International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 10 (2017) pp. 2221-2231
© Research India Publications. https://dx.doi.org/10.37622/IJAER/12.10.2017.2221-2231

Survey on Video Big Data: Analysis Methods and Applications

P.Ushapreethi[1*]
School of Information Technology and Engineering,
VIT University, Vellore, India

[1]ORCID: 0000-0002-3743-0907

G.G. Lakshmipriya [2]
School of Information Technology and Engineering,
VIT University, Vellore, India

Abstract

In recent years, Big Data has become high-focus of researchers as many organizations have been processing huge amount of information. We face a diversity of datasets from different sources in different domains such as Flickr, Twitter, YouTube, Facebook and other local data centers. In this big data era, Video data face its new challenges to store and retrieve, since the size of video data is high and the processing and analysis methods are very complex. Over past years, various ideas and techniques have been proposed towards the effective storage and retrieval of video contents. This paper summarizes the description of the background to common approaches in visual content-based big video storage and retrieval. It also provides a systematic classification of most recent technical works on feature representation, video analysis methods such as video segmentation/video content structuring, video summarization, video abstraction, video indexing, and finally, future challenges and opportunities of big video data. The underlying components for each approach are identified and the details on how they are addressed in specific works were discussed.

Keywords: Feature, Video Segmentation, Video Abstraction, Video Indexing, Video Retrieval, Video Summarization, Video Annotation, Video Content Structuring, Video Categorization and Video Big Data.

INTRODUCTION

In this big data era, Video big data storage and retrieval has its own particularities. Nowadays, a distributed scheme is followed by most of the multimedia environments to reduce the maintenance of multimedia data. But, the multimedia content, especially video content is rapidly increasing due to the highly equipped commercial computers and their usages. Even though many applications and services are available for editing and processing the video data before storing them, the efficiency in content based video retrieval is not yet achieved. In general, big data raises issues of Volume, Velocity, Variety, and Value in short 4V's. More focused view of 4V's in video environment is discussed in the following subsections.

Variety: The variety in big video surveillance data is ensured since different capturing devices such as organizational cameras and wild cameras are used for collecting the real world videos. Most of the fraud detection system, are processing and analyzing variety of surveillance data from distinct cameras to detect the related people, vehicles, or things. The variety of video surveillance devices makes storage and maintenance of distributed video surveillance data the big challenge [1].

Volume: With the rapid development of the surveillance devices, the volume of video surveillance data becomes the big data. For example, NASA's space crafts have sent 1.73GB (gigabytes) of streaming data approximately [2]. Another well-known big challenge is the processing and analyzing video surveillance data with huge volume.

Velocity: The video capturing devices usually work in 24X7 and the amount of data they process is also high. Collecting the video data is faster compared to processing the data when the device has fast in and out. The high velocity of video capturing devices makes processing and analyzing video surveillance data a big challenge [3]. For example, the velocity of collecting the video surveillance data is much larger than analyzing and processing them.

Value: The value of stored real time video data is very high. For example, security systems under many organizations are identifying infrequent and illegal activities on the working environment with the help of surveillance video data. In countries like china the research works are based the traffic surveillance system. These research works are utilized for detecting illegal vehicles or people to ensure the countries safety. On the other hands, the huge volume introduces the challenges for knowledge mining from the video surveillance data.

In this big data era, Video big data storage and retrieval has its own particularities. Nowadays, a distributed scheme is followed by most of the multimedia environments to reduce the maintenance of multimedia data. But, the multimedia content, especially video content is rapidly increasing due to the highly equipped commercial computers and their usages.

Even though many applications and services are available for editing and processing the video data before storing them, the efficiency in content based video retrieval is not yet achieved. In general, big data raises issues of Volume, Velocity, Variety, and Value in short 4V's. More focused view of 4V's in video environment is discussed in the following subsections.

Variety: The variety in big video surveillance data is ensured since different capturing devices such as organizational cameras and wild cameras are used for collecting the real world videos. Most of the fraud detection system, are processing and analyzing variety of surveillance data from distinct cameras to detect the related people, vehicles, or things. The variety of video surveillance devices makes storage and maintenance of distributed video surveillance data the big challenge [1].

Volume: With the rapid development of the surveillance devices, the volume of video surveillance data becomes the big data. For example, NASA's space crafts have sent 1.73GB (gigabytes) of streaming data approximately [2]. Another well-known big challenge is the processing and analyzing video surveillance data with huge volume.

Velocity: The video capturing devices usually work in 24X7 and the amount of data they process is also high. Collecting the video data is faster compared to processing the data when the device has fast in and out. The high velocity of video capturing devices makes processing and analyzing video surveillance data a big challenge [3]. For example, the velocity of collecting the video surveillance data is much larger than analyzing and processing them.

Value: The value of stored real time video data is very high. For example, security systems under many organizations are identifying infrequent and illegal activities on the working environment with the help of surveillance video data. In countries like china the research works are based the traffic surveillance system. These research works are utilized for detecting illegal vehicles or people to ensure the countries safety. On the other hands, the huge volume introduces the challenges for knowledge mining from the video surveillance data.

Meng Wang and Hong-Jiang Zhang provided the details on hierarchical decomposition and representation of video content [4] which is shown in Fig. 1. It visualizes the origin of features, feature extraction, methods of video analysis and application of base methods. Since features are the base for all video analysis methods and applications, study on features is essential. This paper describes various types of features and their usage in various recent research works. Feature representation is the process of extracting features and reduction in the feature dimension to utilize them in the video analysis methods and applications. Video analysis methods are used for tracking the objects and events in such important applications as robot vision, object-based auto-focusing, activity recognition, and intelligent surveillance systems.

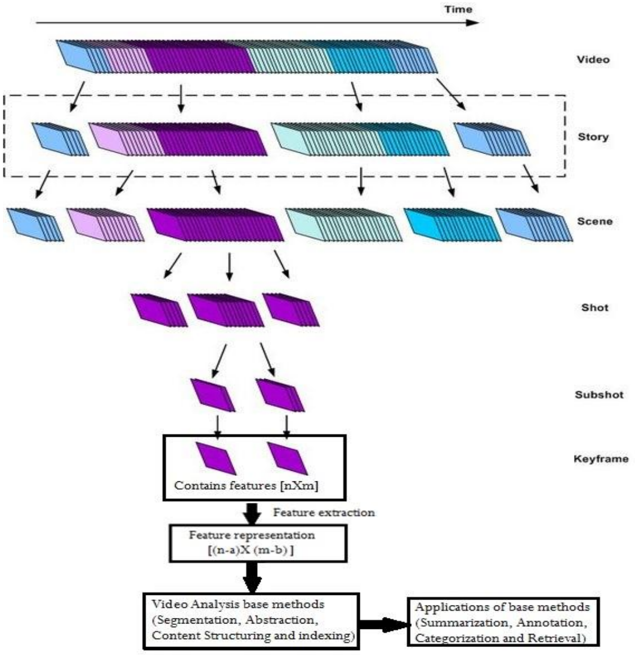
Time
Video
Story
Scene
Shot
Subshot
Keyframe
Contains features [nXm]
Feature extraction
Feature representation
[(n-a)X (m-b)]
Video Analysis base methods
(Segmentation, Abstraction,
Content Structuring and indexing)
Applications of base methods
(Summarization, Annotation,
Categorization and Retrieval)

Figure 1. Features – base for video analysis process.

FEATURES CLASSIFICATION

In this digital era, capturing and sharing videos happens very often in our day-to-day life. Even people without the knowledge of video have their own media devices and increase the amount of daily media content. The data in/out to the big data system can be in different formats, such as text, image, audio or video. Proper feature identification and representation methods are needed for effective storage and retrieval of video data since several types of features are identified by the researchers.

The prominent features provide the successful video categorization systems, which recognize various video categories, eventhough the video is multimodal and it has various intra-domain variations. Video is multimodal, because it contains both visual and acoustic channels for capturing visual and audio information respectively. For effective video categorization these two channels must be interdependent and should be used together. In recent years, several features namely static features, motion features, acoustic features and semantic features have been identified and the same are discussed in the following section.

A. Low level features

Static Features: The video is divided into frames and shots to extract the static features. Static features are used by most of the image and video analysis techniques since it needs low computational cost and provides comparably high performance in many applications.

Color information is the basic feature used in earlier video analysis works [5 -8]. Authors of [5] utilized textural measures along with color feature based on covariance and provides a new feature named Color Wavelet Covariance (CWC). The specificity and sensitivity were estimated for color colonoscopic videos using a Linear Discriminant Analysis (LDA) and the performance is high for that specific dataset.

Efficient object segmentation is achieved in [6] using edge information. Self-Organizing Feature Map (SOFM) neural network is used for transforming multiple feature space to one-dimensional label space.

An edge fusion process is deployed for incorporating edge information with SOFM neural network. Histogram of Orientated Gradients (HOG) [7] is well known among numerous existing image features. Several other features are also widely used. In addition to that Scalable Invariant Feature Transform (SIFT) and color SIFT [8] are also the most popular for modelling the color information.

Recently, dominant color features are taken algorithm for video cut detection in [9].This work is based on block based histogram differences and the performance is comparably high with existing color feature methods.

Convolutional Neural Network (CNN) based models are also used as static appearance features for video analysis. Frame-level features of ImageNet [10] dataset are extracted from proposed CNN model for action recognition and the performance is reported as comparably high [11].

Static appearance features does dot incorporate any temporal information in the video. So, the value of such features is considerably less in activity recognition and moving object recognition. Due to these inconveniences of static features, the attraction towards motion features was increased in video analysis and the same is discussed in the next section

Motion Features: Motion is the key feature incorporates temporal information and object movements of videos. Different from the static appearance features such as color, texture, etc. motion features are more objective and consistent. Because of this motion, features possess high value for understanding object actions and complex events. Apart from static features, the motion features needs to be extracted efficiently to obtain content-based video processing. Even in most of the research works [12, 13] manual motion feature extraction has been taken and ontology model is used for motion feature representation. This approach is compared with dense trajectories and competitive results are exhibited. In [14] motion features are used with ontology representation, however, event extraction is manual.

3D spatial-temporal space is used to get the motion features by extending the frame based image features. Space-Time Interest Points (STIP) is an extension of Harris corner patch detector which provides extracted motion key features but the pixel values vary in both space and time dimensions [15]. Tracking densely sampled patches provides dense trajectory features by computing Local descriptors among densely sampled patches. This feature has been popular and provided promising performance compared to all the popular benchmarks [16]. CNN model for temporal features of videos is proposed and utilized for motion features where stacked optical flow images are taken as inputs instead of video frames [17]. High level features occupies much space comparably low level features, hence the retrieval time gets increased. Acoustic Features: Some researchers [18, 19] made a try with a different feature known as acoustic features, since it is comparably easy to extract and provides valuable and highly complementary information. Instead of visual features, spectral form of an audio signal is utilized and named as Mel-Frequency Cepstral Coefficients (MFCC) descriptors [18]. MFCC descriptors are encoded with the standard straight forward approach Bag-Of-Words (BoW) representation and achieved top-notch performance [19]. Similar representations have been used in recent works but the issue arises when the complexity increases to recognize natural and artificial sounds, hence acoustic features are considered in any recent video analysis researches.

B. High Level Features

Semantic features: A semantic feature is a method for expressing the pre-established semantic properties of the object. Both low-level features and additional knowledge of a specific domain are integrated in high-level semantic features. Because of the compactness of high-level semantic features the video retrieval is fast, easy and more intuitive. Efficient techniques for video analysis according to their high-level semantics are most necessary in many applications. The researches related to video analysis and semantic features are discussed in this section.

Semantic features are used by BilVideo [20], extended-AVIS [21], multiView [22] and classView [23] but no ontology-based models is used for extracting semantic features. A Video Event Recognition Language (VERL) and Video Event Markup Language (VEML) are proposed with semantic features [12]. Domain based ontology designing is achieved by VERL and manual annotation for VERL videos are obtained by Video Event Markup Language (VEML). The manual annotation and lack of low level processing are the identified drawbacks of this study. WordNet consists of Semantic hierarchies, which are used to obtain visual appearance learning by studying the inter-class relationships among objects and events [24]. A systematic approach to address the problem of designing ontologies for semantic feature extraction is presented in [25]. Semantic relations between visual symbols in the key frames are some researchers' attention. [26] presented a framework that generates text descriptions of image and video content based on image understanding namely Image parsing to Text description (I2T).

In paper [27] the author proposed a novel semantic-based heterogeneous Transportation Media Retrieval (TMR). TMR supports the function of retrieving different media types such as image, video, audio and text. The semantic fields are extracted from the user annotation and both media document data and semantic information are stored. The information is retrieved based on the query using the semantics and documentation. Semantic features are utilized for their effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval [28]. They follow below steps and provided an appealing performance in video retrieval. Initially, Heterogeneous multimedia retrieval is done followed by Extracting and representing semantic information for heterogeneous multimedia. NoSQL-based approach to semantic storage and a MapReduce-based retrieval algorithm is used. A semantic based model is proposed for representing and organizing video big data using a video structural description technology which includes the domain knowledge on Computer vision, Semantic web, Semantic link network, Cloud computing [1]. Incremental probabilistic Latent Semantic Analysis for video retrieval is very recent approach in semantic feature representation. They have compared their work with several other benchmarks like pLSA, LDA, FSTM. [29]

According to fig. 1 features are the base for all video analysis methods and applications. We have discussed various research works based on different features. After choosing the specific feature and feature representation method, it can be applied to video analysis methods for processing the video big data. Next section explains recent researches utilizing various video analysis methods

VIDEO ANALYSIS – BASE METHODS

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Video (temporal) Segmentation

Video (temporal) segmentation is the process of partitioning a video sequence into disjoint sets of consecutive frames that are distinct according to particular criteria. The video is partitioned into shots, camera-takes, or scenes in the most common types of segmentation. A camera take is a sequence of frames captured by a video camera during its start and respective stop moment. Shot is a continuous sequence of frames belonging to a single camera take in an edited video [30]. In paper [31], the lower level features L0 gradient is taken for video segmentation. The input video is converted into frames obviously. Then the image patches are over segmented and output descriptors are generated for the image patches. Gradients are identified for this output descriptors and they are minimised. The advantage of this approach is the 3D output video representation which is generated by applying fused co-ordinate descent algorithm on the 2D images. The major steps of [31] are explained in fig. 2.

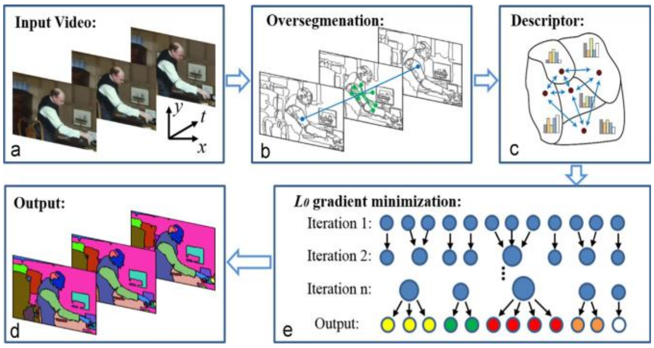

Figure 2. Steps in video segmentation under L0 minimization[31]

An efficient approach to foreground extraction is proposed in [3] using spatio-temporal decorrelated block features. The specific advantage of this approach is it uses compressed images in the video. So, storage and computational resources are not needed for all the stages in the foreground extraction.

Apart from other researchers [32] studied the segmentation for the video with and without illumination changes. Gaussian mixture models are used to capture the lower level features of the frames. Then self- organised maps were created based on the features. Illumination sensitive method is the advantage of this approach. The types of cameras used in this approach hemispheric and normal camera with and without light effects. The basic structure of the segmentation process is described in figure 3.

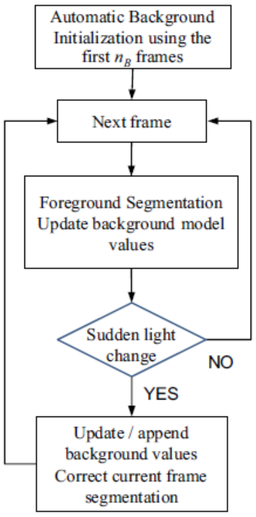
Automatic Background Initialization using the first n_B frames
Next frame
Foreground Segmentation Update background model values
Sudden light change
NO
YES
Update / append background values Correct current frame segmentation

Figure 3. Basic structure of the segmentation process [32]

A new Auto-Adaptive Parallel SOM Architecture (AAPSA) is developed based on Self Organized Maps (SOM), shown in fig. 4 [33].

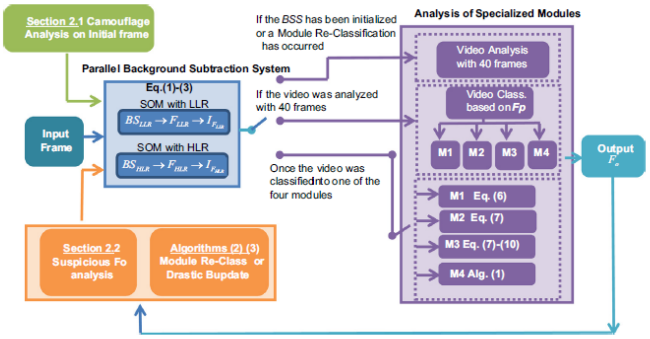
Section 2.1 Camouflage Analysis on Initial frame
If the BSS has been initialized or a Module Re-Classification has occurred
Analysis of Specialized Modules
Parallel Background Subtraction System
Video Analysis with 40 frames
Eq.(1)-(3)
SOM with LLR
$BS_{LLR} \rightarrow F_{LLR} \rightarrow I_{F_{LLR}}$
If the video was analyzed with 40 frames
Video Class. based on Fp
Input Frame
SOM with HLR
$BS_{HLR} \rightarrow F_{HLR} \rightarrow I_{F_{HLR}}$
M1
M2
M3
M4
Output F_o
Once the video was classifiedinto one of the four modules
M1 Eq. (6)
M2 Eq. (7)
M3 Eq. (7)-(10)
M4 Alg. (1)
Section 2.2 Suspicious Fo analysis
Algorithms (2) (3) Module Re-Class or Drastic Bupdate

Figure 4. AAPSA Architecture[33]

In this work also the SOM are created, but the difference between the work in [32] and this work is initial frame monitoring and severe foreground analysis to achieve minimum false positive rates. Moreover commerce approach produced the best segmentation results for both static and dynamic objects compared to all state of art models.In paper [34] the authors developed an efficient method for figure-ground segmentation using Feature Relevance (FR) and active contours. This work needs it training phase for computing foreground segmentation and figure ground separation.

The objects of interest are calculated based on FR and contour information. Appealing performance is achieved through this method but the performance is based on the training set. Steps for training phase and segmentation phase for both image and video frames are clearly organised in fig. 5.

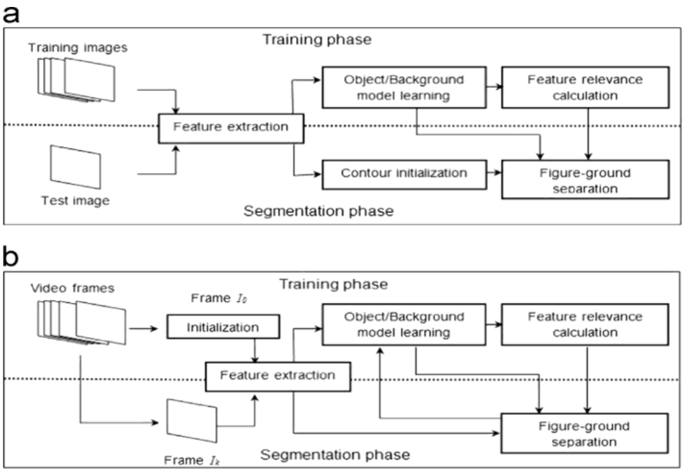

Figure 5. Two basic models of the segmentation process [34]

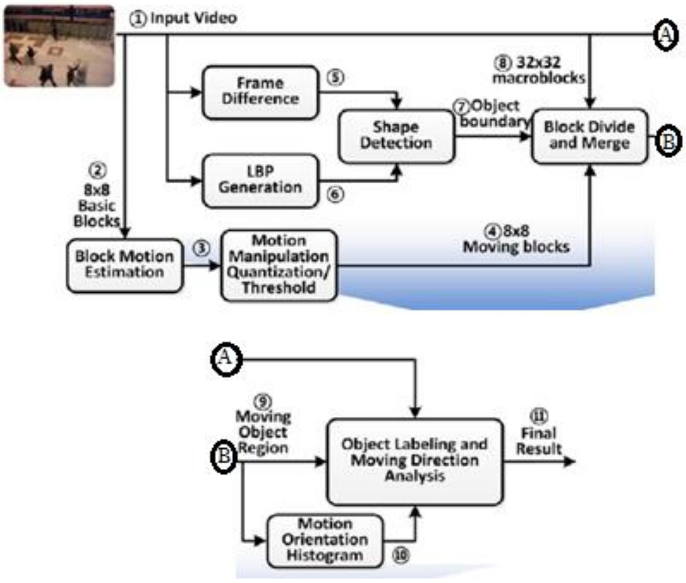

Figure 6. Model for moving object segmentation [35]

A moving object segmentation method is presented in [35] Objects were identified by capturing their block motion vectors and their orientation. Motion manipulation (quantization) was done after identifying the direction of the objects using histograms. The moving objects are represented optimally using inbuilt chip/software. The advantage of this approach is the easy and efficient computation structure for object labelling; In addition, embedding the chip with labelling software is also very easy. The computational structure is shown in fig. 6.

B. Video Content Structuring

Video content structuring is the process of dividing the videos into smaller units. The smaller units are further divided until we get frames/key frames. The major purpose of this video decomposition is to index the lengthy and unstructured videos. The contents are properly structured based on video indexing which leads to efficient content based video processing and retrieval. Generally a video can be structured as the following hierarchical form.

"videos->stories->scenes -> shots->subshots->keyframes"

A hierarchical video summarization is proposed in [37]. The video is converted into shots and key frames using affinity matrix. Both static and dynamic summaries are generated using video content structuring.

Recently, a fully automatic system for extracting the semantic structure is presented in [38]. Academic presentation videos are captured on stage with camera motions such as panning, tilting, and zooming. The advantage of this system is that it keeps track of both projection screen and the presenter and provides efficient summarization than general video summarization methods. But the semantic structure is created only for lecture videos, since it is not a generic approach.

Moreover, video content structuring is the well-known video processing method for the application, video summarization. The summarization produced by using video content structuring is majorly used for fast and efficient video retrieval.

C. Video Abstraction

Video analysis applications are demanded nowadays due to the recent advancements in video technologies and its widespread usage in distributed network infrastructure. A top-notch research area, named as video abstraction is introduced for the efficient and user-friendly storage - retrieval of video contents. The video digest (video abstract) is obtained from the closed-caption information residing in the videos [40]. This method segments the videos and provides the appropriate video abstracts automatically by utilizing low level features. The user will gain the searched video sequence in a specified time constraint during browsing and navigation with the help of good video abstract.

In general, video summary and video skimming are known as two types of video abstraction [41]. Video summary provides the key frame images and skimming provides the prominent images along with the audio information. Video skimming is further classified into two sub-types namely, highlight and summary sequence. Highlight consists of most interest regions of the video and summary sequence provides the short information about whole skimming video.

Authors of [42] introduced a new mixture model for generating video abstraction. The principle method of this work is relational graph representation. The basic steps of this approach are conversion of video into frames, identification of prominent feature vector for each frame, conversion of feature vectors into graphs, partitioning the graphs into connected sub graphs, dimensionality reduction of the data set, applying the mixture model to generate the automatic video abstraction. Regularization fusion object tracking is introduced in [43] objects are identified by utilizing the kernalised confidence and object motion trace regularizes were utilized for label assignments on real time video object motion features are identified using circulant Matrix and the identified objects are presented using graph based representation the advantage of this method is its usage in real time environment without any code Optimization.

Abstraction is a short representation of an original video using the features identified in the video, and widely used in video annotation, summarization, and retrieving. Instead of having unordered group of information for retrieval, ordered information leads to an efficient and effective retrieval. This induced the motive for considering the video analysis method called video indexing and the same is discussed in the next section.

D. Video Indexing

Indexing is the process of sorting the video content for dynamic and efficient video storage and retrieval. Conventional video indexing methods are utilised for creating video indexing guidelines for the data set CORPORA [44]. Both static background region identification and dynamic motion of objects are considered for video indexing. 3 factors such as viewer response, variations among viewers and data set creation are best for the effective indexing in this work.

Recent layered video indexing method is presented in [45]. The author created an innovative mobile video search system by developing an efficient video indexing method the work is well advanced because the search system enables the cloud storage access with the help of audio - video signatures. This method considers user experience and improves it by search accuracy and Lo retrieval latency.

A different manual indexing method is proposed in 46. Likelihood between the low level features Bag Of Words (BOW) and Support Vector Machines (SVM) models are utilised for identifying the prominent regions of the video. The imitations of humans, such as hand movements are taken as major key motion activities and the action is predicted and validated with behavioural and neural levels. These action details are utilized for indexing the social media interaction videos.

Widely used applications of the base video analysis methods are video summarization, annotation and categorization. The next section discussed the recent researches in aforementioned applications.

APPLICATIONS OF BASE VIDEO ANALYSIS METHODS

A. Video Summarization

Video summarisation is the process of creating a summary of digital video. Efficient summary must possess the following principles: 1) the events mentioned in the summary must be key/ prominent events in the video. 2) The continuity of the events should be insured using the summary. 3) The event in the summary should not possess any replica. Two types of summary are physical summary (based on physical property) and semantic summary (based on meaning) [47].

Authors of [48] proposed a basic video summarisation method by utilising both static features and dynamic features. Each shot is extracted on the threshold of the adjacent frames are identified by applying dwt discrete wavelet transform. Pixels of Interest are gathered from the threshold and the high resolution values for the pixels are calculated. The dynamic wavelet features are also gathered and combined with the static features, because static features of dominating in some videos and dynamic features are dominating in other videos. Fig. 7 shows the basic steps followed in [48].

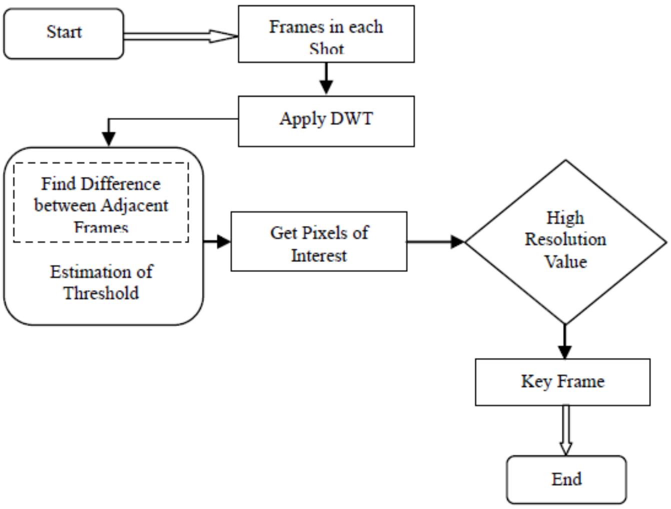

Figure 7. Summarization steps [48]

In papers [49] , authors developed keyframe based approach for video summarisation. Static features, such as mean variance, skew and kurtosis are calculated for each image blocks in the frames. The frame which is having maximum mean and variance selected as keyframe and it is utilised for video summarisation. Semantically meaningful video summaries are generated in [50]. Advantage of this method is gathering the user performance for efficient video summarisation. The video information is stored in the database and it is provided for various video analysis systems like open CV, face recognizer, scene change detection and shot boundary detection.

These systems are working on creating metadata for these videos and the same is passed to the database. The metadata is the semantic information which is utilised for video summarisation. Semi-automatic annotation is also provided to the user. Fig. 8 shows the procedural steps followed in [50]. A divide-and-conquer based framework for an efficient summarization of big video data is proposed in [51]. The original video is divided into shots. The advancement in this approach is that the viewers neuronal signals are also captured for analysing the key frames(prominent regions). The aural and visual information in the video is processed by Teager energy and motion intensity respectively. The information collected using all the above methods are combined to create new aggregate summary for the given video.

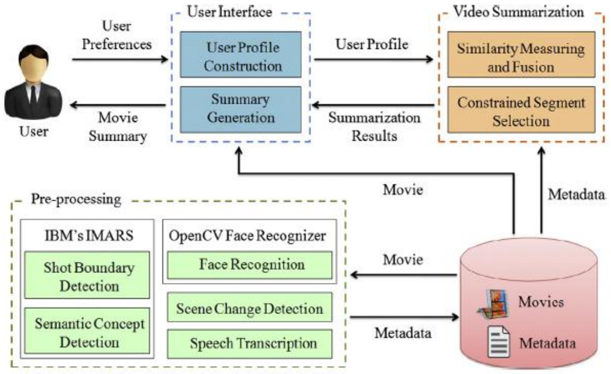

Figure 8. Semantically meaningful video summary [50]

This method outperforms than all other video summarisation methods. Fig. 9 shows the summarization process of the video. Graph-based hierarchical clustering method is used for computing video summary [52].

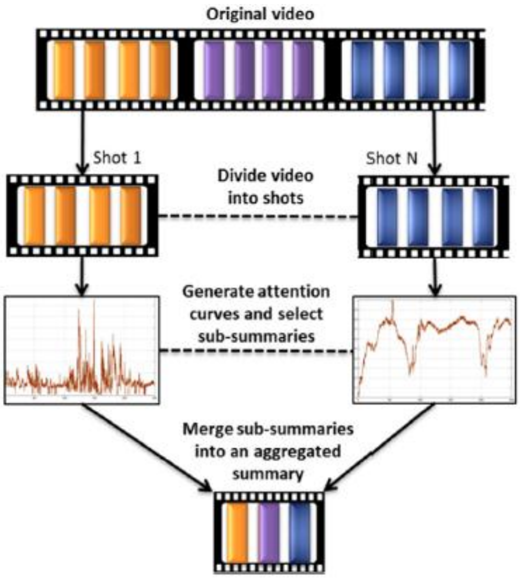
Original video
Shot 1
Divide video into shots
Shot N
Generate attention curves and select sub-summaries
Merge sub-summaries into an aggregated summary

Figure 9. Summarization process of a video [51]

Initially, videos divided into video frames and they are group as clusters. The clusters are considered as connected components of a graph. Minimum spanning tree is calculated for the graphs and the summary is insured based on weight maps. Furthermore user preferences are also taken for producing the summary. These two summaries are compared and analysed to get the better summary. The steps are shown in Fig. 10.

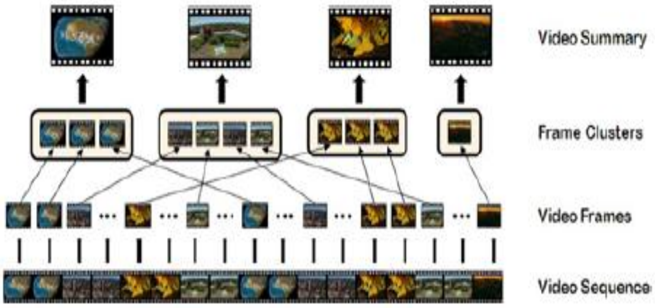

Figure 10. Video Summarization steps [52]

Low level feature is identified by [53], namely heterogeneity image patch (HIP) index. In all the images, different image patches are presented. HIP Curve for all frames' HIP indices are identified for a video using the entropy based method. The summary is generated based on the HIP Curve and the advantage of this method is its lowest complexity.

COMPARISIONS

The features, Feature representation methods and Video analysis methods for various recent research works are mentioned in Table 1. Variants of Static features such as HOG, HIP, Spare feature points are majorly used by most of the recent works. Motion histograms, Motion descriptors are some of the features utilized by several recent research works. But, Semantic feature is majorly used for big video analysis work, since it provides the generic representation, which leads to fast and efficient data retrieval.

The video analysis methods are taken based on the domain and application of the work. In most of the recent papers related to big video data, a specific framework is designed for efficient storage and retrieval. The user feedback and satisfaction analysis are also taken for automatic semantic updation. Even though, semi-automatic semantic annotation and automatic updation are possible in big video retrieval, the works are limited to particular domain such as transportation, web movies, web news, etc.

Based on the wide literature survey, we can say that there are very less works were done on big video storage and retrieval. So, there are plenty of opportunities available in this field such as dimensionality reduction in features, various feature representation methods and video analysis methods. An efficient storage and retrieval framework for big video data can be designed by combining the aforementioned methods in future.

CHALLENGES AND OPPORTUNITIES

Although several progresses have been made in the past years, the current video analysis techniques for big video storage and retrieval are not satisfactory. The major challenge is selecting the features such as mathematical low-level features and the high-level semantic features.

TABLE 1: COMPARISON OF VARIOUS RESEARCH WORKS

Reference	Features used	Feature Representation method	Video analysis method/application used	Applied on Big data? (Yes/No)
[11]	Static feature -Color , HOG	CNN based representation	Video Categorization	No
[17]	Motion feature	CNN based representation		No
[44]	Motion feature- Audio and Visual	Circular and One Dimensional Model	Video Indexing, Video Annotation	No
[45]	Motion features	audio-video descriptors	Video Indexing	Yes
[53]	Static feature - HIP index	Affinity matrix	Video Summarization	No
[1]	Semantic feature	Semantic Based Representation Model – Video Structural Description Technology	Video Annotation	Yes
[3]	Motion feature-block features	Block coding units	Video Segmentation	Yes
[27]	Semantic feature	Ontology	Video Annotation	Yes
[28]	Semantic and Motion feature	Ontology	Video Annotation	Yes
[29]	Semantic feature	Topic model	Video Categorization	No
[34]	Static feature	Feature Relevance (FR) and active contours	Video Segmentation	No
[35]	Motion Feature- Motion Orientation Histogram	Adaptively Partitioned Block Representation	Video Segmentation	No
[39]	Static feature - Sparse Feature Points	Feature Trajectory Labelling	Video Content Structuring	No
[46]	Static feature	BOW and 1-Class SVM models	video indexing	No
[48]	Static and Motion feature	Wavelet features	Video Summarization	No
[49]	Static feature - Mean, Variance, Skew and Kurtosis Histogram	Image - block based Representation	Video Summarization	No
[50]	Semantic feature	Cosine Similarity Metric	Video Summarization, Semi-Automatic Annotation	Yes
[31]	Motion feature - Motion Descriptor	L0 gradient minimization	Video Segmentation	No
[32]	Static feature – pixels	Gaussian mixture models, the Self Organizing Maps	Video Segmentation	No
[33]	Motion feature – IP(Interest Points) and its descriptor	Self-Organized Maps (SOM)	Video Segmentation	No
[36]	Static feature –pixel	Shortest Path Algorithm	video segmentation	No
[51]	Motion feature – Audio and Video	Teager energy, instant amplitude, and instant frequency, multi-scale contrast and motion intensity	Video Summarization	Yes
[52]	Static feature - Global Descriptors – Color Histogram	Bag Of Features (BOF)	Video Summarization	No
[43]	Static feature- Gray Scale Intensity	Guassian Representation	Video Abstraction	Yes

If features are considered, most of the recent big video analysis techniques are based on temporal features of a specific domain. The features are extracted and represented with respect to the analysis method and the application used. Therefore, we envision that a new generic feature identification method is needed for all video analysis process.

But, for videos, there is no impressive performance reported using this approach. Designing a generic architecture for big video retrieval with suitable feature selection, feature representation and video analysis method is needed in this big data era. First, various neural network architectures were designed for big video analysis, but generic and common neural network architecture is not yet achieved; since videos have specific spatial-temporal characteristics. In addition, the architectures are based on sample data for their training phase, but samples are not enough to achieve architecture for analysing large amount of real time video data.

Second, various semantic based architectures were designed using ontology-based representation with recent advantages like the user preferences for better user satisfaction. The deep learning approaches such has semantic representation, semantic feature extraction are becoming popular and may achieve a big hike on the video analysis performance by developing new semantic learning approaches that are suitable for big video storage and retrieval. Semantic learning has exhibited inspiring results on many fields including text analysis and image annotation.

Unfortunately, all of these architectures are domain or application based architectures. Therefore, designing a large, generic and well-defined ontology-based architecture is still a challenging opportunity in this field.

Moreover, results on video surveillance data are also much less because of the storage size and redundant data formats. In summary, there are two areas need to be investigated extensively. They are generic feature identification methods and generic semantic architecture design for big video analysis. The former is the base for all video analysis methods which leads to effective big video storage and retrieval. The later demands a smart design for achieving good performance in real-time applications.

REFERENCES

[1] Z. Xu, , Y. Liu, L. Mei, C. Hu, and L. Chen, “Semantic based representing and organizing surveillance big data using video structural description technology,” Journal of Systems and Software, vol. 102, pp. 217–225, Apr. 2015.

[2] N. Skytland, Big data: What is NASA doing with big data today?, OpenNASA. Oct. 2012.

[3] B. Dey and M. K. Kundu, “Efficient Foreground Extraction From HEVC Compressed Video for Application to Real-Time Analysis of Surveillance ‘Big’ Data,” IEEE Trans. on Image Processing, vol. 24, no. 11, pp. 3574 – 3585, Nov. 2015.

[4] M. Wang and H. Zhang, “Video Content Structuring,” Scholarpedia, Vol. 4, no. 8, pp.9431, Aug.2009.

[5] S.A. Karkanis, D.K. Iakovidis, D.E. Maroulis, D.A.Karras and M.Tzivras, “Computer-aided tumor detection in endoscopic video using color wavelet features,” IEEE Trans. on Info. Tech. in Biomedicine, 2003, vol. 7, no. 3, pp. 141 - 152, Sept. 2003.

[6] J.Kim, and T.Chen, “Combining static and dynamic features using neural networks and edge fusion for video object extraction, Vision, Image and Signal Processing,” IEE Proceedings - Vision, Image and Signal Processing, vol.150 , no.3, pp. 160 – 167, Jun. 2003.

[7] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 2005, Vol. 1, pp. 886 - 893.

[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. Journal of Computer Vision, Vol 60, no 2 , pp 91-110, Nov.2004

[9] G. G. LakshmiPriya, and S. Domnic, “Video Cut Detection using Dominant Color Features” in Proc. of the First Int. Conf. on Intelligent Interactive Technologies and Multimedia, New York, NY, USA, 2010, pp. 130-134.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. F. Fei, “Image Net: A large-scale hierarchical image database,” in proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009, pp. 248 - 255.

[11] M. Jain, J. Van Gemert, and C. G. M. Snoek, “University of Amsterdam at thumos challenge 2014,” 2014.

[12] R. Nevatia, J.Hobbs, R.Bolles, and J.Smith, “VERL: an ontology framework for rep- resenting and annotating video events,” IEEE Multimedia, Vol.12 issue-4, pp.76–86, oct. 2005.

[13] A. Bagdanov, M.Bertini, A. Del Bimbo, C.Torniai, G.Serra, “Semantic annotation and retrieval of video events using multimedia ontologies,” in Proc. of IEEE International Conference on Semantic Computing, Irvine, CA, 2007, pp. 713 – 720.

[14] L. Bai, S. Lao, G. Jones, and A. Smeaton, “Video semantic content analysis based on ontology,” in Proc. of the 11th Int. Machine Vision and Image Processing Conference, pp. 117–12, 2007.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, 2008, pp. 1-8.

[16] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in proc. of IEEE International Conference on Computer Vision (ICCV), Sydney, NSW, 2013 pp. 3551 – 3558.

[17] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. of Twenty-eighth Annual Conference on Neural Information Processing Systems (NIPS), Montreal, CA, 2014.

[18] S. Umesh, and R. Sinha, “A Study of Filter Bank Smoothing in MFCC Features for Recognition of Children's Speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp.2418 – 2430, Nov. 2007

[19] Y.G. Jiang, X. Zeng, G.Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.F. Chang, “Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching,” in TREC video Retrieval Evaluation workshop 2010, TRECVid, 2010.

[20] M. Donderler, E. Saykol, U. Arslan, O. Ulusoy, and U. Gudukbay, “Bilvideo: design and implementation of a video database management system,” Multimedia Tools and Applications, vol.27, no. 1, pp.79–104, Sept.2005.

[21] T. Sevilmis, M.Bastan, U.Gudukbay, O.Ulusoy, “Automatic detection of salient objects and spatial relations in videos for a video database system,” Image and Vision Computing Vol.26 no. 10, pp. 1384–1396, oct. 2008.

[22] J. Fan, W.Aref, A.Elmagarmid, M.Hacid, M.Marzouk, and X.Zhu, “Multiview: multilevel video content representation and retrieval,” Journal of Electronic Imaging, vol. 10, no. 4, pp. 895–908, Oct. 2001.

[23] J. Fan, A. Elmagarmid, X. Zhu, W. Aref, and L.Wu, “Classview: hierarchical video shot classification, indexing, and accessing,” IEEE Transactions on Multimedia, vol.6, no. 1, pp. 70–86, Feb. 2004.

[24] M. Marszalek, C. Schmid, and M.Inria, “Semantic hierarchies for visual object recognition,” in Proc. . IEEE Conference on of Computer Vision and Pattern Recognition, Minneapolis, MN , 2007, pp. 1-7.

[25] U. Akdemir, P.Turaga, R.Chellappa, “An ontology based approach for activity recognition from video,” in Proceedings of the ACM International Conference on Multimedia, pp. 709–712, 2008.

[26] B. Yao, X. Yang, L. Lin, M. Lee, S. Zhu, “I2T: image parsing to text description,” Proc. of IEEE, vol. 98, no. 8, pp.1485–1508, Jun. 2010.

[27] K. Guo, R. Zhang, L. Kuang, “TMR: Towards an efficient semantic-based heterogeneous Transportation media big data retrieval,” Neurocomputing, submitted for publication. 2015.

[28] K. Guo, W. Pan, M. Lu, X. Zhou, J. Ma, “An effective and economical architecture for semantic based heterogeneous multimedia big data retrieval,” The Journal of Systems and Software, vol.102, pp. 207–216, Apr. 2015.

[29] R. F. Beltran, F. Pla, “Incremental probabilistic Latent Semantic Analysis for video retrieval,” Image and Vision Computing, Vol. 38, no. C, pp. 1-12, Jun. 2015

[30] N. Dimitrova, L. Agnihotri, M. Barbieri, H. Weda, “Reference Work Entry: Video Segmentation,” Encyclopedia of Database Systems, pp 3308-3313, 2009, Available: http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_442.

[31] X. Cheng, Y. Fen, M. Zeng, X. Liu, “Video segmentation with L0 gradient minimization,” Computers &Graphics Vol.54, pp. 38–46, Feb. 2016.

[32] K. K. Delibasis, T. Goudas, I. Maglogiannis, “A novel robust approach for handling illumination changes in video segmentation Engineering Applications of Artificial Intelligence,” Vol.49 pp. 43–60 Mar.2016.

[33] G. R. Alonso , M. I. C. Murguía, “Auto-Adaptive Parallel SOM Architecture with a modular analysis for dynamic object segmentation in videos,” Neuro Computing, Vol. 175, Part B, no. 29, pp. 990–1000, Jan. 2016.

[34] M. S. Allili and D. Ziou, “Likelihood-based feature relevance for figure-ground segmentation in images and videos,” Neuro Computing, Vol. 167, no.1, pp. 658–670, Nov. 2015.

[35] S. Lee, N. Kim, K. Jeong, I. Paek, H. Hong and J. Paik, “Multiple moving object segmentation using motion orientation histogram in adaptively partitioned blocks for high-resolution video surveillance systems,” Optik - International Journal for Light and Electron Optics, Vol. 126, no. 19, pp. 2063–2069, Oct. 2015.

[36] X. Cao, F. Wang, B. Zhang, H. Fu, C. Li, “Unsupervised pixel-level video foreground object segmentation via shortest path algorithm,” Neuro computing Vol. 172, no.8, pp. 235–243, Jan. 2016.

[37] M. Wang and H. Zhang, “Video Content Structuring,” Scholarpedia, Vol. 4 no.8, pp. 9431, 2009. Available: http://www.scholarpedia.org/article/Video_Content_Structuring

[38] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid and W. G. Aref, “Exploring video content structure for hierarchical summarization,” Multimedia Systems, Vol. 10, no. 2, pp. 98–115, Aug. 2004.

[39] K. Li, J. Wang, H. Wang, and Q. Dai, “Structuring Lecture Videos by Automatic Projection Screen Localization and Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, no. 6, pp. 1233 – 1246, Jun 2015.

[40] B. T. Truong and S. Venkatesh, “Video Abstraction: A Systematic Review and Classification,” ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 3, no. 1, Article 3, Feb. 2007.

[41] Y. Li, T. Zhang and D. Tretter, “An overview of video abstraction techniques,” 2001. Available: http://www.hpl.hp.com/techreports/2001/HPL-2001-191.html

[42] S. L. Zhai, B. Luo, J. Tang and C. Y. Zhang, “Video Abstraction Based On Relational Graphs,” in Proc. of Fourth International Conference on Image and Graphics (ICIG 2007), Sichuan, 2007, pp. 827 - 832.

[43] P. Zhang, T. Zhuo, Y. Zhang, L. Xie and D. Tao, “Real-time tracking-by-learning with high-order regularization fusion for big video abstraction,” Signal Processing, submitted for publication.

[44] M. Soleymani, M. Larson, T. Pun and A. Hanjalic, “Corpus Development for Affective Video Indexing,” IEEE Transactions on Multimedia, vol. 16, no. 4, pp. 1075 - 1089, Jun 2014.

[45] W. Liu, T. Mei, and Y. Zhang, “Instant Mobile Video Search with Layered Audio-Video Indexing and Progressive Transmission,” IEEE Transactions on Multimedia, Vol. 16, no. 8, pp. 2242 – 2255, Dec. 2014.

[46] E. Delaherche, G. Dumas, J. Nadel, M. Chetouani, “Automatic measure of imitation during social interaction: A behavioral and hyper scanning-EEG benchmark,” Pattern Recognition Letters Vol. 66 pp. 118–126, Nov. 2015.

[47] “Video Summarization - Summary, Events, Create, and Class - JRank Articles,” Available: http://encyclopedia.jrank.org/articles/pages/6930/Video-Summarization.html

[48] J. Kavitha, Dr. P. A. J. Rani, “Static and Multiresolution Feature Extraction for Video Summarization,” Procedia Computer Science, Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014), Vol. 47 pp. 292 – 300, 2015.

[49] P. S. Jadhava and D. S. Jadhav, “Video Summarization using Higher Order Color Moments,” Procedia Computer Science, International Conference on Advanced Computing Technologies and Applications (ICACTA) Vol.45 pp. 275 – 281, 2015.

[50] R. Kannan, G. Ghinea, S. Swaminathan, “What do you wish to see? A summarization system for movies based on user preferences,” Information Processing and Management, Vol. 51. pp. 286–305, May 2015.

[51] I. Mehmood, M. Sajjad, S. Rho, S. WookBaik, “Divide-and-conquer based summarization framework for extracting affective video content,” Neurocomputing, Vol. 174 Part A, pp. 393–403, Jan. 2016.

[52] L. D. S. Belo, C. A. Caetano, Z. K. G. Patrocínio, and S. J. F. Guimaraes, “Summarizing video sequence using a graph-based hierarchical approach,” Neurocomputing, Vol. 173, pp. 1001–1016, Jan. 2016.

[53] C. T. Dang and H. Radha, “Heterogeneity Image Patch Index and Its Application to Consumer Video Summarization,” IEEE transactions on Image Processing, vol. 23, no. 6, pp. 2704 – 2718, Jun. 2014.