Survey on content Based Video Retrieval

Borkar S.V[1], Katariya S.S[2].

[1] M.E. Electronics, Department of Electronics Engineering, Amrutvahini College of Engineering, Sangamner, Maharashtra, India.
[2] Professor in Electronics Department, Department of Electronics Engineering, Amrutvahini College of Engineering, Sangamner, Maharashtra, India.

ABSTRACT

Content based video retrieval and analysis is the one of the most important and recent research area in the Image processing domain. Video retrieval is a wide spectrum of promising applications, motivating the researchers from worldwide. Due to the availability of increased internet bandwidth and transfer media, the multimedia objects usage is widely used. The number of videos is being uploaded through various sources. It requires some kind of retrieval technique which will aid users from tracking videos relatively similar manually. In this paper represents an overview of the general methods used in content based video retrieval. It focuses on the different methods for video structure analysis, including key frame extraction, shot segmentation, scene segmentation, feature extraction, video annotation and video retrieval method.

Keywords: *Feature extraction, shot detection, Scene segmentation, Video retrieval, Video annotation, Video structure analysis.*

1. INTRODUCTION

There has been a widely growth in the usage of digital data. The most wide spread and common among the digital data is digital video which has become most essential part of our lives. Even if there are many tools to retrieve and process the digital data, it is a little difficult, less efficient and less effective. In the present day scenario the technologies that allow easy capture and sharing of digitized video have been rapidly developing, for example internet enabled mobile phones equipped with digital camera [2].

Videos are communicative and powerful media that can capture and present information. The apace expanding digital video information has motivated growth of new technologies for efficacious browsing, annotating and retrieval of video data. The challenges behind the design and implementation of the content based video browsing; indexing and retrieval systems have attracted researchers from much compliance. It is mostly accepted that successful solution to the problem of understanding and indexing the videos requires combination of information from different sources such as images, text, audio, speech etc. Content based video retrieval has a large range of applications such as analysis of visual electronics commerce, quick video browsing, remote instructions, digital museums [1], intelligent management of the web videos, news video analysis and video surveillance. A video may have a visual channel as well as auditory channel.

Many works has been done in video the videos are being processed in terms of video indexing, video classification, shot boundary detection, concept based video retrieval etc. And now this is in the advance of switching from concept to content based video retrieval. The generic framework of content based video retrieval system is as [4] shown in fig.1

2. VIDEO INDEXING

The process of indexes for videos normally involves the following three main steps:

2.1 Video Parsing: Manipulation of whole video for breakdown into key frames [18]. Video parsing is the process detecting scene changes or the boundaries between camera shots in video stream. It consists of temporal segmentation of the video contents into smaller units. These methods extract structural information from the video by identifying significant segments and detecting temporal boundaries, called *shots.*

2.2 Abstraction: It consists of extracting the representative set of video data from the video. The most largely used video abstractions are the key frame and the highlight sequence. The result of video abstraction forms the basis for the video browsing and indexing.

2.3 Content Analysis: It consists of extract visual features from key frames.

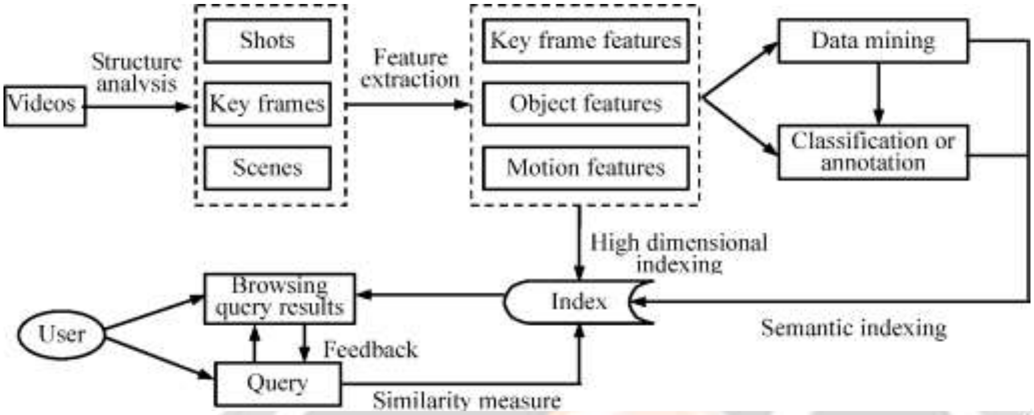

Fig. 1. General framework for content-based video indexing and retrieval.

1. VIDEO PARSING

Video parsing is the process of breaking a video into smaller units. Videos are structured according to a descending manner of video, scenes, shots, and frames.

- *Frame level*: Each frame is treated separately. There is no changeable analysis at this level.
- *Shot-level*: A shot is a set of contiguous frames all acquired through a continuous camera recording. Only the changeable information is used. The basic unit is called as a shot.
- *Scene-level*: A scene is a set of appended shots having a common semantic significance. A *scene* or *sequence* is formally defined as a temporally adjacent shots, and collection of semantically related, depicting and conveying a high-level concept or story
- *Video-level*: The complete video object is treated as a whole [9][18].

Video structure analysis aims at segmenting a video into a number of structural elements that have semantic contents, including shot boundary detection, key frame extraction, and scene segmentation.

3.1 shot boundary detection

These processes of detecting boundaries between two continuously shots, so that a sequence of frames belonging to a shot will be grouped together. First extract visual features from each frame then measure similarities between frames using the extracted features and finally detect shot it is the methods for shot boundary detection. The basis for detecting shot boundaries is the detection of significant changes in contents on consecutive frames lying on either side of a boundary. Automatic shot boundary detection techniques can be classified into pixel, threshold, edges and motion [9].

3.1.1. Pixel-based: The easiest way to detect a shot boundary is to count the number of pixels that change in value more than some threshold. The major problems with this approach are its noise and sensitivity to camera movement.

3.1.2. Histogram-based: The most popular metric for sharp transition detection is the difference between histograms of two continuously frames. In its simplest form, the gray level or color histograms of two continuously frames are computed: if the bin-wise difference between the two histograms is above a threshold then shot boundary is said to be found [7].

3.1.3. Threshold-Based: The threshold-based approach detects shot boundaries by comparing the measured Pair-wise similarities between frames with a pre determined threshold. The boundary is detected when a similarity is less than the threshold. The threshold can be adaptive, global or global and adaptive combined. Transform-based techniques use the compressed DCT coefficients present in an MPEG stream as the boundary measure [8].

3.1.4. Edge-based: Mai and Zabih Miller proposed an edge-based algorithm. The percentage of edges that exit and enter between the two frames is then computed. Dissolves and fades are identified by exiting edge percentages and looking at the relative values of the entering. They adduce that their method was more accurate at detecting cuts than histogram based techniques.

3.1.5. Motion-based: Zhang, Kankanhalli, and Smoliar used motion vectors determined from block matching to detect whether or not a shot was a zoom or a pan. Shahraray used the motion vectors extracted as part of a region-based pixel difference computation to decide if there is a large amount of camera or object motion in a shot. The motion characteristics change this content of frames in time. This change occurs due to camera motion and/or object motion. Camera motion is the major contributor to global motion, is generally due to panning, zooming, tilting etc. The characteristic of dynamic videos that distinguishes them from still images is the motion of objects and motion of background against each other. The foreground motion is caused by moving objects whereas the background motion is caused by camera motion [7].

3.1.6. Others: Other supervised learning algorithms have been employed for shot boundary detection. For instance, Cooper uses the binary *k* nearest-neighbor (kNN) classifier, where the similarities between frames within the particular changeable interval are used as its input. Boreczky and Wilcox apply hidden Markov (HMM) models with separate states to model shot cuts, dissolves, fades, pans, and zooms.

3.2. Scene Boundary Detection:

The automatic detection of opposed to *physical* boundaries (*semantic* boundaries) within a video program is a much more challenging task and the subject of ongoing research. Scene segmentation is also called as story unit segmentation. In general, a scene is nothing but a group of appended shots that are consistent with a certain subject. Scenes have higher level semantics than shots. Scenes are identified by grouping the successive shots into a meaningfull semantic unit with similar content. The grouping may be based on information from audio track, texts or images in the video. According to shot representation, scene segmentation methods can be classified into three categories which is key frame based, Audio and vision integration-based, and background-based [1].

3.2.1. Key frame based : In the same shots there have been huge redundancies in frames that is why those frames which have the best reflection are chosen for the content of the shot as key frames and to concisely demonstrate the shot. As much as possible salient content of the shots must be contained by redundancy must be avoided and the extracted key frames. The characteristics of the key frame extraction are colour (usually the histogram colour), shapes, edges, optical flow, motion temporal intensity and spatial distribution of MPEG-7. The simplest and best way to select the key frame is key frame can be extracted by selecting the first and last frame as the key frame [7].

3.2.2. Audio and vision integration based approach: Audio features are essential in characterizing a video's affective Content. In fact, Wang and Cheong's study shows that audio features are often more informative than visual ones with respect to affective content characterization [12]. This method selects a shot boundary where the visual as well as audio contents change simultaneously as a scene boundary. The limitation of this approach is that it is difficult to determine the relation between visual shots and audio segments [1].

3.2.3. background-based: Background based approach segments the scenes under the acceptance that shots belonging to the same scene often have similar backgrounds. Chen et al. uses a mosaic technique to reconstruct the background of each video frame. Then, the texture and color distributions of all the background images in a shot are estimated to determine the shot similarity and the rules of filmmaking are used to guide the shot grouping process. The limitation of this method is the assumption that shots in the same scene have similar backgrounds: sometimes the backgrounds in shots in a scene can be different [1].

4. VIDEO ABSTRACTION

There are two main approaches to video abstraction: *key-frames* and *highlight sequences.*

4.1 key-frames: *It* is a frame that represents the content of a scene or shot. This content must be the most representative as possible. In the large amount of video data, we initially reduce each video to a set of representative key frames. In practice, often the first frame or center frame of a shot is chosen, which causes information loss in case of long shots containing considerable panning and zooming [3]. Key frame set extracted for a video segment3 by an arbitrary KFE method. For powerful video browsing and video retrieval, the selected key frames should be able to represent the content of the video segment [14].The actual approaches to extract key frames are classified into six categories 1)sequential comparison-based, 2) reference frame-based, 3)clustering based, 4)Global comparison-based, 5)curve simplification-based, and 6) object/event-based.

4.2 Highlight sequence: This approach also known as *video skimming* or *video summaries.* A successful approach is to utilize information from multiple sources (e.g., shot boundaries, human faces, camera and object motions, sound, speech, and text). A successful application of this type of approach has been the informed project, in which text and visual content information are merged to identify video sequences that highlight the important contents of the video. The extension of this skimming approach from documentary video programs to other videos with a soundtrack containing more than just speech remains an open research topic [4].

5. FEATURE EXTRACTION

Once object of interest have been segmented from the input video sequences, feature extraction is used to identify abnormal activities such as excessive human movement that may indicate that fighting is taking place, for instance. A shot of a person walking can be regarded as an instance that is segmented into a compilation of adjacent regions with different criteria like shape, color, Edge and texture, though all the regions may be consistent in their motion attribute. A feature is believed to be good only if, dissimilar objects are far from each other and similar objects are adjacent to each other in the feature space [9][19][20].

5.1. Static key frame features: The static key frame features are useful for video indexing &retrieval and are mainly classified as texture-based, color-based and shape based.

5.1.1. Color-based features: Color-based features include color correlograms, color histograms, color moments, a mixture of Gaussian models, etc. The exaction of color-based features depends on color spaces such as HSV, RGB, YCbCr , YUV, normalized r-g, and HVC [17][8]. The choice of color space depends on the applications. Different color representation schemes include chromaticity, luminance and red-green blue (RGB), system of International Commission on Illumination (CIE), hue-saturation intensity (HSI) and others. The RGB scheme is most commonly used in display devices [15].The limitation of color-based features is that they do not directly describe shape, texture, etc., and thus, ineffective for the applications in which shape or texture is important [1].

5.1.2. Texture-Based Features: Texture is that property of surfaces that shows visual patterns. Co occurrence Matrix method is used for Texture-Based Features extraction. Texture represented by pixels gives relative brightness of consecutive pixels and finds the regularity, coarseness degree of contrast and directionality which classifies textures as 'rough", smooth' etc. Texture is a visual pattern where there are a large number of visible elements evenly arranged and densely [10]. The merit of texture-based features is that they can be effectively applied to applications in which texture information is salient in videos. However, these features are unavailable in non texture video images. Depending on the texture on the foreground and background regions in a given video, we will get the trajectories from both parts from database videos and from input video. Techniques based on Wavelet, FFT, Gabor filters and spatial variance, etc. have been used to detect the textural properties of a text region in an image [20].

5.1.3. Shape-Based Features: Shape goes one step further than texture and color. It needs identification of regions to compare. There have been many shape similarity measures suggested for pattern recognition that can be used to construct shape distance measures [8]. Shape descriptions are an important task in content-based video retrieval. It is a mapping that converts the shape space into a vector space and satisfies the requirement that two similar shapes will also have close-to-identical shape descriptors. Hauptmann *et al*. use the edge histogram descriptor to capture the spatial distribution of edges for the video search task in TRECVid-2005 [10].

5.2. Object Features: Object features include the texture, dominant color, size etc. of the image regions corresponding to the objects. The limitation of object-based features is that identification of objects in videos is difficult and time-consuming. Current algorithms focus on identifying specific types of objects, such as faces, rather than various objects in various scenes [4][11].

5.3. Motion Features: Video is a kind of content-sensitive media carrying rich motion information; the optical flow field is used to give a good estimation of the motion. The motion characteristics change this content of frames in time. This change occurs due to object motion and camera motion. Camera motion is the major contributor to global motion, is generally due to panning, zooming, tilting etc. which leads to the change in background of the scene in the video. Motion of objects or Local motion changes the foreground [6]. A good representation of motion in videos can be used as a query for retrieval of similar videos. For camera-based features, different camera motions, such as "panning right or left," "zooming out or in," and "tilting down or up," are estimated and used for video indexing. Video retrieval using only camera-based features has the limitation that they cannot describe motions of key objects [4].

6. VIDEO ANNOTATION

Video classification and Video annotation share similar methodologies: initially low- level features are extracted then certain classifiers are trained and employed to map the features to the concept or category labels. Corresponding to the fact that a video may be annotated with multiple concepts, the methods for video annotation can be classified as concept based isolated annotation; context based annotation, and integrated based annotation [1].

6.1. Concept based Isolated Annotation: In a visual lexicon, this procedure of annotation has been used for every concept as a statistical detector trainee. For the discovery of multiple concepts of semantic, the classifiers of isolated binary have been utilized separately and autonomously although correlation among the perceptions has not been measured. The distribution of multiple Bernoulli has been used by Feng for the video and image annotation sculpturing. For every concept, the accuracy of different classifiers like HMM, GMM, kNN, and Adaboos t have been inspected by Naphade and Smith [4].

6.2. Context Based Annotation: By using different contexts for different concepts the concept detection performance can be improved. By using the context based concept fusion approach the context based annotation concepts of deduced higher level concepts or purifies the results of detection of the binary classifiers individually. Ontology hierarchy has been used to make the accuracy of the detection of the individual binary classifiers up to the mark. Model vectors have been made by Naphade and Smith which is dependent on the scores detection of those classifiers which are individual [1].

6.3. Integration Based Annotation: It covers the concepts of individual and its correlation at the same time. Concurrently, learning and optimization have been done. Along with that, all the samples have also been used for the individual concepts modelling and its correlation at the same time. A new feature vector is constructed, which grabs the concept's characteristics and concept's correlation, with the help a correlative algorithm of multi label proposed by Qi. The high computational complexity has been the drawback of the integration based annotation. The accurate amount of labelled training samples is required for the effective learning as well as robust detectors and with the help of feature dimensions the required numbers enhance exponentially [11].

7. QUERY AND VIDEO RETRIEVAL

The video retrieval which is content based starts its performance when the video indices have been attained. For the search of the user's video in accordance to the query sent by the user, the method of similarity measurement, which comprises on the indices, is used. In reference to the feedback, the repossessed results have been optimized. Bellow similarity matching and the feedback relevance type queries have been reassessed. In the following, we review query types, similarity matching, and relevance feedback.

7.1. Query types*:* Those video queries which are non-semantic based, they are for instance: query by objects and query by subject. Those types of video queries which are semantic based are: query by natural language and by keywords [5][21].

7.2. Similarity measurements: Video similarity measurement plays an important role in content-based video retrieval. Measurement methods of video similarities can be classified into text matching, feature matching, combination-based matching and ontology-based matching. The choice of method depends on the query type [4].

7.3. Relevance Feedback: In this approach, those videos which are gathered in reply to the searched queries are graded automatically or by the user. For the refinement of the advance searches this ranking has been used. This method of refinement involves the optimization of the query point, weight adjustment feature, as well as embedding of information. The distance among the semantic notions of low level video content representation and search relevance, are decreased by the relevance feedback. Like relevance feedback for video retrieval, relevance feedback for image retrieval can be divided into three categories: implicit, explicit, and pseudo feedback [22].

CONCLUSION

We have presented a review on recent developments in visual content-based video indexing and retrieval. The state of the art of existing approaches in each major issue has been described with the focus on the following tasks: video structure analysis including shot boundary detection, key frame extraction and scene segmentation, features extraction of static key frames, objects and motions, video annotation, query type and video retrieval methods, video search including interface, similarity measure and relevance feedback.

5. ACKNOWLEDGEMENT

We would like to acknowledge Electronics Engineering Department of Amrutvahini College of Engineering and all the people who provided with the facilities being required and conductive conditions for completion of the review paper.

6. REFERENCES

[1]. Mr. Amit Fegade1, Prof. Vipul Dalal2 “A Survey on Content Based Video Retrieval,” International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 7 July, 2014 Page No. 7271-7279.

[2]. Shripad A.Bhat, Omkar V.Sardessai, Preetesh P.Kunde and Sarvesh S.Shirodkar,“Overview of Existing Content Based Video Retrieval Systems,” International Journal of Advanced Engineering and Global Technology Vol-2, Issue-2, February 2014

[3]. P. Geetha and Vasumathi Narayanan, “A Survey of Content-Based Video Retrieva,” Journal of Computer Science 4 (6): 474-486, 2008 ISSN 1549-3636

[4]. Weiming Hu, Senior Member, IEEE, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval,”IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6, November 2011.

[5]. Aasif Ansari, Muzammil H Mohammed,“ Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges,” International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 7, February 2015

[6]. Dipika H Patel,“ Content based Video Retrieval: A Survey,” International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 13, January 2015

[7]. Muhammad Nabeel Asghar, Fiaz Hussain, Rob Manton,“Video Indexing: A survey,” International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 01, January 2014

[8]. Shweta Ghodeswar1, B.B. Meshram2 “Content Based Video Retrieval,”

[9]. Prof. Priyadarshani A. Bandgar,“Feature Extraction in Content based Video Retrieval,” International Journal of Latest Trends in Engineering and Technology (IJLTET) Vol. 3 Issue 3 January 2014.

[10]. Dr. H.B. Kekre, Dr. Dhirendra Mishra, Ms. P. R. Rege,“ Survey on Recent Techniques in Content Based Video Retrieval International Journal of Engineering and Technical Research,” (IJETR) ISSN: 2321-0869, Volume-3, Issue-5, May 2015.

[11]. Shangfei Wang, Member, IEEE and Qiang Ji, Fellow, IEEE, “Video Affective Content Analysis: A Survey of State-of-the-Art Methods,” IEEE Transactions On Affective Computing, Vol. 6, No. 4,October-December 2015.

[12]. Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan Thomas, HuangAvideh Zakhor, “Applications of Video-Content Analysis and Retrieval,” IEEE MultiMedia July–September 2002.

[13]. Hyun Sung Chang, Member, IEEE, Sanghoon Sull, Member, IEEE, and Sang Uk Lee, Senior Member, IEEE,“Efficient Video Indexing Scheme for Content-Based Retrieval,” Ieee Transactions on Circuits And Systems For Video Technology, Vol. 9, No. 8, December 1999.

[14]. Y. Alp Aslandogan and Clement T. Yu, Senior Member, IEEE,“Techniques and Systems for Image and Video Retrieval,” IEEE Transactions on Knowledge And Data Engineering, Vol. 11, No. 1, January/February 1999.

[15]. Muhammad Nabeel Asghar, Fiaz Hussain, “Rob Manton Video Indexing: A Survey,” International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 01, January 2014.

[16]. Lu Weigan, Weng Wengan,“The Reaserch on Video Retrieval Based Content for Industrial Application,” Sciverse ScienceDirect IERI Procedia 3 (2012) 148 – 155.

[17]. Dipali Patil#1, Mrs. M. A. Potey*2 “Survey of Content Based Lecture Video Retrieval,” International Journal of Computer Trends and Technology (IJCTT) – Volume 19 Number 1 – Jan 2015.

[18]. Bhagwant B. Handge1, Prof. N.R.Wankhade2, “Survey: Retrieval of Video Using Content (Speech &Text) Information,” Bhagwant B Handge et al, Int.J.Computer Technology & Applications,Vol 5 (6),1939-1944.

[19]. Miss. Poonam V. Patil 1, Prof. S. V. Bodake,“Video Retrieval by Extracting Key Frames in CBVR System,” Innovative Research in Science,Engineering and Technology Vol. 4, Issue 12, December 2015.

[20]. T.N.Shanmugam And Priya Rajendran,“ An Enhanced Content-Based Video Retrieval System Based On Query Clip,” International Journal of Research and Reviews in Applied Sciences ISSN: 2076-734X, EISSN: 2076-7366 Volume 1, Issue 3(December 2009).

[21]. D.Saravanan,“Segment Based Indexing Technique For Video Data File,” Science Direct Procedia Computer Science 87 (2016) 12 – 17.

[22]. Simon Jones, Ling Shao,“Content-based retrieval of human actions from realistic video databases,” Information Sciences 236 (2013) 56–65.