MULTI-VIEW VIDEO SUMMARIZATION USING TARGET DETECTION AND CLUSTERING

A DISSERTATION

SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Atul Aman (2K19/CSE/05)

Under the supervision of

Dr. Anil Singh Parihar

(Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

JUNE, 2021

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

DECLARATION

I, Atul Aman, Roll No. 2K19/CSE/05 student of M.Tech (Computer Science & Engineering), hereby declare that the Project Dissertation titled "**Multiview video summarization using target detection and clustering**" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi. Report of the Major II, which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place: DTU, Delhi
Date: 06-09-2021

Atul Aman
(2K19/CSE/05)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "***Multiview video summarization using target detection and clustering***" submitted by Atul Aman, Roll No. 2K19/CSE/05, Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology (Computer Science and Engineering) is a record of the original work carried out by him under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 06-09-2021

Dr. Anil Singh Parihar
Professor
Department of Computer Science & Engineering
Delhi Technological University

ACKNOWLEDGEMENT

During these hard times of COVID, one thing every human being required was a great source of support that not only supports the person but stands by him/her in all the circumstances. The source of ample support for me was Dr. Anil Singh Parihar, Professor, Department of CSE, DTU, Delhi. He not only guided me for the project but also was always there for me in all the circumstances.

I present my sincere gratitude to Dr. Rajni Jindal, HOD, Department of CSE, DTU, Delhi for her immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Atul Aman
Roll No - 2K19/CSE/05

ABSTRACT

In the era of booming technology, with the advancement of mobile phones and camera-enabled devices the application and purpose of digital data have been exponentially increased. The data collected by these devices are in trillions to larger units of digital data. Therefore, it has become quite uneasy to retrieve valuable information from these videos. Here, it is multiview video summarization, the concept of understanding and finding out the important information from a large video file when inspected through different angles and projections. In this project, we proposed deep learning techniques with a clustering algorithm in three phases for multi-view video summarization. In the first phase, shot segmentation is done using target/object-based along with eliminating redundant frames. The second phase extracts frame-level features using the ResNet50 CNN model and passes them to the final step. In the third or final phase, the visual features are clustered using HDBSCAN and select the final keyframes based on entropy value (informativeness). Experimental results on the popular datasets clearly shows that the proposed methodology performs better than the existing methods.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

1. VS: Video Summarization
2. MVS: Multiview Video Summarization
3. CNN: Convolutional Neural Networks
4. HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise
5. GAN: Generative Adversarial Networks
6. LSTM: Long Short Term Memory
7. DETR: DEtection TRansformer
8. YOLO: You Look Only Once
9. SSIM: Structural Similarity Index Measure
10. SIFT: Scale-invariant feature transform

CHAPTER-1

INTRODUCTION

1.1 Background

In the current world scenario, application and purpose of digital data or media such as text, video etc., has exponentially increased. The digital data is not only for the purpose of research or entertainment but also for record maintenance, security, etc. The data collected from social media, security cameras, offices, institutions, etc. is so huge that it is becoming too difficult to store and retrieve quality information from this data. Hence, this is the reason it has become an apple pie for all researchers to research and find out economical tools to manage this data. Here, not entire, but a part of this data, the streaming video data is considered for research and retrieval of valuable information.

Video summarization is the process of retrieving the important data from these videos and store and use it for various purposes. The summarization tools can be categorized as- Static and Dynamic (Video Skimming) [1]. In dynamic, motion pictures also called skims are used whereas in static, a set of interesting frames are chosen for the objective of video summarization. In static video summarization, the video is divided into various smaller frames on which the clustering algorithms are implemented in order to merge or group the related frames also called keyframes. These frames are selected on the basis of various metric features such as color, shape to name a few. Whereas in dynamic summarization, major methodology and techniques select and fetch the video clips from the original video itself. A few techniques used for video summarization are motion model, semantic analysis, etc.

1.1.1 Single View Video Summarization

There are various studies and efficient tools available for the single view video summarization that efficiently and economically gives the best outcomes. In single view video summarization, the frames lack a large amount of interview correlation making the work less complex and easy for the researcher to retrieve or summarize the data. The methods or techniques fail to expose the different angles and unbind the data correlation from Multiview video summarization methods.

1.1.2 Multi View Video Summarization

These methods introspect the videos properly and in depth trying to resolve various content correlations and try to reduce the redundancy and find out the best shot from the huge data collected using different angles. This work is based on finding out the results using SSIM for the Multiview summarization.

1.2 Motivation of research

The need and purpose of the research comes from the changing lifestyle of people that is leading to the growth of the technologies to the next level. People are more exposed to the security threats they have ever been which has resulted in the development of multiple security measures one of which is security cameras. These multi angle security cameras capture trillions of data which should be completely checked in order to get the required information. The management of this data in order to get the best information in simplified and refined format is the key reason for this research work.

1.3 Objective of research work

The project aims to multi-view video summarization using target detection and clustering.

1.4 Thesis layout

In this thesis, the next sections talks about the following :-

- **Literature Review:** This section talks about various research works carried out by different researchers in the field of video summarization along with their shortcomings.
- **Video Summarization:** This section talks about the different video summarization techniques that are existing as well as the proposed technology.
- **Proposed Work:** The section talks about the proposed technology and proposed algorithm.
- **Implementation and Results:** The implementation of proposed model and various results obtained are observed and noted for the further comparisons and study purpose.
- **Conclusion:** It talks about the entire research and its outcomes in short.

CHAPTER-2

LITERATURE REVIEW

In this section, we will be discussing and referring to a few available researches in order to get the appropriate knowledge about the topic. These research works helped in clarifying the objective and henceforth are referred while doing this research.

2.1 Single View Video Summarization

Publish Year	Paper Title	Algorithms /Method used	Key Points
2019	Video summarization based on motion detection for surveillance systems [2]	• Background Subtraction	Presented a motion detection technique for summarizing video that mainly focuses on overcoming the limitations due to the illumination changes. The concept presented here covers two main methods, i.e., background subtraction and Structure-texture-noise decomposition. The entropy evaluation method was incorporated to tackle the illumination change problems.
2017	Equal Partition Based Clustering Approach for Event Summarization in Videos [3]	• K – means Clustering	Gave an approach of equal partition-based clustering that proved to work efficiently over the real-time data. A cluster validation technique, namely Bouldin index, was implemented to get an adequate cluster of clusters, enabling a user to freely select the keyframe without worrying about extra computational cost. It resulted to be better

			than the baseline methods with good precision and F-measures.
2019	Attentive and adversarial learning for video summarization [4]	• Ptr – Net Generator • 3 D CNN	The combination of the supervised and unsupervised learning methods were incorporated as a GAN-based model to summarize the video and ptr-net generator was used to generate the cutting points of summarization fragments.
2019	Video summarization via spatio-temporal deep architecture, Neurocomputing [5]	• VGG – 16 • Two stream Deep ConvNets	Introduced a method that makes efficient use of the spatial and temporal information of a video for dynamic video summarization.
2018	Video Shot Detection based on SIFT Features and Video Summarization using Expectation-Maximization [6]	• SIFT Feature • Keypoint Matching • EM clustering	The proposed approach consists of 2 main tasks i.e. the shot detection using the SIFT technique to extract features from the frame and expert-maximization method for video summarization.
2018	F-DES: Fast and deep event summarization [7]	• Nucleotide Sequence • Fast Algorithm • Event ordering	To overcome various issues of the video summarization on real-time data such as illumination changes, , a local alignment-based model FASTA is proposed in this work. The proposed model successfully kept the required components of the video

			in the form of events and met the real-time applications.
2016	Context-Aware Surveillance Video Summarization [8]	• Sparse Coding • Spatio – temporal feature • Dictionary	The proposed approach aims to study the information from individual local motion regions and the interaction between these regions. It is a context-awareness video summarization technique that acquires sparse coding features to learn from the dictionary of video features and spatial correlation graphs.
2017	A review on domain adaptive video summarization algorithm [9]	• Local Binary Pattern • HOG	Gave a method finely based on the extraction of high-level features from a video that is to be summarized and classified using any learning technique.
2020	A Novel Key-Frames Selection Framework for Comprehensive Video Summarization [10]	• CapsNet features • Attention Curve • Transition Effects Detection	Proposed a method to capture the contents of the motion videos. Capsule Networks as a spatiotemporal feature extractor that generates a curve for inter-frame motion representation. For automatic shot segments on the curve, a method is proposed here: transition effects detection.
2018	A Key Frame Based Video Summarization using Color Features [11]	• HOG • Histograms	Proposed a method where extraction of color features from the segments of the video frame in place of the entire video is performed, based on this the shot boundaries are identified.

Table 2.1 Literature review of single-view video summarization

2.2 Multi View Video Summarization

Publish Year	Paper Title	Technology /Method Used	Key Points
2015	A Multi-view Video Synopsis Framework [12]	• Video Synopsis • Action Recognition	Presented a simple framework for the task of video summarization and multi-view video synopsis. For synopsis, we need to get the important motions from the video at different intervals of time and arrange them according to the priority. The frames with the greater priority adding much value to the synopsis are added into the synopsis.
2010	Multi-view video summarization [13]	• Spatio Temporal shot graph • Random walks	The paper is based on identifying the interview correlations and extracting valuable information from attribute relations. A Spatio-temporal graph is obtained from a hypergraph that tries to find out the relation between various attributes in multi-view video shots.
2014	Multicamera video summarization and anomaly detection from activity motifs [14]	• Activity Detection	The proposed method aims to summarize the videos based on the activities and gave the summary video by compressing the representation of every individual activity and it also turned out to be beneficial from removing the redundancy.
2015	On-line multi-view video summarization	• GMM	Battery lifetime or energy consumption plays a key role in video summarization.

	for wireless video sensor network [15]	• Online Clustering	The multi-view video cameras capture a huge amount of data and use a huge battery life or energy. In order to optimize the battery consumption, the redundant data is removed in such a way that it can lower the energy consumption and no valuable information is lost.
2016	Embedded sparse coding for summarizing multi-view videos [16]	• Frame embedding • Majorization Minimization Algorithm	Considered both inter and intra view correlations for video summarization in a joint embedded space using majorization-minimization algorithm that decreases the cost function for each iteration in a monotonic manner and therefore, employing a sparse representation approach. The underlying approach proved to be successful for unfolding the data correlations in multi-view videos.
2017	Event BAGGING: A novel event summarization approach in multiview surveillance videos [17]	• Bagging • Learning Algorithm • Euclidean Distance	Presented a method of machine learning. The ensembles are trained using a meta approach where illumination changes and interdependencies are considered the two important feature extraction things. These test sets are further processed to generate the keyframes. Bagging is applied to get accurate results for active and inactive frames. The model resulted in the extraction of keyframes with quality data and met the requirement of real-time applications.

2015	Multi-View Video Summarization Using Bipartite Matching Constrained Optimum-Path Forest Clustering [18]	• HSV feature Extraction • Gaussian Entropy • Bipartite Matching • OPF Clustering	Proposed a graph-theoretic solution to the problem. Semantic features such as text, shape, color are taken into account in a form of visual bag. There and then Gaussian entropy is implemented for the filtration of low activity frames. Bipartite graph is used for mapping the inter-view dependencies and last but not least clustering is done using optimum-path forest algorithm. It shows some improvement over available single and multi-view summarization technologies. The future work is based on the integration of a more extensive set of video features.
2016	Multi-view Metric Learning for Multi-view Video Summarization [19]	• MMC Algorithm • Clustering	Proposed metric learning based multi-view video summarization. The proposed framework used maximum margin clustering (MMC) algorithm along with a disagreement minimization criterion. The high dimensional features of frames are clustered and the final keyframes are extracted from each cluster.
2020	Cloud-assisted multiview video summarization using CNN and bidirectional LSTM [20]	• CNN • Bi-LSTM	Introduced a cloud-based two-tier approach for video summarization. The first online tier performed the target-based shot segmentation and stored the shots in the table for further processing. In the second tier, they first extracted deep

			features from each shot and then passed these features to Bi-LSTM to compute the probability score achieved by the frames.
2021	A comprehensive survey of multi-view video summarization [21]	-	Gave a survey paper comparing various multi-view video summarization technologies available, including their pros and cons.

Table 2.2 Literature review of multi-view video summarization

2.3 Limitations

Although, the work discussed above added a lot of value and scope in the field of video summarization but still there exists a few loopholes to be considered and worked upon. A few suspected limitations of the above works are as follows: -

- The summarized video turned out to be highly redundant, i.e., it contained multiple copies of a frame that unnecessarily increased the output video length.
- While going through the researches, a common issue suspected was that the multi-view summarization in the above works the segmentation is done on the basis of motion, spatio temporal, tracking, etc. but these methods have limitations when we talk about the efficiency and accuracy due to the use of weakly presented activity recognition algorithms.
- The frame difference measure used in the prior research is unable to detect all the events properly because of factors such as camera movements, etc.
- Higher Space complexity is also an additional problem to be taken care of.
- The methods used prior are insufficient for the purpose of MVS when we talk in terms of surveillance camera and their video summarization because the final summary may contain frames without person or still images, which does not have any importance in summarized video.

2.4 Problem Statement

In the era of booming technology, with the advancement of mobile phones and camera-enabled devices the application and purpose of digital data have been exponentially increased. The data collected by these devices are in trillions to larger units of digital data. Therefore, it has become quite uneasy to retrieve valuable information from these videos. The project aims to summarize a multiview video using the target identification and clustering as the proposed technology and view as well as compare its results with baseline methodologies.

CHAPTER-3

VIDEO SUMMARIZATION

3.1 Definition

The video summarization is a technique or process of extracting the important features and details from a large set of videos for various reasons or purposes such as in security, the video summarization technique may help in deriving the best information from a huge pile of data. The video summarization is done using various machine learning and deep learning algorithms such as clustering, CNN, etc.

3.2 Types of Video Summarization

The task of video summarization can be categorized into two forms:

i) Static video summarization

ii) Dynamic video summarization

3.2.1 Static video summarization

Static video summarization is done by extraction keyframes from the original video. Key frame extraction, as the name suggests, is to choose the most informative frames from the video. These indexed frames are supposed to be the best ones that summarize the video. The key frame extraction is primarily used to obtain static summaries.

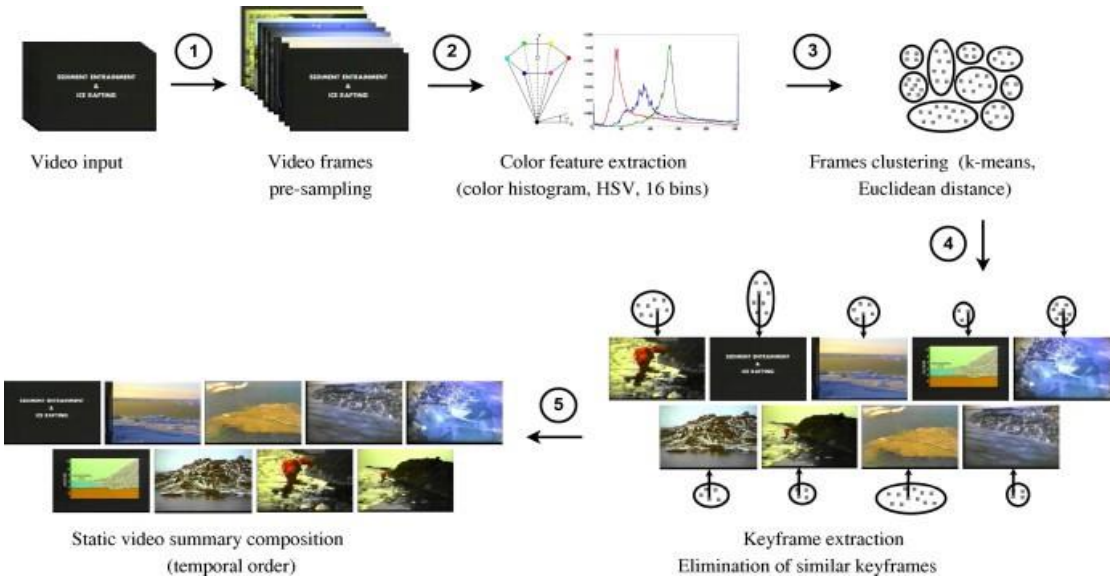

Fig 3.1 Static Video Summarization

3.2.2 Dynamic video summarization

Using keyframes to summarize a video might be useful for automatically analyzing the content of the video, but it produces a discontinuous and rather unpleasant summary for human viewing. This calls for summarizing a video in the form of skims of the video. This however is a complex task and often is more difficult to achieve for user videos which lack structure. The semantic meaning is frequently required in such cases. Dynamic video summarization is also known as video skimming.

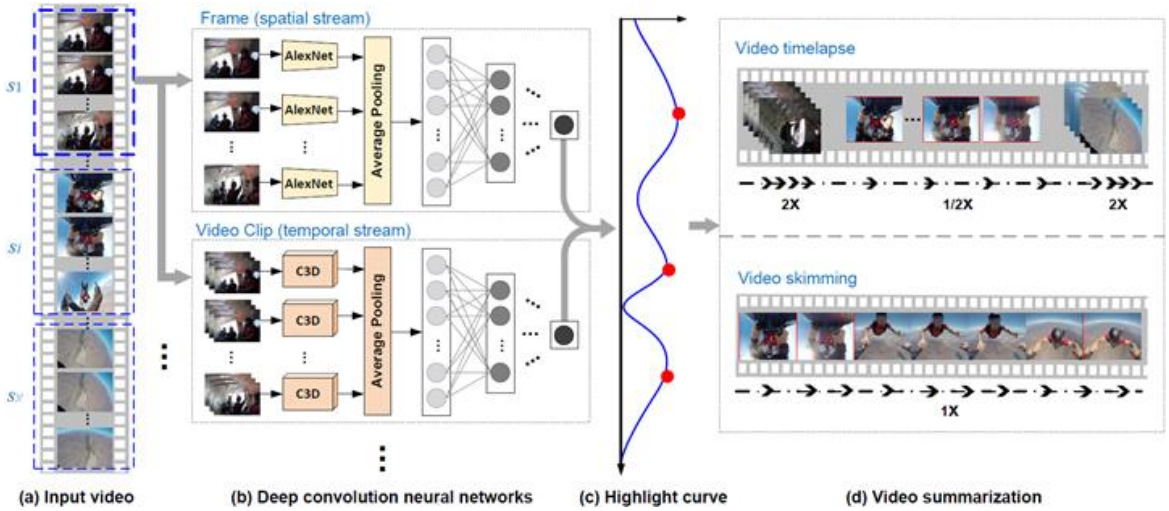

Fig 3.2 Dynamic Video Summarization

3.3 Video summarization techniques

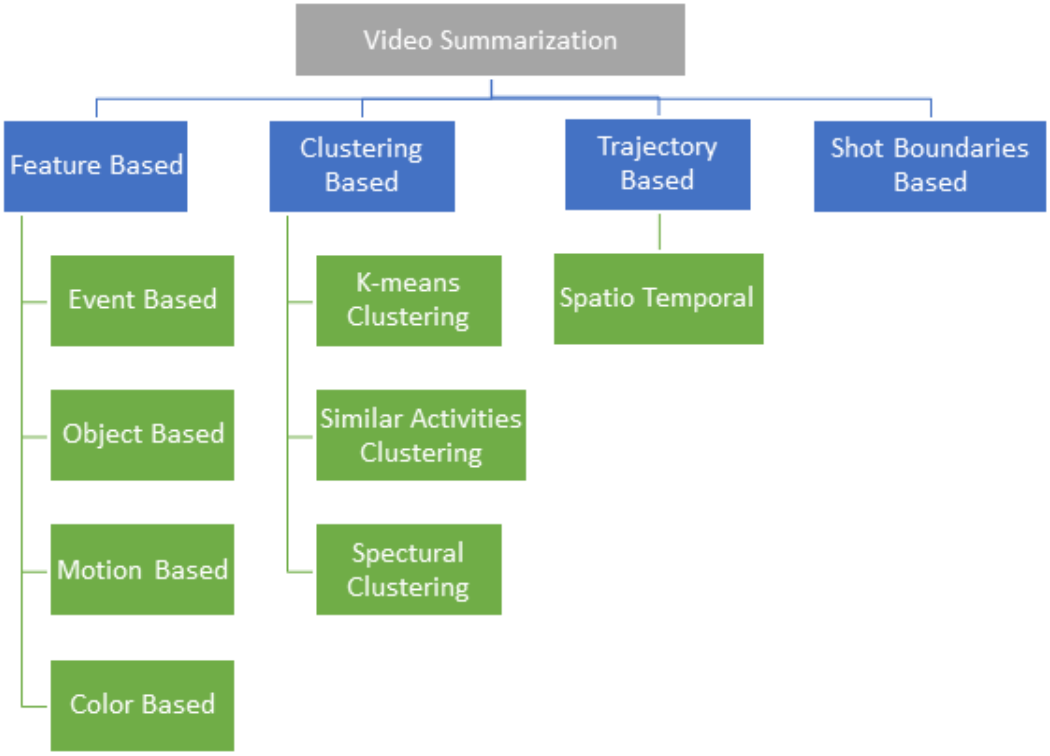

Fig 3.3 Video Summarization Techniques

3.3.1 Feature Based Summarization

The video summarization can be done using various features such as motion, color, event or object based. These features are important for the representation of work efficiently and in a neat manner. The various feature based techniques are explained as follows:-

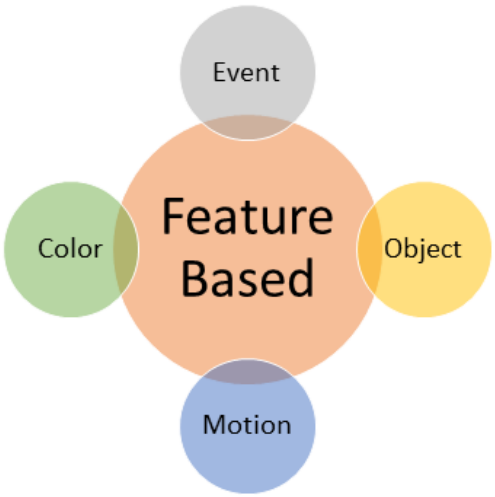

Fig 3.4 Categories of Feature based Summarization

- **Event Based Summarization technique:** These are important to identify normal or abnormal events that are present in a video.
- **Object Based Summarization Technique:** This technique is helpful in detecting objects from a video such as person, place or thing. The summarization is achieved by collecting all the frames from the video that contains the required object.
- **Motion Based and Color Based Summarization Technique:** The capturing of motioned frames is a quite tough form of summarization technique whereas for color, a particular color object or thing can be captured from various frames and used for summarization.

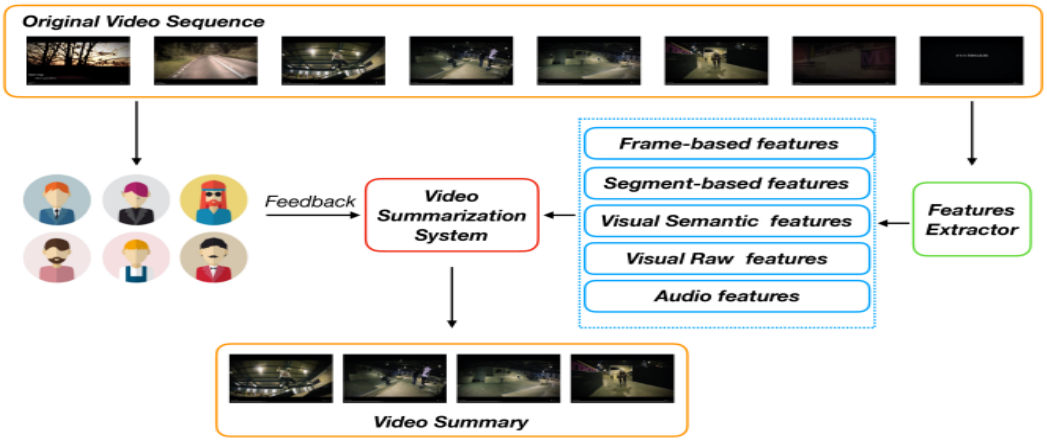

Fig 3.5 Feature Based Video Summarization

3.3.2 Clustering Based Summarization

The video summarization can be done using clustering technique where clusters of frames with the desired features are chosen and used for the video optimization and summarization. The various clustering based summarization techniques are explained as follows:-

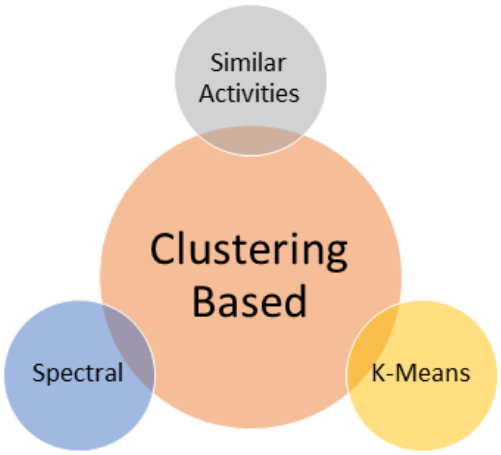

Fig 3.6 Categories of Clustering based Summarization

- **Similar Activities Clustering:** The videos can be summarized based on the similar activities present in various frames. The activity is described beforehand then a similar set of activities are identified based on small aspects present across various frames. At last, these frames are compared based on distance and how similar they are to each other for getting a summarized video.
- **K-Means Clustering:** First the video is divided into multiple segments where the first frame acts as a representative for that segment. These frames can be derived using a histogram. Lastly, the histograms are clustered using the K-means algorithm.
- **Spectral Clustering:** This technique can be used to detect the human face by using the number of faces, sizes and locations. This technique is used only for detecting the desired face, not for multiple faces.

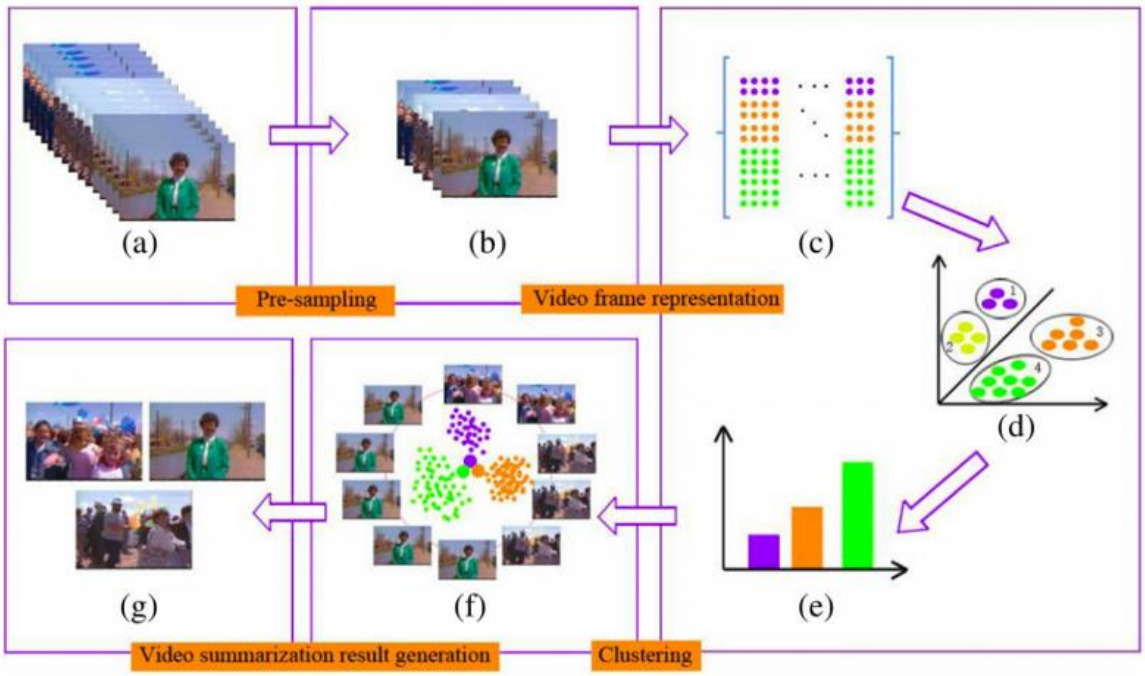

Fig 3.7 Clustering Based video Summarization

3.3.3 Shot boundaries Based Summarization

This technique fetches the key frames from the video where the extraction of the first image is considered to be a shot key frame.

3.3.4 Trajectory Based Summarization

This method is used for analysis in the dynamic environment. It is good and efficient to summarize surveillance camera videos etc.

CHAPTER-4

PROPOSED WORK

This section gives detailed information on the basic concepts, problem formulation and the overall working of the proposed method. The approach mainly consists of three main components i.e. Shot Segmentation, Deep Feature Extraction and Keyframe Selection Mechanism. Table 4.1 depicts and gives all the details of the workflow for the method. The steps are clearly explained in Fig. 4.2.

4.1 Basic Concept

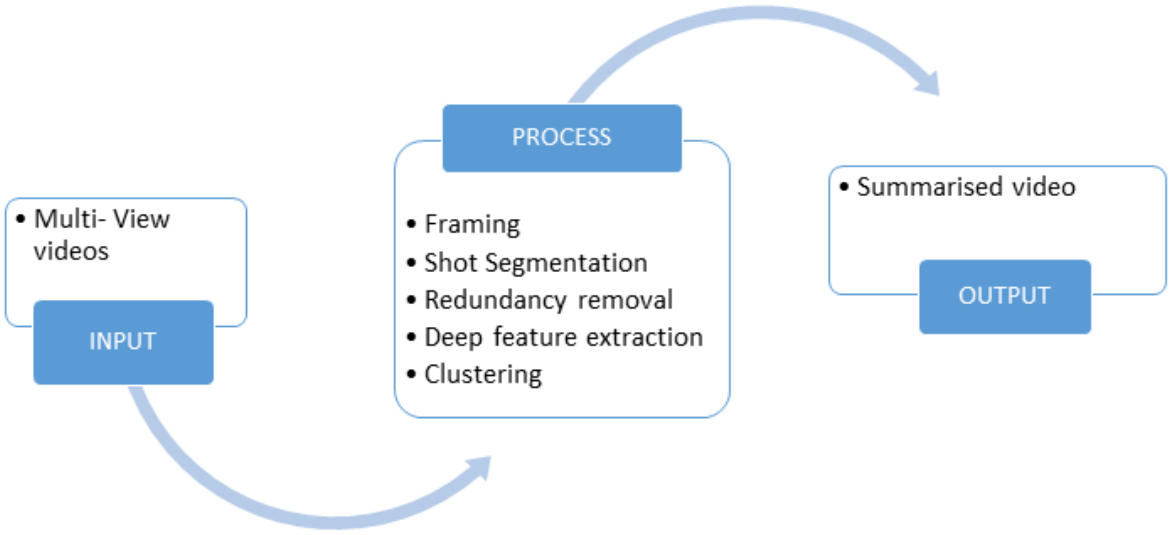

Fig 4.1 Process Flow

As seen in the process flow diagram **fig 2**, the requirements for the proposed work are categories in three main categories i.e input, detailed processing and the output.

These categories are:

Input: Multi-view surveillance video

Process: -

- Framing
- Shot Segmentation

- Redundancy Removal
- Deep feature extraction
- Clustering

Output: - Summarized final video

4.1.1 Input

4.1.2 Process

a. Framing

In this process the frames of video are extracted based on the sampling rate that is decided by experimenting several times to handle the enormous number of frames present in the video.

b. Shot Segmentation

This is the pre-processing step and an important step in multiview video summary generation. Shot segmentation is done based on target based detection to capture human actions and various important motions and factors.

c. Redundancy Removal

The surveillance videos turn out to be highly redundant and in many of the cases there may exist static frames. These redundant and static frames do not provide any useful information. So to overcome these challenges a redundancy removal algorithm is introduced to eliminate the similar frames.

d. Deep Feature Extraction

This process involves the use of CNN model to extract the frame level deep features that will help in deciding whether the frame is informative or not.

e. Clustering

This is the final step of the proposed work where the final keyframes are obtained by using a clustering algorithm which clusters the similar frames and selecting the best frame from the cluster based on an informative score as the final keyframe.

4.1.3 Output

After the entire process is done the selected final keyframes are further transformed to a summarized video.

4.2 Proposed methodology

4.2.1 Problem Formulation

First, frames are extracted from the N multi-view videos captured by N cameras. In the proposed approach, we have used a sampling rate of 15 frames per second to handle the enormous number of frames. F^1, F^2, F^3, ..., F^N represents the set of frames of each view, where $F^i = \{f^i_{(1)}, f^i_{(2)}, f^i_{(3)}, ..., f^i_{(n)}\}$, n denotes the total number frames in the i^{th} view.

4.2.2 Proposed Architecture

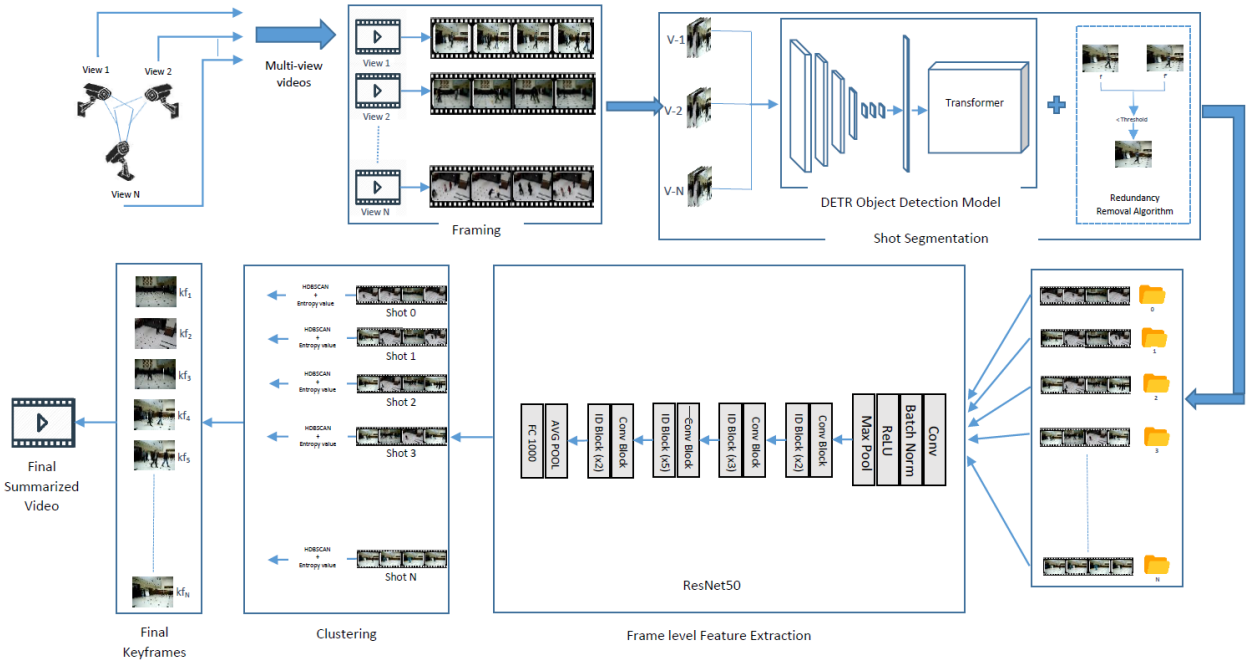

Fig 4.2 Proposed Architecture

4.2.3 Proposed Algorithm

Algorithm 1: Proposed algorithm
Notations: kf_{final} = final set of keyframes, S_i = i^{th} shot, S_c = current shot, DF_{2048} = 2048 dimensional features, kf_s^N = selected keyframe of view N, V_s = Summarized video
Input: N - view videos **Output:** Final summarized video Dataset: $view_1$, $view_2$, ……,$view_N$ // $view_i$ = Input video for the i^{th} view 1. Extract frames according to the sampling rate from $view_i$ to get $F^i = \{f^i_{(1)}, f^i_{(2)}, f^i_{(3)}, ..., f^i_{(n)}\}\ \forall\ i\ \varepsilon\ \{1, N\}$ 2. Apply object detection model to detect targets on F^i. 3. Do while $f^i_{(j)}$ has target objects: 4. Compute similarity score $SS^i_{(j)}$ using Eq. (1) for each $f^i_{(j)}$ and $f^i_{(j-1)}$ in the particular shot S_i. 5. if $SS^i_{(j)}$ < threshold then 6. Select $f^i_{(j-1)}$ in the current shot S_c. 7. for each shot S_i do 8. DF_{2048} are extracted using ResNet50 CNN model and store the feature vector. 9. Feature vector is input to HDBSCAN clustering algorithm to find out the cluster. 10. Generate kf_s^i for each cluster by computing entropy value of each frame in the particular cluster and pick the frame with the highest value. 11. Obtain final keyframe set kf_{final} = [kf_1, kf_2, ……., kf_r] by combining the set of keyframes obtained in step 10. r: total number of keyframes obtained 12. Convert kf_{final} into final summarized video V_s.

Table 4.1 Proposed Algorithm

4.2.4 Shot Segmentation

The shot segmentation is a pre-processing step and an important step in multiview video summary generation. In the multiview summarization theories, there are multiple techniques used for shot segmentation. In some literature, shot segmentation is based on motion in the video data [18],

spatio-temporal C3D features [22] and activity-based video segmentation [13] . The discussed approaches are not upto to the mark when it comes to fill in the room for surveillance videos. Surveillance video summarization requires capturing person, place, vehicle, etc. from different angles and in a quite informative manner that these mentioned approaches fail to detect. To capture the human actions and various other important motions and factors, we tried and implemented an object detection model to segment the video into shots such that a significant part can be captured and made available in the final video.

In the proposed approach, shot segmentation consists of two major steps: object detection algorithm and redundancy removal algorithm.

i) Object detection model

The object detection model has been widely used in video surveillance and object tracking applications in recent years. Object detection primarily identifies and classifies the numerous objects in an image using a bounding box around the object.

Prior approaches [20] used YOLOv3 object detection model for shot segmentation. YOLOv3 [23] is extremely fast (processes images at 30 fps) and achieves better accuracy than state-of-the-art methods for target detection. Here, we implemented a generator-based method called DETR [24] as well as the baseline method (i.e. YOLO) and compared and contrasted the results. DETR is developed by the Facebook research team, which utilizes transformers [25] and a bipartite matching loss. It contains a CNN backbone for feature extraction, a transformer and a feed-forward network head for final detection predictions. Compared to YOLO, DETR has better average precision for small and medium object sizes.

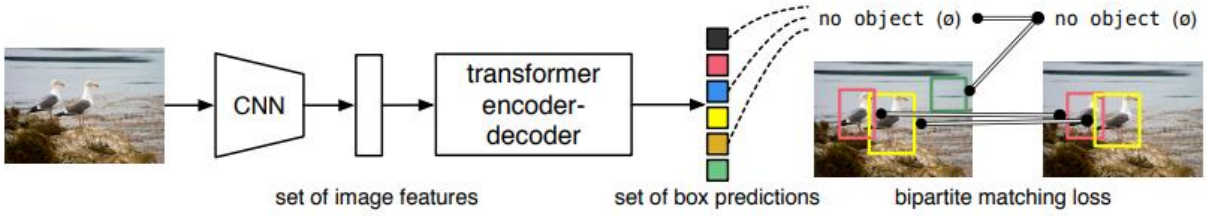

Fig 4.3 DETR

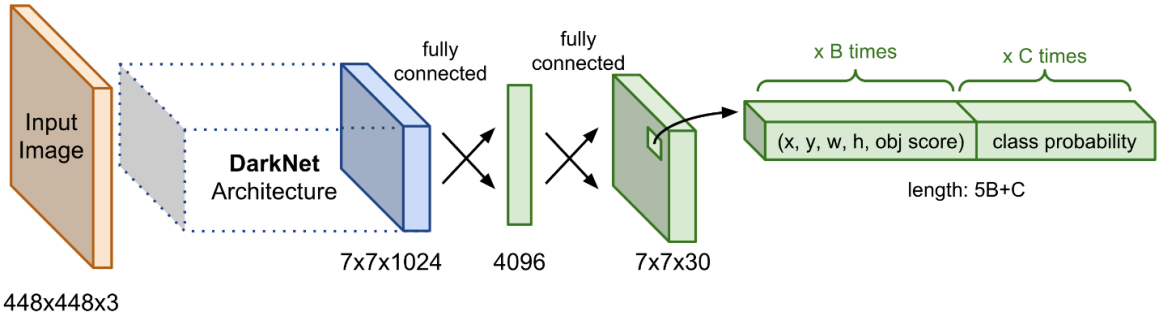

Fig 4.4 YOLO

From the results in Table 4.2, it was as clear as water that the target detection done using DETR performs better than YOLO on the given dataset. DETR works well on low-quality images and detecting small objects as compared to YOLO

Dataset	Object detection model used	Total number of frames	Number of frames after target detection
Office	YOLO v3	5558	1210
	DETR	5558	1922
Lobby	YOLO v3	2964	1810
	DETR	2964	2072

Table 4.2 Performance comparison of shot segmentation mechanism

ii) Redundancy removal algorithm

The surveillance video turns out to be highly redundant and in many of the cases there may exists static frames. These redundant and static frames do not provide any useful information and thus increase the output video length. Therefore, the structural similarity index measure (SSIM) [26] is applied to the frames with the target objects. It reduces the redundancy by computing the SSIM score between two frames. The two frames are compared by extracting 3 features i.e. brightness, contrast and structure from an frame. The SSIM score lies in the scale of -1 to +1 where +1 indicates that the two frames are the same while -1 indicates different frames.

The SSIM score can be calculated by the formula mentioned below:

$$S^{i}_{(j)}(x_1, x_2) = \frac{(2\mu_{x_1}\mu_{x_2} + C_1)\,(2\sigma_{x_1x_2} + C_2)}{(\mu^2_{x_1} + \mu^2_{x_2} + C_1)\,(\sigma^2_{x_1} + \sigma^2_{x_2} + C_2)} \quad (1)$$

where $x_1 = f^{i}_{(j)}$ and $x_2 = f^{i}_{(j-1)}$

For each shot, the mean of SSIM score is obtained and is considered as the threshold for that shot. The frames with a target whose similarity score is less than threshold value is considered as the candidate frames.

To maintain the inter-view correlation, we store segmented shots with given targets in a synchronized format for all the available views. The frames are selected in a timely manner i.e. for a particular time interval 't'. A time interval of 15 seconds is chosen for each shot. Each shot may contain a different number of frames from all the views. Sample frames of our proposed shot segmentation are shown in Fig 4.5 & Fig from office and lobby dataset.

Fig 4.5. Sample frames using proposed shot segmentation on Lobby dataset.

Fig 4.6. Sample frames using proposed shot segmentation on Office dataset.

4.2.5 Feature Extraction

The next step of the proposed approach involves frame-level deep feature extraction. It works by extracting useful features that help in deciding whether the frame is informative or not. Deep learning model extracts features that express images in much more details rather than making the use of handcrafted features such as histogram features [27] , SIFT features [18], and a few lower level traits to suspect the informativity of frame [28].

We used the learned traits of the ResNet50 [29] CNN model. It is trained on the ImageNet [30] dataset. We also tried the extraction phase using a lighter CNN model like MobileNet [31], but outcomes were not as good as the ResNet50. Compared to Hussain et al. [20], who utilized the AlexNet [32] CNN model for deep feature extraction, ResNet50 has lesser parameters (25 M) which means less computation and less training time with improved top-5 accuracy than AlexNet. We used the average pool layer features for frame representation, which produce a 1 x 2048 feature vector for each frame. For each shot, the frames' features are obtained, and the resultant visual features are fed as an input for the pipelined processes.

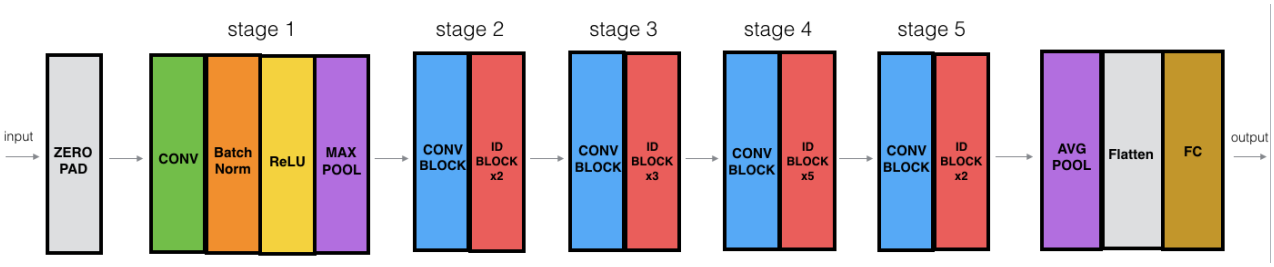

Fig 4.7 ResNet 50 Model

4.2.6 Keyframe Selection Mechanism

After the features are extracted from the frames, the final step is done in two levels to each shot's feature vector to get the final keyframes kf_{final}. The feature vector of frames of each shot is first input to the HDBSCAN clustering algorithm [33] to find out the various clusters from each shot by locating the regions that have higher density than the surroundings. Compared to the partition based K-means clustering algorithm used in prior approaches [27,34], HDBSCAN does not require

a number of clusters at the initial stage. It works well when there is some noise and clusters with different size and densities. After finding the clusters, the final step includes finding each frame's entropy value using Eq. (2) in a particular cluster. These entropy values describe whether the frame is informative or not. The higher the value tells that the frame is largely valuable (highly informative) and the lesser value represents less important. The process of entropy calculation is described in Algorithm 2.

$$E(frame) = -\sum_{i=0}^{n} p_i \, log_b \, p_i \qquad (2)$$

where n = number of RGB color space, p_i = probability of a pixel having RGB color space

Algorithm 2: Entropy score calculation
Input: single frame $f^i_{(j)}$ **Output**: Entropy score E(j) 1. Calculate histogram of $f^i_{(j)}$ in RGB space. 2. Compute probability of histogram (obtained in step 1) in frame 3. Calculate entropy of each color space using Eq. (2)

Table 4.3 Entropy Score Calculation

4.2.7 Summarized video

In the final step, the best keyframes from each cluster of each shot are combined into a single set of keyframes final which is further transformed to a summarized video. The keyframes that are part of final summary video of dataset are as shown in Fig 4.8 and Fig. respectively

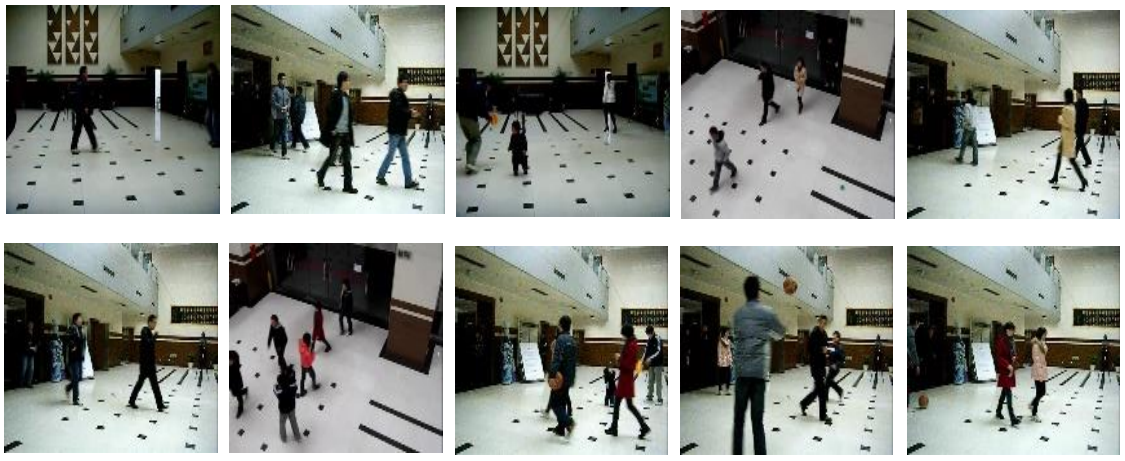

Fig. 4.8 Sample keyframes of the final summary using our proposed approach on Lobby dataset.

Fig. 4.9 Sample keyframes of the final summary using our proposed approach on Office dataset.

CHAPTER-5

IMPLEMENTATION AND RESULTS

5.1 Experimental settings and dataset description

The proposed system is implemented in Python language and all the experiments are carried out on an i5 processor with 12 GB of DDR3-RAM laptop. We conducted experiments on the most widely used publicly available dataset for multiview video summarization.

The details of these surveillance video datasets are as follows:

(a) Office [13]: This is one of the popular datasets that contains 4 surveillance videos recorded by stable and unsynchronized cameras in an office

(b) Lobby [13]: This dataset contains 3 surveillance videos recorded by fixed and unstable cameras in a lobby area. These videos are synchronized with each other and capture various events.

5.2 Quantitative analysis

This section explores the various parameters to measure the performance of the final summarized video. Though there is no optimal technique to measure the performance of the video summarization algorithm. Thus, we chose the standard metrics which were used by the prior approaches to evaluate the final summarized video.

These are as follows:

(a) Video Length Reduction (LR):

It represents the percentage of reduction of video length of summarized video compared to the original video length.

$$LR = (1 - S_{len} / O_{len}) \times 100 \quad (3)$$

where S_{len} denotes summary length and O_{len} denotes original video length. Results of LR are given in Table 5.1

Dataset	Method	Input Video (s)	Summary Video (s)	Percentage reduced Length (LR)
Lobby	GMM	1482	484	67.34
	Bipartite Matching	1482	176	88.12
	Muti-view Metric	1482	158	89.33
	Event Bagging	1482	152	89.74
	Proposed	**1482**	**110**	**92.57**
Office	GMM	2779	402	85.53
	Bipartite Matching	2779	59	97.87
	Muti-view Metric	2779	-	-
	Event Bagging	2779	80	97.12
	Proposed	**2779**	**50**	**98.20**

Table 5.1 Performance Results I

(b) Event Detection ratio (ER):

It represents the ratio of events in the summarized video to events given in ground truth.

$$ER = Es/Eo \quad (4)$$

where E_s denotes number of events detected in video summary and E_o denotes number of events given in the ground truth. Results of ER are given in Table 5.2.

Dataset	Method	Input Video (s)	Number of events detected in summary	Event Detection Ratio (ER)
Lobby	GMM	1482	-	-
	Bipartite Matching	1482	33/35	0.94
	Muti-view Metric	1482	34/35	0.97
	Event Bagging	1482	34/35	0.97
	Proposed	**1482**	**35/35**	**1.00**
Office	GMM	2779	-	-
	Bipartite Matching	2779	18/25	0.72
	Muti-view Metric	2779	20/25	0.80
	Event Bagging	2779	21/25	0.84
	Proposed	**2779**	**24/25**	**0.96**

Table 5.2 Performance Results II

5.3 Qualitative analysis

G. Money et al. [36] suggested that only quantitative analysis is not enough to evaluate the video summarization, subjective evaluations are necessary for further evaluation. Video Summarization is quite extensively based on the viewer or user's imagination or perception. For example, if we have a glass of water that is filled upto the 50% of its limit, for some viewers it will be half empty and for some half full. Each user has its own way to look towards an object or aim or subject let's say. This is the reason that brings the visual pleasantness and informativeness of the final video as significantly important. To tackle such a situation, a group of random 10 people who had no information about the project were asked to review the summarized video on a scale of 5 on the basis of the informativeness and goodness of this summarized video where a score of 5 was an excellent score and 1 as a poor grade for the video and its quality. The results of user study are shown in Table 5.3.

Measures	Dataset	Multi-view [13]	Bipartite Matching [18]	Cloud Assisted [20]	Ours
Visual Pleasantness	Lobby	4.1	4.5	-	4.9
	Office	4.2	4.5	4.5	4.7
Informativeness	Lobby	4.0	4.3	-	4.7
	Office	4.1	4.5	4.2	4.7

Table 5.3. Statistical data of qualitative analysis

5.4 Analysis of results

From Table 5.1, We can clearly see that the summary length for the lobby data in the table as per the baseline models, only event bagging turns out to be the best performing algorithm here which reduced the input length of video to 152 s from 1482 s whereas our proposed method out performed all the state-of-the-art methods and resulted in a summarized video of length 110 s with about 3% improvement as compared to the baseline methods. Similarly, for the office dataset, the summarized video length given by event bagging was 80 from 2779(input length) and the summarized video produced by the proposed method is of length 50.

The event detection efficiency of the proposed method comes out to be 100% whereas for the best state-of-the-art method it was 97% for the lobby dataset.

CHAPTER-6

CONCLUSION

We compared the proposed method with a few state-of-the-art methods [17], [18], [19], [15]. As per the results out of all these state of the art method the best method comes out to be event bagging which out performs all the baseline methods with the better results i.e. it produced a summarized video of length 110 s and 50 s for the Lobby and Office dataset respectively with an efficiency of 97% to detect the events. The proposed method used the frame redundancy removal metric during the shot segmentation that resulted in the reduction of redundancy at a larger scale as compared to the baseline methodology which helped it in outperforming all the state-of-art methods by producing an efficiency of 100% for event detection and produced a better summarized length.

APPENDIX A

Appendices

A.1 Publication (Communicated)

A.1.1 Multi-view video summarization using target detection and clustering

Abstract

In the era of booming technology, with the advancement of mobile phones and camera-enabled devices the application and purpose of digital data have been exponentially increased. The data collected by these devices are in trillions to larger units of digital data. Therefore, it has become quite uneasy to retrieve valuable information from these videos. Here, it is Multiview video summarization, the concept of understanding and finding out the important information from a large video file when inspected through different angles and projections. In this paper, we proposed deep learning techniques with a clustering algorithm in three phases for multi-view video summarization. In the first phase, shot segmentation is done using target/object-based along with eliminating redundant frames. The second phase extracts frame-level features using the ResNet50 CNN model and passes them to the final step. In the third or final phase, the visual features are clustered using HDBSCAN and select the final keyframes based on entropy value (informativeness). Experimental results on the popular datasets clearly shows that the proposed methodology performs better than the existing methods.

Keywords: Surveillance videos, Target detection, Multi-view video summarization, Density based clustering, Entropy, Structural similarity

REFERENCES

[1] L. Zhang, L. Sun, W. Wang, Y. Tian, KaaS: A Standard Framework Proposal on Video Skimming, IEEE Internet Comput. 20 (2016) 54–59. https://doi.org/10.1109/MIC.2016.83.

[2] O. Elharrouss, N. Al-Maadeed, S. Al-Maadeed, Video summarization based on motion detection for surveillance systems, 2019 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2019. (2019) 366–371. https://doi.org/10.1109/IWCMC.2019.8766541.

[3] K. Kumar, D.D. Shrimankar, N. Singh, Equal Partition Based Clustering Approach for Event Summarization in Videos, Proc. - 12th Int. Conf. Signal Image Technol. Internet-Based Syst. SITIS 2016. (2017) 119–126. https://doi.org/10.1109/SITIS.2016.27.

[4] T.J. Fu, S.H. Tai, H.T. Chen, Attentive and adversarial learning for video summarization, Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vision, WACV 2019. (2019) 1579–1587. https://doi.org/10.1109/WACV.2019.00173.

[5] S. hua Zhong, J. Wu, J. Jiang, Video summarization via spatio-temporal deep architecture, Neurocomputing. 332 (2019) 224–235. https://doi.org/10.1016/j.neucom.2018.12.040.

[6] J. Majumdar, M. Awale, K.L.K. Santhosh, Video Shot Detection based on SIFT Features and Video Summarization using Expectation-Maximization, 2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018. (2018) 1033–1037. https://doi.org/10.1109/ICACCI.2018.8554662.

[7] K. Kumar, D.D. Shrimankar, F-DES: Fast and deep event summarization, IEEE Trans. Multimed. 20 (2018) 323–334. https://doi.org/10.1109/TMM.2017.2741423.

[8] S. Zhang, Y. Zhu, A.K. Roy-Chowdhury, Context-Aware Surveillance Video Summarization, IEEE Trans. Image Process. 25 (2016) 5469–5478. https://doi.org/10.1109/TIP.2016.2601493.

[9] N.R. Aiswarya, P.S. Smitha, A review on domain adaptive video summarization algorithm, 2017 Int. Conf. Networks Adv. Comput. Technol. NetACT 2017. (2017) 412–415. https://doi.org/10.1109/NETACT.2017.8076806.

[10] C. Huang, H. Wang, A Novel Key-Frames Selection Framework for Comprehensive Video Summarization, IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 577–589. https://doi.org/10.1109/TCSVT.2019.2890899.

[11] M. Asim, N. Almaadeed, S. Al-Maadeed, A. Bouridane, A. Beghdadi, A Key Frame Based Video Summarization using Color Features, 2018 Colour Vis. Comput. Symp. CVCS 2018. (2018) 1–6. https://doi.org/10.1109/CVCS.2018.8496473.

[12] I. Terms, A MULTI-VIEW VIDEO SYNOPSIS FRAMEWORK Ansuman Mahapatra , Pankaj K Sa , and Banshidhar Majhi Department of Computer Science and Engineering National Institute of Technology Rourkela, Int. Conf. Image Process. (2015) 1–5.

[13] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, Z.H. Zhou, Multi-view video summarization, IEEE Trans. Multimed. 12 (2010) 717–729. https://doi.org/10.1109/TMM.2010.2052025.

[14] C. De Leo, B.S. Manjunath, Multicamera video summarization and anomaly detection from activity motifs, ACM Trans. Sens. Networks. 10 (2014) 1–30. https://doi.org/10.1145/2530285.

[15] S.H. Ou, C.H. Lee, V.S. Somayazulu, Y.K. Chen, S.Y. Chien, On-line multi-view video summarization for wireless video sensor network, IEEE J. Sel. Top. Signal Process. 9 (2015) 165–179. https://doi.org/10.1109/JSTSP.2014.2331916.

[16] R. Panda, A. Das, A.K. Roy-Chowdhury, Embedded sparse coding for summarizing multi-view videos, Proc. - Int. Conf. Image Process. ICIP. 2016-Augus (2016) 191–195. https://doi.org/10.1109/ICIP.2016.7532345.

[17] K. Kumar, D.D. Shrimankar, N. Singh, Event BAGGING: A novel event summarization approach in multiview surveillance videos, Proc. 2017 Int. Conf. Innov. Electron. Signal Process. Commun. IESC 2017. (2017) 106–111. https://doi.org/10.1109/IESPC.2017.8071874.

[18] S.K. Kuanar, K.B. Ranga, A.S. Chowdhury, Multi-View Video Summarization Using Bipartite Matching Constrained Optimum-Path Forest Clustering, IEEE Trans. Multimed. 17 (2015) 1166–1173. https://doi.org/10.1109/TMM.2015.2443558.

[19] L. Wang, X. Fang, Y. Guo, Y. Fu, Multi-view Metric Learning for Multi-view Video

Summarization, Proc. - 2016 Int. Conf. Cyberworlds, CW 2016. (2016) 179–182. https://doi.org/10.1109/CW.2016.38.

[20] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S.W. Baik, V.H.C. De Albuquerque, Cloud-assisted multiview video summarization using CNN and bidirectional LSTM, IEEE Trans. Ind. Informatics. 16 (2020) 77–86. https://doi.org/10.1109/TII.2019.2929228.

[21] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S.W. Baik, V.H.C. de Albuquerque, A comprehensive survey of multi-view video summarization, Pattern Recognit. 109 (2021). https://doi.org/10.1016/j.patcog.2020.107567.

[22] R. Panda, A.K. Roy-Chowdhury, Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization, IEEE Trans. Multimed. 19 (2017) 2010–2021. https://doi.org/10.1109/TMM.2017.2708981.

[23] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 779–788. https://doi.org/10.1109/CVPR.2016.91.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-End Object Detection with Transformers, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 12346 LNCS (2020) 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 2017-Decem (2017) 5999–6009.

[26] A.O. Daoud, A.A. Tsehayae, A.R. Fayek, A guided evaluation of the impact of research and development partnerships on university, industry, and government, Can. J. Civ. Eng. 44 (2017) 253–263. https://doi.org/10.1139/cjce-2016-0381.

[27] S.H. Ou, Y.C. Lu, J.P. Wang, S.Y. Chien, S. De Lin, M.Y. Yeti, C.H. Lee, P.B. Gibbons, V.S. Somayazulu, Y.K. Chen, Communication-efficient multi-view keyframe extraction in distributed video sensors, 2014 IEEE Vis. Commun. Image Process. Conf. VCIP 2014. (2015) 13–16. https://doi.org/10.1109/VCIP.2014.7051492.

[28] T. Hussain, K. Muhammad, J. Del Ser, S.W. Baik, V.H.C. De Albuquerque, Intelligent Embedded Vision for Summarization of Multiview Videos in IIoT, IEEE Trans. Ind. Informatics. 16 (2020) 2592–2602. https://doi.org/10.1109/TII.2019.2937905.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December (2016) 770–778. https://doi.org/10.1109/CVPR.2016.90.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115 (2015) 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[31] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, (2017). http://arxiv.org/abs/1704.04861.

[32] T.F. Gonzalez, Handbook of approximation algorithms and metaheuristics, Handb. Approx. Algorithms Metaheuristics. (2007) 1–1432. https://doi.org/10.1201/9781420010749.

[33] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 7819 LNAI (2013) 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.

[34] X. Zhu, J. Liu, J. Wang, H. Lu, Key observation selection-based effective video synopsis for camera network, Mach. Vis. Appl. 25 (2014) 145–157. https://doi.org/10.1007/s00138-013-0519-8.

[35] A. Singh Parihar, J. Pal, I. Sharma, Multiview video summarization using video partitioning and clustering, J. Vis. Commun. Image Represent. 74 (2021) 102991. https://doi.org/10.1016/j.jvcir.2020.102991.

[36] A.G. Money, H. Agius, Video summarisation: A conceptual framework and survey of the state of the art, J. Vis. Commun. Image Represent. 19 (2008) 121–143. https://doi.org/10.1016/j.jvcir.2007.04.002.