International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Aug-2022)

https://dx.doi.org/10.12785/ijcds/120150

Intelligent Video Analytic Based Framework for Multi-View Video Summarization

Vishal Parikh[1] **and Priyanka Sharma**[2]

[1]*Department of Computer Science and Engineering, Institute of Technology, Nirma University, Gujarat, India*
[2]*Samyak Infotech Pvt Ltd., Ahmedabad, Gujarat, India*

Received 7 Jan. 2021, Revised 16 Jul. 2022, Accepted 31 Jul. 2022, Published 6 Aug. 2022

Abstract: A multi-view surveillance system captures the scenic details from a different perspective, defined by camera placements. The recorded data is used for feature extraction, which can be further utilized for various pattern-based analytic processes like object detection, event identification, and object tracking. In this proposed work, we present a method for creating a network of the optimal number of video cameras, to cover the maximum overlapping area under surveillance. In the proposed work, the focus is on developing algorithms for deciding efficient camera placement of multiple cameras at various junctions and intersections to generate a video summary based on the multiple views. Deep learning models like YOLO have been used for object detection based on the generation of a large number of bounding boxes and the associated search technique for generating rankings based on the views of the multiple cameras. Based on the view quality, the dominant views will be located. Further, keyframes are selected based on maximum frame coverage from these views. A video summary will be generated based on these keyframes. Thus, the video summary is generated through solving a multi-objective optimization problem based on keyframe importance evaluated using a maximum frame coverage.

Keywords: Multi-view Video, Video Summarization, Camera Placement, Keyframe Extraction

I. Introduction and Overview

The volume of video data captured, stored, and seen has increased dramatically due to rapid advancements in video surveillance technology. This necessitates efficient and consistent video information retrieval methods for analysis [1]. There is a requirement to summarize a large number of videos by extracting the relevant features and events from the videos and maintaining just those videos that are required for reference in the future, i.e., the storage capacity can be reduced. As a result, we may enhance storage efficiency and bandwidth consumption by maintaining crucial information in the actual video. The goal of the video summarization method is to create a meaningful segment of the complete video that explains everything that has transpired so far in a shorter amount of time [2].

The need for security and monitoring requires the setting up of multiple surveillance cameras that captures a single area from different angles to obtain an image to totality, by reducing or removing the number of blind zones of the area under surveillance. Multi-view surveillance systems target a broad application area, and so it has been attracting the researchers. The placement of different cameras in multi-view surveillance systems is crucial. The primary issue that is required to be addressed in the multi-camera network is to filter out the necessary information from a set of different camera angles of the same place. When required, the overlapping and unimportant information becomes costly and time-consuming to store and transfer. Extracting information from the captured videos is the main challenge. Generation of the automatic composite summary from all the views by incorporating important contents is crucial in a multi-view scenario.

There are many non-overlapping and overlapping Field of View (FoV) in a camera network, as shown in Fig. 1. The necessity for sifting out correlating information from several perspectives to build a multi-view summary is due to the overlapping field of vision [3].

Truong et al. [1] and Money et al. [4] have given comprehensive reviews of single view video summarization. According to Truong et al. in [1], a keyframe sequence and a video skim are the two ways of video summarization. Nowadays, due to multiple cameras capturing an overlapping FoV, multi-view summarization is gaining popularity. Hence the multi-view video summarization has become an active research topic.

E-mail address: vishalparikh@nirmauni.ac.in, drpriyankasharma.ai@gmail.com

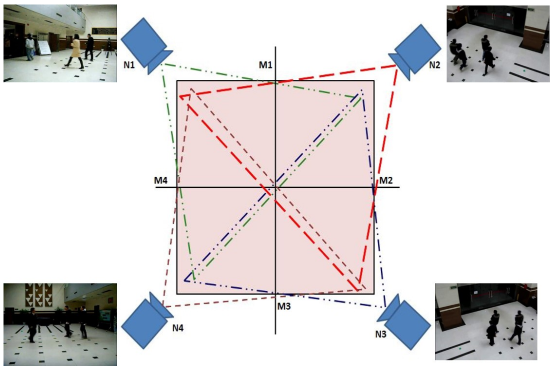


Figure 1. For a multi-view scenario a camera network is used where 4 cameras N1,N2, N3, and N4 are observing the area from different viewpoints. M1, M2, M3, and M4 are midpoint of each side. The overlapping views are there, so the concise summary from multiple views are required to be generated by using the correlations in each view.

In this paper, we suggest a camera placement strategy for closed room surveillance area. Deep neural networks are used to detect the existence of objects so that further analysis can be performed. Object detection results are then fed into keyframe extraction algorithms, which produce a summary video. Also, we propose a keyframe based technique by preserving intra-view and inter-view correlation for creating video summaries. A View Quality (VQ) is identified, and a spatio-temporal shot graph is constructed and the summarization problem is formulated as a graph labeling task. A hypergraph is used to derive the spatio-temporal shot graph. Using random walks, the shot graph is partitioned and clusters of object-centered shots with overlapping or similar contents are found. We also provide a comparison based on F-Measure, Recall, and Precision to see how accurate each case is in context.

The paper's scope extends up to the placement of video cameras by which we can maximize the area under surveillance. By calculating the view quality, the salient views are identified based on the object under consideration. Keyframes are extracted from multiple views, and inter-view and intra-view keyframes are selected and video summaries from multiple views are generated.

The method overview is shown in Fig. 2. As shown, the decision about the placement of cameras for the surveillance area is taken. The position of the cameras ensures that the maximum surveillance area is covered, and we also get an overlapping region so we can have multiple views of the same events. Based on the view angle information and object under consideration of all the views, the keyframes are extracted for each view. The intra-view summary for each view is created and from the same, the inter-view summary is generated.

The section II addresses similar work in the literature, section III discusses the methods of camera placement, view angle computation, and video summarization, section IV discusses the results of the suggested approaches compared to existing state-of-the-art methods.

II. Related Works

Video summarization detailed review is present in [1] and [4]. T. Hussain et al. in [5] represented the detailed survey of multi-view summarization. Only some representative work is present here. Fu et al. in [6] first to propose the multi-view summarization method. To represent the multi-view summarization problem in graph theory, a spatio-temporal shot graph is created, and a random walk is used to cluster the event-center shot clusters. A multi-view event board and storyboard are presented for generating a multi-view video summary. Li et al. in [7] proposed a correlation map to model the correlation with attributes among keyframe importance, multi-keyframe, and to construct the map, and weighted correlations are computed. Li et al. in [8] proposed a motion-focusing method for keyframes extraction and summarized surveillance videos. J Almeida et al. in [9] present video summarization for online application for video summarization that operates directly in the compressed domain. They suggested the grouping of similar frames based on small intragroup differences and large intergroup differences.

Kuanar et al. in [10] use Delaunay graphs for clustering the keyframes for better content coverage in summary. By using the calculation of consecutive frames, they suggest that reciprocal information among two frames shows the connection between these frames. Using joint entropy calculation for successive frames, the mutual information is estimated. Mutual information between frame F_t frame F_{t-1} is represented by $MI(F_t, F_{t-1})$.

$$\frac{MI(F_t, F_{t-1})}{MI(F_{t-1}, F_{t-2})} + \frac{MI(F_t, F_{t-1})}{MI(F_{t+1}, F_t)} \leq 2(1-\epsilon) \tag{1}$$

where ϵ is a significant valley identification threshold. They then used colour and texture feature extraction techniques to achieve higher semantic interdependence between video frames in order to remove spatial redundancy between the frames; PCA was used to reduce dimensionality, and the Delaunay graph was constructed with the goal of preserving intra-cluster edges while removing inter-cluster edges.

The majority of the work in video summarization is based on offline video summarization. Ou et al. in [11] proposed on-line summarization for wireless video sensor networks. For the intra-view stage in [11], they use color and edge histogram for feature extraction; then, clustering is done by Gaussian Mixture Model (GMM) to group related content. The frame with more considerable weight is not

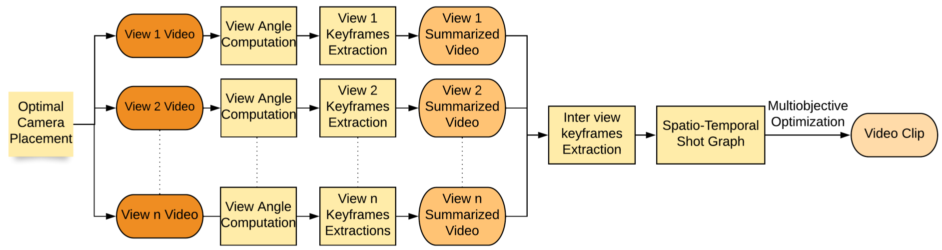


Figure 2. Multi-view summarization overview

selected in summary as that frame belongs to the frequently appeared cluster. For inter-view stage frame selection in [11], uses view selection, so if at time t, N cameras are capturing the area then the importance of view can be calculated as

$$c^* = \arg\max_c \{s_t^c|1, ..., N\} \quad (2)$$

where s_t^c is the important score of the view c at time t.

An automatic camera placement algorithms are proposed by A. van den Hengel, et al. in [12] and E. Yildiz, et al. in [13].

For summarizing multi-view videos, R. Panda et al. in [14], generates inter and intra-view similarities by modeling multi-view correlations using a sparse representative selection method. S. Liu et al. in [15] have created a video summary for finding suspicious movements in a building by visualizing object trajectories.

Applications of video summarization are in various disciplines like generating soccer match highlights [16], cattle behavior analysis from video [17], information retrieval and surveillance systems [18], multi-person tracking using representation learning [19], disaster management [20], and abnormal event detection [21]. Based on the application requirements, video summarization techniques can be segregated but are primarily classified into two types: keyframe extraction techniques and shot selection techniques. In [22], a combination of shot segmentation and keyframe selection is used to effectively summarize video content by automatically splitting the video stream into shots and extracting keyframes from the shots. V. Parikh et al. in [23] discussed about the key-frame extraction techniques for video summarization for close-room scenarios.

In recent years, a range of approaches are used to attain optimum results of video summarization. M. Ajmal et al. in [24] provides information on video summarization techniques in depth. DeMenthon et al. in [25], represented a video stream as a trajectory curve. Kawashima et al. in [26] generated important highlights of a baseball game by using content-based summarization. Li et al. in [27] patented the summarization technique based on a multiplicity of video clips. Padmavathi Mundur et al. in [28] expanded on standard clustering approaches that rely on input data by creating multi-dimensional point data from frame content and clustering using Delaunay Triangulation. The optimal camera placement algorithm is presented by Parikh et al. [29], in order to cover the most area under surveillance. S.K. Kuanar et al. in [30] uses a method of bipartite matching the correlation in a multi-view environment by using a texture, color, a visual bag of words and so on. R. Panda et al. in [31] work on shot-based video synopsis creation by identifying C3D attributes out of each shot.

For object detection, a number of techniques have been identified, including Color-based Determination [32], Template Matching [33], Background Saturation and Foreground Masking [34], Edge detection Techniques [35], Haar Featurization [36], and Cascade Classifiers [37]. By using Feature extraction and a series of ML and DL Models, identification of the objects can be determined using SVM and Histogram of Gradients [37] can also help in the identification process. On the other front, the recognition process can be a combination of the appearance-based, model-based, part-based, region-based, or contour-based approach.

Shen et al. in [38] suggest the automatic camera selection methods. Based on their work, View Quality has majored. The point o in Fig. 3 reflects the subject's head's Center of Gravity (COG). View of Angle is (θ, ϕ), where θ is the angle between the projection of camera optical axis on a plane (o,i,j) and subject's body orientation. At the same time, ϕ is the angle between a line passing through the camera center and COG of plane (o,i,j) and the subject's head.

III. PROPOSED METHODOLOGY

A. *CAMERA SETUP*

For generating summaries from multi-view, we first need to set up the cameras in such a manner that we have an overlapping view for our region of interest, so we have multiple views of the same event, and we can generate the summaries for the same. For this, we have an optimal

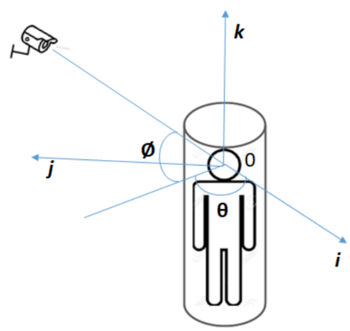


Figure 3. View Angle Interpretation

camera placement proposal wherein we have almost zero blind zones, and a particular region is covered with at least two cameras, as shown in Fig. 1.

Fig. 4 depicts the coverage of camera N in three dimensions. Center of gravity is shown by point G and Point V shows the camera V(x,y,z) position. The points A, B, C, and D in the FoV of V are computed by using the position of the video camera, horizontal Angle of View (AoV), and vertical AoV. Arbitrary point X is in the FoV of V and is required to be observed by V.

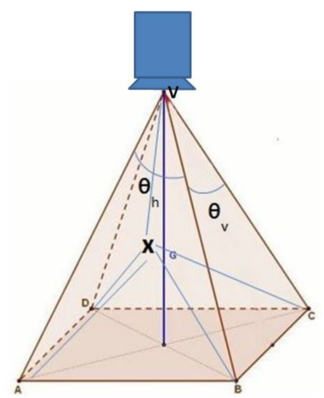


Figure 4. Coverage of a Video Camera

Algorithm 1 represents the systematic steps to place the cameras at intersections and junctions in a multi-view surveillance system. The cameras are placed in such a way that the overlapping region of the surveillance area is captured. Here midpoint M_j needs to be found for each side of the region LxB of A.

Algorithm 1 Optimal Camera Placement

$N_i(x, y, z)$

Require: Surveillance area A of size LxBxH
Cameras N_i, where i = 1 to 4

function: FieldOfView(N_i,M_j)
  Input: $N_i(x, y, z)$, M_j
  Output: TRUE/FALSE
  result = FALSE
  Find the coverage area V_i for camera N_i
  Find individually volume of four tetrahedrons $V_n^{M_j}$ formed by M_j with $N_i(x, y, z)$ as apex and each of the four sides of the coverage area of $N_i(x, y, z)$, where n =1 to 4 for point M_j
  Find volume $V_{base}^{M_j}$ of the pyramid formed with M_j as apex
  Find the total volumes as: $V_{total}^{M_j} = V_{base}^{M_j} + \sum V_n^{M_j}$, where n = 1 to 4
  if $V_{total} == V_i$
result = *TRUE*
end if
  return *result*
end function
Initialization
Divide the region LxB of A into four equal regions R_i; where i = 1 to 4 and find midpoint M_j for each side
for i = 1 to 4 **do**
  Execute **function** FieldOfView(N_i,M_j) to find whether $N_i(x, y, z)$ captures the region $R_i(x, y, z)$
end for

B. Object Detection for Summary Generation

Object detection is useful in the context of multi-view surveillance systems, especially in the surveillance systems as object tracking can be helpful in scenarios like traffic rules offenders, anomaly detection, etc. Furthermore, a user or human expert is asked to give input to locate a particular object from the video obtained through object detection. Researchers nowadays prefer deep learning algorithms for object detection. Multiple pre-trained deep learning models for object recognition and object detection are available. Such as Region-based Convolutional Neural Networks(R-CNN) [39], Fast R-CNN [40], Faster R-CNN [41], Mobilenet [42], Single Shot Detector (SSD) [43], You Only Look Once (YOLO) [44], etc. For object detection, YOLO has been implemented and reviewed. YOLO can process 45 frames per second which is very fast. Then, keyframe extraction techniques are applied to identify the best keyframes for closed room scenarios. Similarly, Haar feature-based algorithms for object detection are also available.

C. Keyframe Selection

Various image processing techniques, such as colour histogram difference, can be used to detect shot transitions [45], pairwise pixel difference [46], and Edge Change Ratio [47][48][49]. As shown in Fig. 5 three videos are captured of the same scene from various angles. We have applied shot boundary detection for all these three videos. As it is evident from Fig. 5 that all the videos though it is captured of the same scene, still generates different shot boundaries. According to [1], video summary can be either a series of still images (keyframes) or a series of moving images (shots or video skims). As shown in Fig. 5 the shots are overlapping, which is visible if we compare shot 1 of 1 with shot 2 of 1 and shot 3 of 1. So if we only consider shots of videos for generating multi-view summaries than either we have to take shot 1 of 1 or shot 1 of 2 or shot 1 of 3 for generating summary and if we choose shot 1 of 3, and next we pick shot 2 of 2 then we are losing intermediate shot between t1 and t2. Thus we need to convert the shots into keyframes and take the keyframes for generating the video summaries. Hence we have opted for implementing keyframe selection techniques for multi-view video summarization.

The goal of keyframe extraction is to map a video's complete content into a series of representative frames known as keyframes [50]. For the object under consideration, various experiments are performed for keyframe selection techniques to identify which keyframe selection techniques will be best suited for closed room scenarios.

The sufficient content change method can be best described using the following equation:

$$f_k = argmin\left\{C(f_t, f_i) > \epsilon, i < n\right\} \quad (3)$$

$$f_{kj+1} = argmin\left\{C(f_{kj}, f_i) > \epsilon, i < n\right\} \quad (4)$$

where, f_i = input frame, f_t = threshold frame, C = Content change function, ϵ= threshold, n = number of frames. The object identification algorithm is used to construct the content change function for this method.

The maximum frame coverage method can be best describe using following equation

$$r_1, r_2, .., r_k = argmax\left\{(r_i)C_{r_1}(\epsilon) \cup C_{r_2}(\epsilon) \cup .. \cup C_{r_k}(\epsilon) = V\right\} \quad (5)$$

where, r_i = Key-frame set , argmax()= maximum argument function checked on the each r_i frame, C_{r_j}= union of probabilistic belonging to the C^{th} class, ϵ= probability value of the defined class, and V = Video.

Minimum correlation can be best describe using following equation

$$r_1, r_2, .., r_k = argmin\left\{[r_{ik}]Corr(f_{r_1}, f_{r_i+1})\right\} \quad (6)$$

where, r_i = Key-frame set , minimum argument function argmin() checked on the each combination of r_{ik} frame, Corr = correlation function on f_{r_1} and f_{r_i+1} frames.

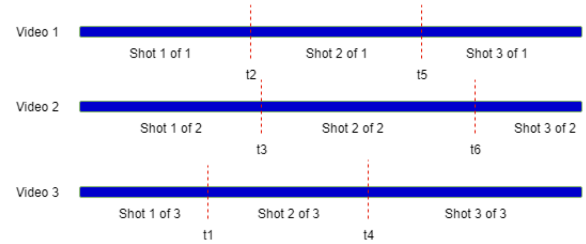


Figure 5. Number of Shots of Videos

Along with these three curve simplification and clustering method is also implemented and observed.

D. Video Summarization

The literature for video summarization is mainly focused on off-line video summarization; there is very little literature that talks about on-line video summarization. Ou et al. in [11] proposed on-line summarization in a wireless video sensor network. They break the task into two parts intra-view phase and the inter-view phase. In their approach for the intra-view phase, for every frame, they extract the representative feature. Then they cluster the features by using the Gaussian Mixture Model (GMM) for grouping the similar contents and removing the redundant information. The frames that belong to the cluster with a small weight represent the rare event, and this frame should be selected for summarization. Then the said frame is passed for the inter-view stage. Their main aim is to generate an energy-efficient on-line video summarization system. However, they have not taken care of multi-objective optimization. In our approach, instead of selecting the frames from the videos, we have identified the best view, which is capturing the object into consideration. Shen et al. in [38] suggest the automatic camera selection methods. Based on their work, we major the View Quality (VQ) as below:

$$VQ = Q_i \cap (\omega_\theta * (1 - |\frac{\theta_i}{\pi}|) + \omega_\phi * (1 - |\frac{2\phi_i}{\pi}|) + \omega_l * (1 - \frac{D_i}{D_{Bi}})) \quad (7)$$

Here $Q \in \{0, 1\}$ represents whether the subject is occluded or out of view, θ represents the orientation angle of the camera to the subject body, and D represents the distance between subject and camera and D_B represents best distance. ω represents weight which depends on application. $\theta_i \in (-\pi, \pi], \phi_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]$

Algorithm 2 depicts systematic steps for view quality calculation.

For off-line video summarization, the keyframes are selected based on sufficient content change method for the

Algorithm 2 View Angle Calculation

Require: At least two cameras which captures a common view at substantially at the same time

1: Initialization
2: Foreground and background detection
3: Occlusion Check, Distance Detection and Angle of View (AoV) Detection
4: View Quality is calculated by
$VQ = Q_i \cap (\omega_\theta * (1 - |\frac{\theta_i}{\pi}|) + \omega_\phi * (1 - |\frac{2\phi_i}{\pi}|) + \omega_l * (1 - \frac{D_i}{D_{Bi}}))$

intra-view stage, however other frame selection techniques can also be applied based on the application. Algorithm 3 shows the proposed systematic method for creating intra-view and inter-view video summaries.

Algorithm 3 Video Summarization

Require: At least two videos of a common view at substantially at the same time

1: Initialization
2: **for** i = 1 to N **do**
3: ConvVideotoFrames($view_i$,$path$)
4: **end for**
5: **for** i = 1 to N **do**
6: ViewQuality($view_i$)
7: **end for**
8: frame importance based on object detection is calculated by
$f_k = argmin\Big\{C(f_t, f_i) > \epsilon, i < n\Big\}$
$f_{kj+1} = argmin\Big\{C(f_{kj}, f_i) > \epsilon, i < n\Big\}$
where, f_i = input frame, f_t = threshold frame, C = Content change function, ϵ= threshold, n = number of frames.
9: Generate the Video Similarity Graph (VSG)
10: Generate the Spatio-Temporal Graph
11: Apply clustering method to obtain the cluster
12: To obtain the final cluster, propagate the original cluster result into the VSG using the Random Walk
13: The cluster significant factor is used to organise the keyframes

E. GRAPH CONSTRUCTION

For the representation of a multi-view video in graph theory, a Spatio-temporal frame graph is used. Object-based frame clustering is done via random walks is used for multi-view summary generation. The construction of the Spatio-temporal frame graph in a multi-view summarization is significant. The graph is constructed by parsing the input video. The importance of frames is calculated based on the amount of information in the frame for a given object. The said frames are stored as a result. These frames are then employed as nodes in a graph, with the importance value serving as a node value. Many works of literature use a Gaussian entropy fusion model for importance evaluation, which was created to fuse a set of video features. Different attributes, such as content similarity and temporal adjacency, are having diverse correlations in the frames. For systematically characterizing the correlation among frames, a hypergraph is used. A hyperedge of hypergraph can be used to link a subset of nodes. The hyperedge represents a kind of correlation among multi-view frames. The hypergraph is transformed into a spatio-temporal graph, with frame correlations across multiple views mapped to edge weights.

The random walk is employed to cluster the object-centered similar frames for implementing multi-view summarization. Multi-objective summarization is achieved by using object-centered similar frames as an anchor points, which caters the requirements of different users along with multi-level summarization.

IV. Experimental Results and Discussion

A. Simulation Environment and Parameters

The OMNET++ Network Simulator was used to simulate the Algorithm 1. The simulation parameters used for validating the results are shown in Table I. Since four equal parts are created from the surveillance area, the network topology of camera placement was configured by placing each camera either on the boundaries of the one-fourth portion or randomly within the surveillance area. For the best possible coverage of the area under observation, every camera's FOV should have two midpoints on adjacent sides.

B. Object Detection

The video was fed to the object detection algorithm, and objects are identified. The Fig. 6 show the person as an object (while in motion) detected using the algorithm. The target is to employ object detection for video summarization such that the foundation for all the keyframe extraction techniques parameters will be prediction outputs from the Object Detection Algorithm in use, and this will act as the threshold for deciding which frame, out of a set of frames will be abstracted as a keyframe.

Hence, if we take an object to be our key and set x as a threshold value which shows objects' presence, for the frame to classify as a keyframe, then only those frames where the key objects' presence probability will be greater than x, will be a keyframe.

To gain a clear understanding of what object detection is, and how it may be utilised for keyframe extraction, we must first understand classification and localization. The challenge of localization is viewed as a regression problem. The input

Parameter	**Value**
Simulation Time	300 s
Area under Surveillance	25m x 20m x Depth of Area under surveillance(mentioned below)
Depth of Area under Surveillance	10m to 15m
No. of Cameras	4
Angle of View (AoV) of Cameras	90 to 120
Focal Length	4.0 mm
Deployment of Cameras (Co-ordinates are in the three-dimensional space)	Random

TABLE I. Optimal Camera Placement Algorithm's Simulation Parameters

frame is sent over two separate networks. One of the network models is trained for class identification, which categorizes the frame class. The other network model calculates x and y coordinates for the bounding box's left top for the classified object, as well as the box's width and height, to produce the best local maxima for the presence probability of the class object. As a result, the overall output is the target class value as well as a 4-valued set of attributes comprising the bounding box's height, width, and x, y position, as well as the presence probability with the highest proportion of matching features to total features, as determined by the proportion of matching features to total features.

Figure 6. Detection and Recognition a person as an object

C. Results

Various simulations were performed by placing cameras at the midpoints at the vertex where two boundaries are met, and at random places inside the surveillance area. The AoV and depth of fields was identical and fixed, but it is in the range specified in Table I.

It is clear from Table II that the best position of the cameras in a multi-view surveillance system was near the vertex joining boundaries of the surveillance area. The results show that a wider AoV, along with a higher depth of field covers the surveillance area better by leaving less than 5% of the area uncovered. Also, from the results, we can infer that a wider AoV along with a higher depth of field has a maximum overlapping area, so multiple views of the same area can be captured and processed for a multi-view summary generation. It is evident from the result that when the cameras are placed at the vertices where two boundaries are joined maximum area is covered with the maximum overlapping area, which is required for reducing the blind zone and occlusion.

Table III shows the performance comparison with baseline multi-view methods applied to two indoor multi-view datasets, namely office and lobby datasets As seen from the result obtained, View Quality Based Multi-view Video Summarization (VQBMVS) produces a similar result as BiparitieOPF and Embedded Sparse Coding for both the datasets. As seen from the result obtained, our method, View Quality Based Multi-view Video Summarization (VQBMVS) produces a similar result as BiparitieOPF and Embedded Sparse Coding for both the datasets. It is evident from Table III that a higher recall value illustrates that our method is better in retaining important information in the generated summary than RandomWalk, and RoughSets for both the datasets. Also, these datasets have multiple views, so in our approach, before generating the keyframes, we first find the view quality, and then the keyframes are generated. Hence the computation processing was greatly reduced without loss in the information.

V. Conclusion

The proposed approach can be used to position multiple surveillance cameras at intersections and junctions which covers maximum area under the surveillance by eliminating or reducing the number of blind zones. The proposed approach aims to optimize subject visibility while minimizing occlusion in the surveillance environment. A summary of large-sized and lengthy videos of closed room scenarios is generated using a multi-view summarization method. The summary generation based on view-quality greatly reduces computational processing. The proposed methods reduce the computational costs, and the number of cameras required for surveillance which provide enhanced result for the production of a multi-view summary.

Angle Of View (degree)	Depth Of Field (m)	Camera Placement	Area Covered (%)	Overlapped Area (%)
90	10	Random-inside the Surveillance Area	29	25
90	11		34	30
100	12		45	40
100	13		50	46
105	14		48	43
100	15		53	50
90	10	On the borders on one-fourth part in Surveillance Area	43	39
95	11		50	47
110	12		61	56
100	13		70	66
105	14		73	69
120	15		74	70
95	10	Adjacent to midpoints	23	22
110	11		28	26
120	12		39	38
100	13		38	37
100	14		27	26
95	15		24	22
95	10	At the vertices where two boundaries are joined	94	90
110	11		97	95
105	12		95	93
120	13		98	96
105	14		95	93
120	15		97	93

TABLE II. Optimal Camera Placement Simulation Results

Methods	Office			Lobby		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
RandomWalk [6]	100	61	76.19	100	77	86.81
RoughSets [51]	100	61	76.19	97	74	84.17
BipartiteOPF [30]	100	69	81.79	100	79	88.26
Embedded Sparse Coding [14]	100	70	81.79	100	79	88.26
VQBMVS	100	70	81.77	100	79	88.26

TABLE III. Performance comparison with baseline multi-view methods applied on two multi-view datasets.

References

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 3, no. 1, p. 3, 2007.

[2] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video manga: generating semantically meaningful video summaries," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 383–392.

[3] R. Panda, A. Dasy, and A. K. Roy-Chowdhury, "Video summarization in a multi-view camera network," pp. 2971–2976, 2016.

[4] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[5] T. Hussain, K. Muhammad, W. Ding, J. Lloret, S. W. Baik, and V. H. C. de Albuquerque, "A comprehensive survey of multi-view video summarization," *Pattern Recognition*, vol. 109, p. 107567, 2020.

[6] Y. Z. F. L. C. S. Yanwei Fu, Yanwen Guo and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, pp. 717–729, 2010.

[7] Y. G. Ping Li and H. Sun, "Multi-keyframe abstraction from videos," *International conference on Image Processing*, vol. 18, pp. 2473–2476, 2011.

[8] S. Y. C Li, Y Wu and T. Chen, "Motion-focusing key frame extraction

and video summarization for lane surveillance system," *Proc. of the 16th IEEE ICIP*, pp. 4273–4276, 2009.

[9] N. J. JurandyAlmeida and R. da S.Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, pp. 397–409, 2012.

[10] S. K. Kuanar, R. Panda, and A. S. Chowdhury, "Video key frame extraction through dynamic delaunay clustering with a structural constraint," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 1212–1227, 2013.

[11] S.-H. Ou, C.-H. Lee, V. S. Somayazulu, Y.-K. Chen, and S.-Y. Chien, "On-line multi-view video summarization for wireless video sensor network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 1, pp. 165–179, 2015.

[12] A. van den Hengel, R. Hill, B. Ward, A. Cichowski, H. Detmold, C. Madden, A. Dick, and J. Bastian, "Automatic camera placement for large scale surveillance networks," in *2009 Workshop on Applications of Computer Vision (WACV)*, Dec. 2009, pp. 1–6.

[13] E. Yildiz, K. Akkaya, E. Sisikoglu, and M. Y. Sir, "Optimal camera placement for providing angular coverage in wireless video sensor networks," *IEEE Transactions on Computers*, vol. 63, no. 7, pp. 1812–1825, Jul. 2014.

[14] A. D. Rameswar Panda and A. K. Roy-Chowdhury, "Embedded sparse coding for summarizing multi-view videos," *IEEE International Conference on Image Processing (ICIP)*, pp. 191–195, 2016.

[15] S. Liu and S. Lai, "Schematic visualization of object trajectories across multiple cameras for indoor surveillances," in *2009 Fifth International Conference on Image and Graphics*, Sep. 2009, pp. 406–411.

[16] B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 3. IEEE, 2003, pp. III–169.

[17] Q. Zhi, T. Saitoh, M. Nakajima, and T. Saitoh, "A development of content-based video summarization system using machine-learning and its application to analysis of livestock behavior," in *International Workshop on Advanced Image Technology (IWAIT) 2019*, vol. 11049. International Society for Optics and Photonics, 2019, p. 1104911.

[18] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient cnn based summarization of surveillance videos for resource-constrained devices," *Pattern Recognition Letters*, 2018.

[19] H. Wu, Y. Hu, K. Wang, H. Li, L. Nie, and H. Cheng, "Instance-aware representation learning and association for online multi-person tracking," *Pattern Recognition*, vol. 94, pp. 25–34, 2019.

[20] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[21] Y. Yuan, Y. Feng, and X. Lu, "Structured dictionary learning for abnormal event detection in crowded scenes," *Pattern Recognition*, vol. 73, pp. 99–110, 2018.

[22] C. Huang and H. Wang, "Novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[23] V. Parikh, J. Mehta, S. Shah, and P. Sharma, "Comparative analysis of keyframe extraction techniques for video summarization," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 14, no. 9, pp. 2761–2771, 2021.

[24] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video summarization: techniques and classification," in *International Conference on Computer Vision and Graphics*. Springer, 2012, pp. 1–13.

[25] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *Proceedings of the sixth ACM international conference on Multimedia*, 1998, pp. 211–218.

[26] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," in *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, vol. 1. IEEE, 1998, pp. 871–874.

[27] B. Li and J. B. Sampsell, "Summarization of baseball video content," Nov. 28 2006, uS Patent 7,143,354.

[28] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.

[29] V. Parikh, P. Sharma, V. Shah, and V. Ukani, "Optimal camera placement for multimodal video summarization," in *International Conference on Futuristic Trends in Network and Communication Technologies*. Springer, 2018, pp. 123–134.

[30] K. B. R. Sanjay K. Kuanar and A. S. Chowdhury, "Multi-view video summarization using bipartite matching constrained optimum path forest clustering," *IEEE Transactions on Multimedia*, vol. 17, pp. 1166–1173, 2015.

[31] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Transactions on Multimedia*, vol. 19, pp. 2010–2021, 2017.

[32] A. Vailaya, M. A. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image classification for content-based indexing," *IEEE transactions on image processing*, vol. 10, no. 1, pp. 117–130, 2001.

[33] J. P. Lewis, "Fast template matching," in *Vision interface*, vol. 95, no. 120123, 1995, pp. 15–19.

[34] C.-B. Liu and A. K. Bhattacharjya, "Systems and methods for generating background and foreground images for document compression," Aug. 24 2010, uS Patent 7,783,117.

[35] L. S. Davis, "A survey of edge detection techniques," *Computer graphics and image processing*, vol. 4, no. 3, pp. 248–270, 1975.

[36] B. Paterson and A. Lacoste, "Cs231a project final report character identification in tv series from partially labeled data," 2014.

[37] W.-C. Cheng and D.-M. Jhan, "Triaxial accelerometer-based fall detection method using a self-constructing cascade-adaboost-svm classifier," *IEEE journal of biomedical and health informatics*, vol. 17, no. 2, pp. 411–419, 2012.

[38] C. Shen, C. Zhang, and S. Fels, "A multi-camera surveillance system that estimates quality-of-view measurement," in *Image processing, 2007. ICIP 2007. IEEE international conference on*, vol. 3. IEEE, 2007, pp. III–193.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[40] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[41] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[43] A. Womg, M. J. Shafiee, F. Li, and B. Chwyl, “Tiny ssd: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection,” in *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 2018, pp. 95–101.

[44] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[45] R. Kasturi, “Dynamic vision,” *Computer Vision: Principles*, 1991.

[46] J. S. Boreczky and L. A. Rowe, “Comparison of video shot boundary detection techniques,” in *Journal of Electronic Imaging*, vol. 5, April 1996, pp. 122–128.

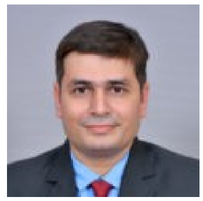
[47] T. Liu, H.-J. Zhang, and F. Qi, “A novel video key-frame-extraction algorithm based on perceived motion energy model,” *IEEE transactions on circuits and systems for video technology*, vol. 13, no. 10, pp. 1006–1013, 2003.

[48] R. W. Lienhart, “Comparison of automatic shot boundary detection algorithms,” in *Storage and Retrieval for Image and Video Databases VII*, vol. 3656. International Society for Optics and Photonics, 1998, pp. 290–302.

[49] S. Konishi, A. L. Yuille, J. Coughlan, and S. C. Zhu, “Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues,” in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1. IEEE, 1999, pp. 573–579.

[50] S. D. Thepade and A. A. Tonge, “An optimized key frame extraction for detection of near duplicates in content based video retrieval,” in *2014 International Conference on Communication and Signal Processing*. IEEE, 2014, pp. 1087–1091.

[51] P. Li, Y. Guo, and H. Sun, “Multi-keyframe abstraction from videos,” in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 2473–2476.



Vishal Parikh Mr Vishal Parikh is working as an Assistant Professor in CSE Department. His research interests include Machine Learning, and Multimedia Communication. He has many publications to his credit. Prof. Parikh has conducted various workshops and lecture series at the department. He has also been invited for expert talk at various institutions.

Priyanka Sharma Dr Priyanka Sharma is a Vice President Projects - Artificial Intelligence at Samyak Infotech Pvt. Ltd. at Ahmedabad. She is NVIDIA Deep Learning Ambassador and has over 21 years of teaching and industrial experience. Number of research projects funded by GUJCOST-DST, IPR, Department of Atomic Energy, and Shastri Indo-Canadian Research Grant have been completed by her. She has also served as the NVIDIA Research Center’s Principal Investigator (2014-16). Dr Sharma is also collaborating with Gujarat Cancer Society, Medical and Research Center on Covid-19 based research.