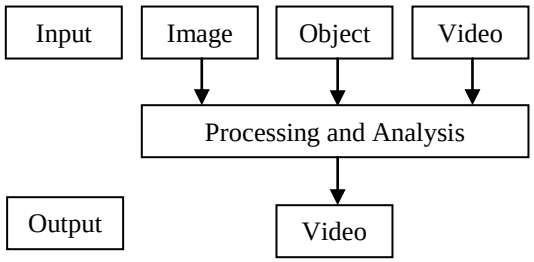

Figure 1. Basic Retrieval System

A video is formed by scenes, shots and sequence of frames. A scene is a series of shots depicting a continuous event. A sequence of consecutive frames from a stable place is termed as a shot. A frame is any 2-dimensional image. So, a sequence of number of frames forms a shot. A number of shots form a scene. Multiple scenes form a video.

Further in the system, the frames are processed instead of videos.

There are mainly three steps in a video retrieval system: (a) Key-frame extraction (b) Feature extraction (c) Checking the similarity measures with classification.

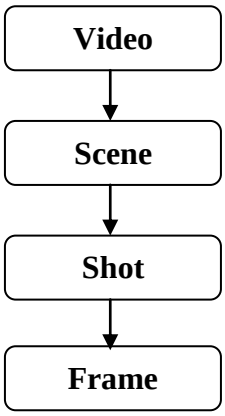

Figure 2. Fragments of a video [2]

The flow diagram of a video retrieval system is as shown in the following figure.

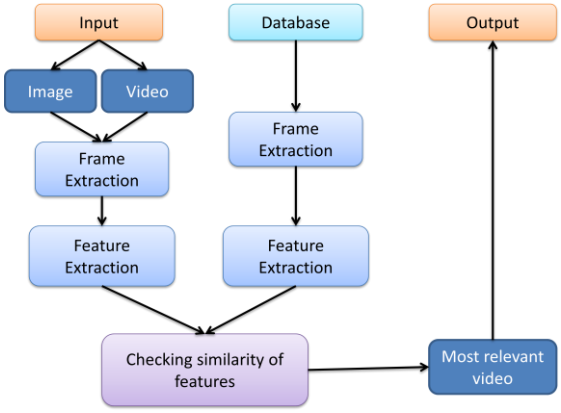

Figure 3. Flow of the system (CBVR) [2]

A. Key-Frame Extraction:

Key-frame extraction is the fundamental step of video retrieval system. A key frame is a frame that represents salient contents of a shot. It should be selected such that the contents of the shot are reflected in the best possible way with maximum content and minimum redundancy. Key-frames are also known as R-frames (Representative frames). Processing of unwanted frames is of no use. By appropriate selection of the key-frames, the amount of processing is also reduced. There are different approaches for the selection of key-frames from the total number of frames extracted. One of the approaches is to select any random frame or the first frame of every shot but in such an approach there may be loss of visual information which may represent the entire shot in a better way. Another approach is to set some static threshold for measuring the similarity between frames but this may not be suitable for all the videos and may lead to loss of meaningful information. The other proposed method is to first identify the shots by shot detection algorithm. Then mean is calculated for each frame and stored in a vector. For that vector, the local maxima and minima is calculated and then compared with the mean value. If it matches, then that frame index is observed and the frame is selected as the key-frame. This method is preferable over others as it provides better efficiency than other methods [7].

The following table shows the number of actual frames of a video along with the number of frames extracted as key-frames, the size of the video and its duration.

Table 1. A view at the number of extracted key frames [3]

Query Video	**Frames**	**Video Dura-tion (sec)**	**Vid-eo size**	**No. of extrac-ted frames**
Starwars .mpg	2919	120.2	1.66 Mb	89
Shoab3. mpg	95	04	748 Kb	03
Baryrich ard.mpg	316	12	2.16 Mb	03

Shot Detection:

The Shot Detection is a fundamental step for content-based video retrieval applications. This can also be referred to as cut detection method. A shot is a sequence of frames from one camera only without any interruption in between them. When there is a sudden change between two frames, it is termed as a hard-cut. When one frame gradually replaces another frame, it is termed as a soft transition. A shot can be detected by approaches such as sum of absolute differences, histogram difference or edge change ratio. The sum of absolute differences mainly used for hard-cuts and rarely identifies soft-cuts. It is more sensitive to soft-cuts. The edge change ratio as a score is sensitive to both hard-transitions as well as soft-transitions. Thus edge change ratio is preferable over others for identifying a shot. After identifying the shots, it is necessary to be measured whether that is correctly a shot or not. So, measures such as precision and recall can be used for this. Precision identifies that a correct cut is detected. Recall identifies that a cut assumed is really a cut or not [7].

B. Feature Extraction:

Feature extraction extracts the features from the key-frames selected. The features may be colour, shape, texture, motion, etc. For extracting colour features, there is a technique known as Block Truncation Coding which further improved to Thepade's Sorted Ternary Block Truncation Coding. Vector Quantization technique used

the hybrid features including colour features as well as transform features. It used the concept of codebook generation. Gabor filters are also used to extract edges as well as the regions containing objects. These are included in category of Gabor features. These features can be obtained at different scales and orientations. In this strategy, first the image is divided into sub-blocks. Then set of magnitudes is calculated from different angles and scales. The mean and standard deviation is calculated to obtain Gabor feature vector containing the texture features. For the measurement of the performance of better feature extraction technique, the values obtained for precision and recall are much higher for BTC and KFCG (Kekre's Fast Codebook Generation) in comparison to Gabor features. KFCG is basically used for image compression. It requires less time to generate the codebook with the use of vector quantization method. The vector quantization is used for lossy data compression.

Other techniques are BRISK (Binary Robust Invariant Scalable Keypoints), SURF (Speeded Up Robust Features), FAST (Features from Accelerated Segment Test), SIFT (Scale Invariant Feature Transform), HOG (Histogram of Oriented Gradients). SURF is a technique basically used for object recognition and classification. It is an improved technique than SIFT. Its feature descriptor is based on Haar wavelet response around the point of interest. FAST (Features from Accelerated Segmented Test) is used to detect the corner features for object tracking and mapping. It is used for real-time video processing. HOG (Histogram of Object Gradient) is mainly for object detection. It can be more improved with normalization method. BRISK (Binary Robust Invariant Scalable Key-points) has low complexity and contains bit-string vector. It makes use of Hamming distance rather than Euclidean distance [9].

SIFT:

It is a feature detection algorithm used to detect the local features of an image. It is widely for applications such as robotic mapping and navigation, object recognition, video tracking, gesture recognition, image stitching. As this is a scale-invariant algorithm, it is invariant to orientation, scale and rotation and robust to changes in illumination, noise and minor changes in the viewpoint. The main stages are scale-space extrema detection, keypoint localization, orientation assignment and keypoint descriptor. The scale-space step involves the construction of a scale-space and approximation of Laplacian of Gaussian representation of images. In the step of keypoint localization, we find the keypoints by using minima and maxima of Difference of Gaussian images. Then we eliminate the bad keypoints. For orientation assignment, an orientation is assigned for each keypoint and then is made rotation invariant. At last the SIFT features are generated as the descriptors which uniquely identifies the features [11][12].

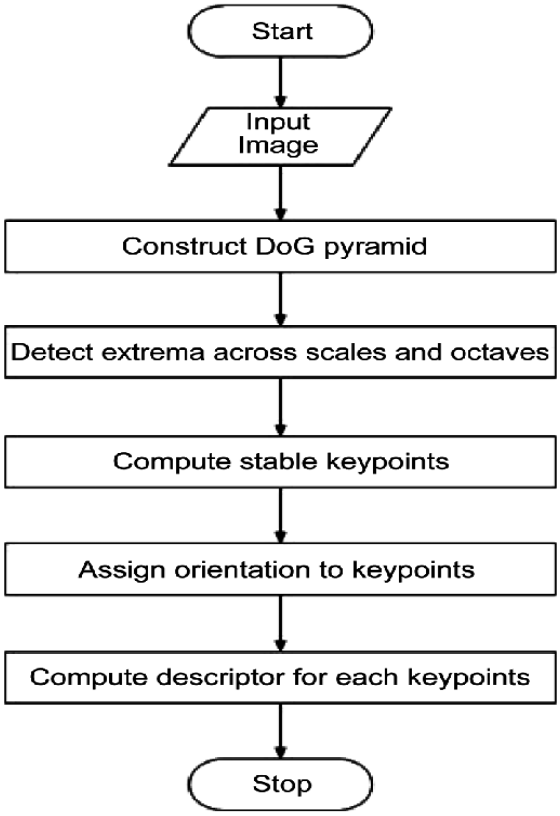

Figure 4. SIFT algorithmic steps [15]

SURF:

The following figure shows us the basic steps of the SURF algorithm:
It is a local feature detector and descriptor. It works much faster than SIFT. It is used for object recognition, image registration, classification and 3D reconstruction.

The working steps are similar to that of SIFT but the methods used for each stage is different. After generating the scale-space, for the approximation, it uses square-shaped features instead of the cascaded filters used by SIFT algorithm. For finding the interesting keypoints, it uses Blob detector which works on the basis of Hessian matrix. For the descriptor building, the description of intensity distribution of each pixel is used. These descriptors are based on the responses of Haar wavelet. Then the descriptors of different images are used for matching [13].

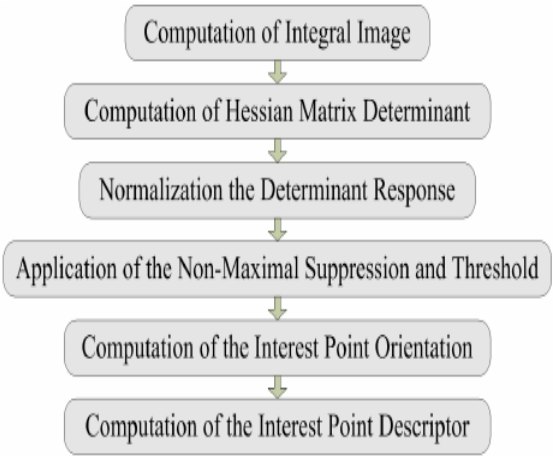

Figure 5. SURF algorithmic steps [14]

BRISK:

With comparison to SIFT and SURF, BRISK achieves much better quality of matching and at lower computation time. The procedure involves key stages as feature detection, descriptor composition and key point matching. For each keypoint, true scale is estimated over the scale-space. The descriptor formed consists of binary strings. The characteristic direction of each key point is identified for allowance of orientation- normalized descriptor to make it rotation-invariant. For maintaining the prevention of aliasing effect after sampling, Gaussian smoothing is applied for choosing fewer key points. The bit string obtained is of length 512 bits. For matching the two descriptors, instead of using Euclidean distance for comparison, Hamming distance is used which only involves XORing of two vectors that leads to less computation time. The following table shows comparison between the SIFT, SURF and BRISK when compared for a graffiti image:

Table 2. Comparison between SIFT, SURF, BRISK for an image [10]

	SIFT	**SURF**	**BRISK**
Points in first image	1851	1557	1051
Points in second image	2347	1888	1385
Total time [ms]	291.6	194.6	29.92
Time per comparison [ns]	67.12	66.20	20.55

Figure 6. Matching with BRISK descriptor [10]

C. Comparison by Similarity Measures:

Checking the similarity measures involve comparison of the features of frames in the database with the features of frames given as input query. The video with which maximum features are matched is the most relevant video which is provided as the output. Some of the similarity measures are Euclidean distance, Manhattan distance, Minkowski distance, Kullback-Leibler distance, SVM (Support Vector Machine). Amongst these, SVM can work more efficiently as it can be used for automatic classification of the videos.

SVM:

SVM is a supervised learning machine learning algorithm mainly used for classification and regression analysis. It may be linear as well as non-linear. When SVM uses non-kernel, it makes the boundary which is not a straight line. It can handle complexity. It works perfectly fine with a clear line for separation. But for larger datasets, it requires more processing time. When it actually works in a video retrieval system, it makes use of the features of the video. It extracts the feature vector from the frames or shots. While training the data with the use of SVM, some videos according to the category are provided as a training set. Other videos are used for classification. For computing the similarity between the features of frames, selection of features also plays a vital role. This is because the selection of features only helps in calculating the similarity. After classification, it measures that which video is most relevant from the dataset. For this, Euclidean distance is also used such that the video query having minimum distance with the video from dataset has to be chosen as the most relevant one to be provided as the output.

III. DATASETS

For the testing of the techniques mentioned in above sections, the dataset that was used consists of many

categories of data in which further each category consists of multiple videos.

IV. PARAMETERS

For the measurement of performance of efficient classification, the parameters that are taken into consideration are precision and recall. Both the parameters are in trade-off. If one parameter improves, the other one degrades. Recall refers to the ratio of the total number of similar clips detected correctly to the total number of similar clips in the database. Precision refers to the ratio of the total number of similar clips detected correctly to the total number of detected clips.

$$Recall = \frac{DC}{DB}$$

$$Precision = \frac{DC}{DT}$$

where DC denotes similar clips detected correctly, DB denotes similar clips in database and DT denotes number of detected clips.

There might be some challenges that the performance and efficiency becomes much acceptable when the features are present which are not similar to others. Also, the performance degrades when features belonging to other videos are identical.

V. CONCLUSION

Content-Based Video Retrieval systems are much useful than text-based retrieval systems as it becomes more accurate and efficient to use. Its different stages also include many different approaches from which accurate results can be achieved as per requirement. Histogram method works well for extracting the key frames. For feature extraction and comparison, BRISK and SURF is better as per the results observed from the survey till now. SVM is preferable over using other similarity measures.

REFERENCES

[1] Prof. Rahul Gaikwad and Jitesh R. Neve, "A Comprehensive Study in Novel Content Based Video Retrieval Using Vector Quantization over a Diversity of Color Spaces", in the Proceedings of 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.

[2] Prof. Dipak R. Pardhi and Jitesh R. Neve, "Performance Rise in Novel Content Based Video Retrieval using Vector Quantization", in the Proceedings of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[3] Andre Araujo And Bernd Girod , "*Large-scale Video Retrieval Using Image Queries*", IEEE Transactions On Circuits And Systems For Video Technology, Vol. 28, No. 6, June 2018.

[4] Aasif Ansari, Muzammil H Mohammed, "Content-based video retrieval systems-methods, techniques, trends and challenges", in the Proceedings of International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 7, February 2015.

[5] Dr. Parag Kulkarni, Bhagyashri Patil, Bela Joglekar, "An effective content based video analysis and retrieval using pattern indexing techniques", in the Proceedings of 2015 International Conference on Industrial Instrumentation and Control, College of Engineering Pune, India, May 28-30, 2015.

[6] Mohd.Aasif Ansari, HemlataVasishtha, "*Content-based video retrieval systems performance based on multiple features and multiple frames using SVM*", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, 2016.

[7] K.S.Thakre, A.M.Rajurkar, R.R.Manthalkar, "Video Partitioning and Secured Keyframe Extraction of MPEG Video", in the Proceedings of International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

[8] Jun Xu , Tao Mei , Ting Yao and Yong Rui, "*MSR-VTT: A Large Video Description Dataset for Bridging Video and Language*"

[9] Ashwini B, Verina, Dr.Yuvaraju B N, "*Feature Extraction Techniques for Video Processing in MATLAB*", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization),Vol. 4, Issue 4, April 2016.

[10] Stefan Leutenegger, Margarita Chli and Roland Y. Siegwart, "*BRISK: Binary Robust Invariant Scalable Keypoints*"

[11] Wikipedia contributors. (2019, February 19). Scale-invariant feature transform. In Wikipedia, The Free Encyclopedia. Retrieved 08:53, February 28, 2019, from https://en.wikipedia.org/w/index.php?title=Scale-invariant_feature_transform&oldid=884107628

[12] AI Shack, SIFT algorithm steps from http://aishack.in/tutorials/sift-scale-invariant-feature-transform-introduction/

[13] Wikipedia contributors. (2017, August 20). Speeded up robust features. In Wikipedia, The Free Encyclopedia. Retrieved 08:58, February 28, 2019, from https://en.wikipedia.org/w/index.php?title=Speeded_up_robust_features&oldid=796404867

[14] Sledevič, Tomyslav & Serackis, Artūras. (2012). SURF algorithm implementation on FPGA. 291-294. 10.1109/BEC.2012.6376874.

[15] Raj Prasanna Kumar, Raghu & Muknahallipatna, Suresh & McInroy, John. (2016). "*An Approach to Parallelization of SIFT Algorithm on GPUs for Real-Time Applications*". Journal of Computer and Communications. 04. 18-50. 10.4236/jcc.2016.417002.

Authors' Profile

Miss Nagariya Maitree R. pursued Bachelor of Engineering in Computer from Vadodara Institute of Engineering, Kotambi, Vadodara, Gujarat, India in 2017.She is currently pursuing Master of Technology in course of Computer Engineering (Software Engineering) from Birla Vishvakarma Mahavidhyalaya (BVM), VVnagar, Anand, Gujarat, India. Her main research work focuses on Image Processing-Computer Vision, Video Retrieval Techniques.

Mr. Udesang K Jaliya pursued Bachelor of Engineering in Computer Engineering and Master of Engineering in Computer Engineering. He has also pursued Ph.D. and currently working as Assistant Professor in Department of Computer Engineering, in Birla Vishvakarma Mahavidhyalaya (BVM), VVnagar, Anand, Gujarat, India. He has published more than 35 research papers in reputed international journals and national journals conferences including IEEE and it's also available online.. He also received Best Paper Award in IEEE Sponsored International Conference on Data Mining and Advanced Computing. He has 15 years of teaching experience . He also Guided more than 20 Project of master level .He also published 2 Books related to their field area. He is also registered as a PhD supervisor at GTU.

Mrs. Mehfuza S Holia pursued Bachelor of Engineering in Electronics Engineering and Master of Engineering in Electronics and Communications Engineering. She has also pursued Ph.D. and currently working as Assistant Professor in Department of Electronics Engineering, in Birla Vishvakarma Mahavidyalaya (BVM), VVNagar, Anand, Gujarat, India. She has 15 years of teaching experience.

