Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.032506
Article

A Personalized Video Synopsis Framework for Spherical Surveillance Video

S. Priyadharshini* and Ansuman Mahapatra

Department of Computer Science and Engineering, National Institute of Technology Puducherry, NITPY Campus, Thiruvettakudy, Puducherry 609609, India
*Corresponding Author: S. Priyadharshini. Email: priya81818@gmail.com
Received: 20 May 2022; Accepted: 13 July 2022

Abstract: Video synopsis is an effective way to easily summarize long-recorded surveillance videos. The omnidirectional view allows the observer to select the desired fields of view (FoV) from the different FoV available for spherical surveillance video. By choosing to watch one portion, the observer misses out on the events occurring somewhere else in the spherical scene. This causes the observer to experience fear of missing out (FOMO). Hence, a novel personalized video synopsis approach for the generation of non-spherical videos has been introduced to address this issue. It also includes an action recognition module that makes it easy to display necessary actions by prioritizing them. This work minimizes and maximizes multiple goals such as loss of activity, collision, temporal consistency, length, show, and important action cost respectively. The performance of the proposed framework is evaluated through extensive simulation and compared with the state-of-art video synopsis optimization algorithms. Experimental results suggest that some constraints are better optimized by using the latest metaheuristic optimization algorithms to generate compact personalized synopsis videos from spherical surveillance videos.

Keywords: Immersive video; non-spherical video synopsis; spherical video; panoramic surveillance video; 360° video

1 Introduction

Spherical videos are also known as omnidirectional, 360°, or panoramic videos. They are recorded using a spherical camera that captures the environment on a spherical canvas. Due to the unlimited FoV, the process of identifying key events is challenging. Fig. 1 illustrates an observer viewing three different interested FoVs from the input spherical surveillance video independently. The object-based video synopsis approach is handy to solve this issue. Despite tremendous efforts devoted to the object-based video synopsis on the non-spherical videos [1–6], they cannot be directly applied to the spherical videos. Traditionally, 360-degree video summarization offers fixed FoV-based summarization, the spherical video summarization in a personalized manner is not focused. This motivates to generate an object-based video synopsis for spherical surveillance video in a personalized manner.

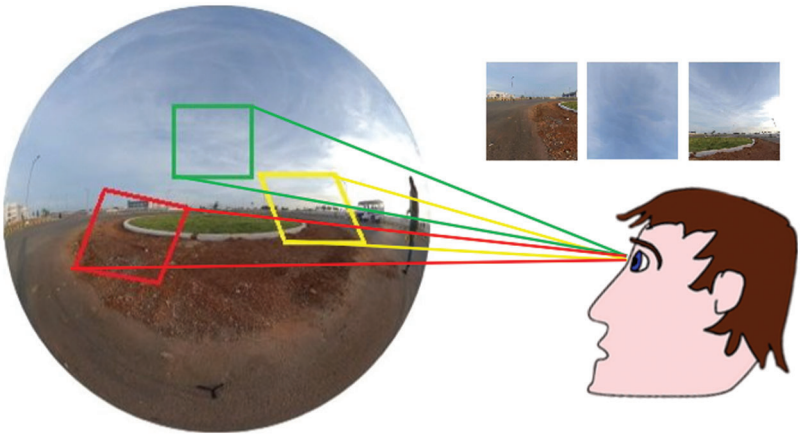


Figure 1: An illustration of an observer viewing the three different interested FoVs in the spherical surveillance video

The novelty of the proffered framework is that it eliminates FOMO and displays the viewer constraint number of objects per frame. FOMO refers to the observer's concerns about missing the salient part occurring outside the observer's viewing angle [7]. In this work, our contribution and focus are on the process of grouping objects and rearranging the personalized tubes in the recorded spherical surveillance video. Therefore, we used state-of-the-art methods for preprocessing, action recognition, viewport prediction, and tube stitching.

The key contributions of this work are:

1. A novel **personalized video synopsis approach** is proposed to generate dynamically non-spherical synopsis video for the spherical surveillance video to eliminate FOMO.

2. An **action recognition module** is incorporated into the proposed framework to prioritize the actions based on their importance.

3. To precisely understand the scene in the synopsis video the suggested framework provides the observer constraint **number of objects** for each frame in the non-spherical synopsis video.

4. The comparative **performance analysis of the proposed framework using the latest metaheuristic optimization algorithms** with the state-of-art video synopsis approach is performed for the generation of personalized non-spherical synopsis videos.

The remainder of this work is organized as follows. In Section 2, related works on 360 ° video summarization and non-spherical synopsis videos are presented. In Section 3, our proposed framework for the generation of personalized synopsis video from spherical surveillance video is introduced. Section 4 presents the results and analysis. Finally, Section 5 concludes this work.

2 Related Works

This section discusses the literature review for 360° video summarization and the generation of classical synopsis video.

2.1 360° Video Summarization

Su et al. [8] proposed autocam, a data-driven methodology to provide automatic cinematography in panoramic videos. Su et al. [9] generalize the introduced Pano2Vid task by allowing it to control the FoV dynamically. It uses a coarse to fine optimization process. Hu et al. [10] proposed a deep learning-based

approach, namely, a deep 360 pilot for piloting in 360-degree sports videos. Based on the knowledge of the previous viewing angles, it predicts the current viewing angle. Yu et al. [11] generate a 360-degree score map to identify the suitable view for 360-degree video highlight. This approach outperforms the existing autocam framework. Lee et al. [12] suggested a memory network model to generate summarized non-spherical video. It uses two memories for already selected subshots and future subshots that are likely to be selected, respectively. Tab. 1 gives the summary of related works on 360° video summarization.

Table 1: Related works on 360° video summarization

Author and year	Model	Summarization technique		FoV glimpses
		Spatial	Temporal	
Su et al. 2016 [8]	AutoCam	Yes	No	198
Su et al. 2017 [9]	AutoCam with zooming	Yes	No	198
Hu et al. 2017 [10]	Recurrent Neural Network	Yes	No	–
Yu et al. 2018 [11]	Composition View Score	Yes	Yes	12
Lee et al. 2018 [12]	Past Future Memory Network	Yes	Yes	81

2.2 Non-Spherical Synopsis Video Methods

Pritch et al. [1] proposed two methods for synopsis video generation namely, low-level graph optimization and an object-based approach. Mahapatra et al. [2] proposed a synopsis generation framework for a multi-camera setup. The synopsis generation problem is formulated as a scheduling problem. An action recognition module is incorporated to recognize and prioritize the actions performed by the objects. Ahmed et al. [13] presented a methodology for synopsis generation concerning a user's query. Ghatak et al. [4] presented a hybridization of Simulated Annealing and Teaching Learning based Optimization algorithms to generate an efficient synopsis video. Ghatak et al. [5] suggested an improved optimization scheme using the hybridization of Simulated Annealing and Jaya algorithms for the generation of video synopsis. Namitha et al. [6] proposed a recursive tube grouping methodology to preserve interacting objects. A spatiotemporal cube voting approach is used to arrange the objects optimally. The length of the synopsis is minimized by introducing the length estimation method. The systematic review of the video synopsis method used in the non-spherical videos is given by [14–16]. Tab. 2 summarizes the video synopsis-related works on non-spherical videos.

Table 2: Related works on non-spherical synopsis video

Author and year	Application	Tube shifting	Action recognition
Pritch et al., 2008 [1]	Generates non-chronological synopsis video	Yes	No
Mahapatra et al., 2016 [2]	Activity-based video synopsis	Yes	Yes
Ahmed et al., 2019 [13]	Query based synopsis video	Yes	No
Ghatak et al., 2020 [4]	Generation of surveillance video synopsis	Yes	No
Ghatak et al., 2020 [5]	Generates consumer surveillance synopsis video	Yes	No
Namitha and Narayanan, 2020 [6]	Generates synopsis video with interaction preservation	Yes	No

3 Generation of Personalized Non-spherical Video Synopsis

The generation of personalized non-spherical synopsis video from the spherical surveillance video is illustrated in Fig. 2. Spherical videos can be generally projected in two ways [17]. This work is based on the equirectangular projection [18] of spherical video. It comprises seven steps. The steps involved in the generation of a personalized spherical synopsis video are explained in detail as follows,

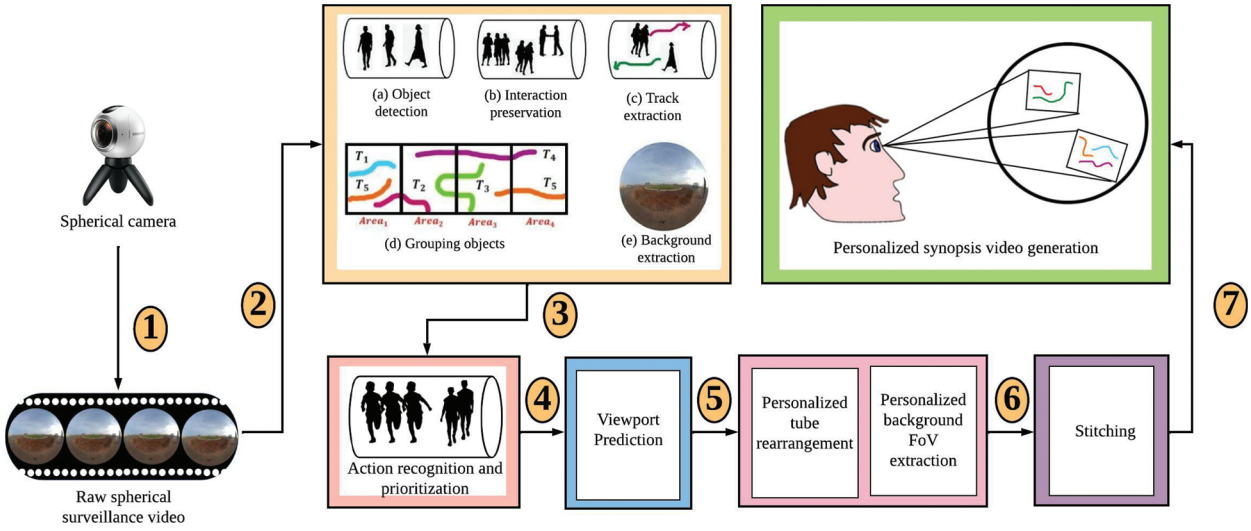


Figure 2: Proposed framework for the generation of personalized non-spherical synopsis video

Step 1: Recording the spherical surveillance video

This step uses a static spherical camera placed in the surveillance area to be monitored in a spherical environment. Due to the capability to record the scene in spherical form, the recorded video allows viewing the video with multiple FoVs.

Step 2: Pre-processing the raw spherical surveillance video

This step detects the moving objects in the raw spherical surveillance video using Faster Region-based Convolutional Neural Network (R-CNN) [19]. Followed by object tube grouping using a recursive tube-grouping algorithm [6]. The movement of the objects is tracked to obtain the track extraction of moving objects by using Deep Simple Online and Real-time Tracking (Deep SORT) [20]. The background of the raw input spherical video is extracted using the timelapse background video generation method [1]. Fig. 3 illustrates the pre-processing step of raw spherical surveillance video. The equirectangular area is partitioned into four areas with equal width of 1440, and then objects are grouped based on their occurrence concerning the area and stored in a lookup table as given in Algorithm 1.

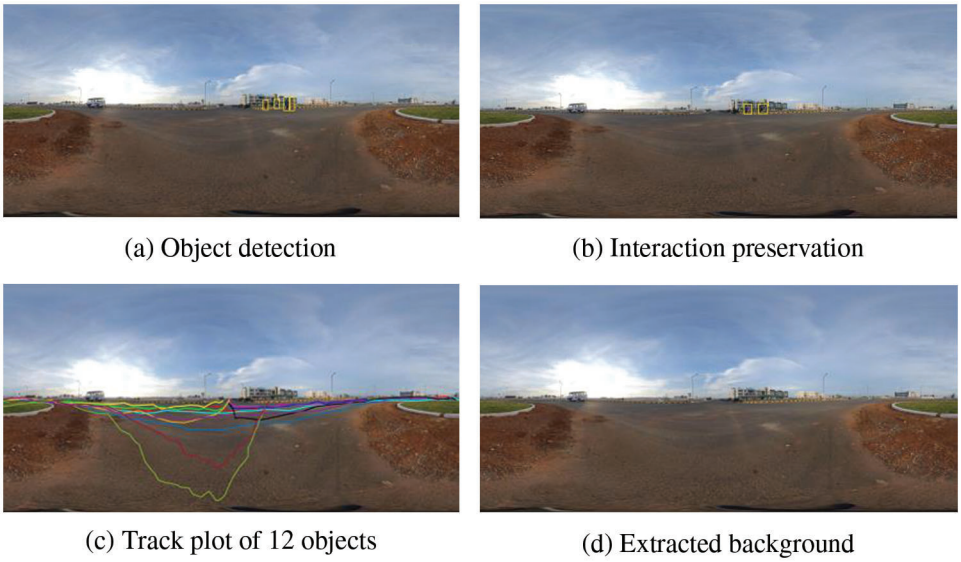
Algorithm 1 Lookup Table Generation

1. **procedure** GENERATE(Area Interval)
2. m←10% of object tube length
3. **for all** k←1:m:n do
4. $T_{partition}(k)=(x_k,y_k)$ of object track T_i
5. **if** $T_{partition}(k)$ lies within [0,A1] **then**

(Continued)

Algorithm 1 (continued)

6. Insert T_i under the index 1 in the Lookup_Table
7. **else if** $T_{partition}(k)$ lies within [A1,A2] **then**
8. Insert T_i under the index 2 in the Lookup_Table
9. **else if** $T_{partition}(k)$ lies within [A2,A3] **then**
10. Insert T_i under the index 3 in the Lookup_Table
11. **else**
12. Insert T_i under the index 4 in the Lookup_Table
13. **end if**
14. **end for**
15. **return** Lookup_Table with unique list of tubes in all index
16. **end procedure**



(a) Object detection (b) Interaction preservation

(c) Track plot of 12 objects (d) Extracted background

Figure 3: Pre-processing of raw spherical surveillance video

Step 3: Recognizing and prioritizing the actions performed

In this step, the raw spherical video is converted into a normal FoV video by using rectilinear projection [21] and segmented into a sequence of overlapping P-frames of long action segments. The time-series features of these sequence frames are extracted using CNN trained using a pre-trained network, Densenet 201 [22]. KTH dataset [23] is used to train the CNN model. Long Short Term Memory (LSTM) [24] is used to perform the sequence classification. Once the action sequences are classified they are prioritized based on the importance of the action performed [2]. This work involves only two actions namely, running and walking. Running is given higher priority compared to walking.

Step 4: Predicting the future viewports of the viewer

To generate a personalized synopsis video, viewport prediction of the future frames is vital. For training, the dataset given by [25] is used. Using imtool in MATLAB, the cartesian coordinates of the moving objects in the viewport for all frames in the raw spherical surveillance video is simulated on the desktop for testing purpose. Spherical coordinates such as pitch and yaw angles are computed using the simulated cartesian coordinates. The viewport prediction using both the position and content data in the spherical video was introduced by [26]. Densenet 201 [22] is used to extract the features of all the frames in the raw input video, and the upcoming viewports are predicted using LSTM [24]. This step outputs the future viewpoints that the observer will likely view while watching the synopsis video.

Step 5: Personalizing tube rearrangement and FoV background extraction

Once the viewport is predicted, the corresponding FoV background is extracted. The objects within the predicted viewport are selected dynamically from the lookup table created in Step 1. It is performed as per Algorithm 2. The selected objects undergo an optimization process to identify optimal rearrangement of tubes in the synopsis video.

Algorithm 2 Lookup Table Search

1. **procedure** SEARCH
2. **if** V_P^f lies within [0,A1] **then**
3. return the list of tubes in the index 1
4. **else if** V_P^f lies within [A1,A2] **then**
5. return the list of tubes in the index 2
6. **else if** V_P^f lies within [A2,A3] **then**
7. return the list of tubes in the index 3
8. **else if** V_P^f lies within [A3,A4] then
9. return the list of tubes in the index 4
10. **else**
11. return the union result for list of tubes // When V_P^f lies between two Area Interval
12. **end if**
13. **return** List of tubes that lies in the interested viewport
14. **end procedure**

The energy cost function to be minimized is,

$$E = k_1A + k_2C + k_3T + k_4L + k_5S - k_6R_A \tag{1}$$

Activity loss cost (A) ensures that no activities are lost by adding a penalty. Collision cost (C) adds a penalty for virtual collisions of objects caused by tube shifts. Temporal inconsistency (T) and synopsis length cost (L) add a penalty for not having the temporal consistency and shorter length in comparison with the raw input video respectively. The show cost (S) adds a penalty for showing more than the observer specified the number of objects per frame. The action recognition cost (R_A) adds a penalty for not including a maximum number of high-priority activities. These individual cost functions are defined as follows,

$$A = \sum_{j=1}^{m} O_j \tag{2}$$

where O is the set of all objects in the tube varying from 1 to m. If all activities are preserved then activity loss cost is 0.

$$C = \sum_{a=1}^{m} \sum_{b=1}^{n} Area(bbox(O_a) \cap bbox(O_b)) \tag{3}$$

where O_a, O_b denotes the two temporally shifted objects a and b in the synopsis.

$$T = \sum_{j=1}^{m} abs\left(order\left(o_j\right) - order\left(O_j\right)\right) \tag{4}$$

where o_j and O_j are the object tubes from the raws input and synopsis video respectively.

$$L = Length(Syn) \tag{5}$$

where Syn is the synopsis video.

$$S = \sum_{j=1}^{m} O_i(F) \tag{6}$$

where m denotes the total number of objects per frame F in the spherical synopsis video in this work utmost seven objects per frame is shown.

$$R_A = \sum_{j=1}^{m} PriorityScore_{O_j} \tag{7}$$

It is the sum of the priority score of each object in the synopsis.

It is solved using various latest optimization algorithms such as Aquila Optimizer (AO) [27], Archimedes Optimization Algorithm (AOA) [28], Dynamic Differential Annealing Optimization (DDAO) [29], Giza Pyramids Construction (GPC) [30], Heap-Based Optimizer (HBO) [31], Hybrid Whale Optimization with Seagull Algorithm (HWSOA) [32] as well as existing works on video synopsis such as Hybrid Simulated Annealing-Jaya (HSAJaya) [5], Hybridization of Simulated Annealing and Teaching Learning based Optimization (HSATLBO) [4], and Simulated Annealing (SA) [1]. The personalized tube rearrangement is given in Algorithm 3.

Algorithm 3 Personalized Tube Rearrangement

1. **procedure** Rearrange (T_i^f)
2. Generate initial population
3. Evaluate the objective function for initial population
4. Update the fitness value with the best value
5. Select the optimal rearrangement result
6. **return** optimal tube rearrangement for T_i^f
7. **end procedure**

Step 6: Stitching the object tubes to the FoV background

In this step, the optimal tube shifting results obtained from optimizing the multi-objective function are used. The objects are stitched to the FoV background based on the shifting results using Poisson image editing [33].

Step 7: Generating the personalized synopsis video

After stitching, the personalized synopsis video is generated in this step. Algorithm 4 gives the complete workflow of generating a personalized synopsis video.

Algorithm 4 Personalized Synopsis Video Generation

Input: Objects Tubes $T_i^f \leftarrow \left[x_i^f, y_i^f, w_i^f, h_i^f\right]$, $i \in 1$ to object tube length;
f is the spherical frame number; Area Interval$\leftarrow[A_1, A_2, A_3, A_4]$;
Future predicted viewpoints $V_P^f \leftarrow \left(\theta_P^f, \phi_P^f\right)$ where $\theta_P^f \in \{-90°, \ldots, 90°\}$ and $\phi_P^f = \{-180°, \ldots, 180°\}$;

Output Personalized synopsis video P_S;

Initialization Personalized FoV background FoV_P is empty;
$S_B \leftarrow$ Spherical background;

1. **for all** Object track T_i do
2. Generate(Area Interval)
3. **end for**
4. **for all** V_P^f do
5. Search in Lookup Table
6. Filter the object tubes that lies in the V_P^f
7. REARRANGE(T_i^f)
8. $FoV_P^f \leftarrow$ Extracted Background of V_P^f from S_B
9. Stitch rearranged object tubes to FoV_P^f
10. **end for**

4 Results and Analysis

Due to the unavailability of real-time spherical surveillance video, in this work Insta360 ONE X is used to record a spherical surveillance video for 01:03:11 (HH:MM:SS) from the National Institute of Technology Puducherry with 24fps. The recorded raw spherical video has a resolution of 5760 × 2880 and includes 110 spherical objects. Among them, 20 are interacting objects. The optimization algorithms used experiments with the number of population as 10 and the number of iterations as 100. A comparative analysis of the state-of-the-art metaheuristic optimization algorithms such as AO [27], AOA [28], DDAO [29], GPC [30], HBO [31], and HWSOA [32] and the existing synopsis generation optimization algorithms like HSAJaya [5], HSATLBO [4], and SA [1] are performed. Evaluation metrics used are a non chronology, collision, and inclusion rate.

a. Non chronology rate (N_R):

It is the rate of the sum of all objects that are not in chronological order (o_d) to the total number of objects (T_O).

$$N_R = \frac{o_d}{T_O} \tag{8}$$

b. Collision rate (C_R):

It is the rate of collision that occurred due to the temporal shifting of two objects.

$$C_R = \begin{cases} Coll_S - Coll_I & if\ STime_I = STime_S \\ Coll_S & otherwise \end{cases} \tag{9}$$

where $Coll_S$ and $Coll_I$ are the area of intersection between two tubes while $STime_I$ and $STime_S$ are the start time of two tubes in the synopsis and raw input video respectively.

$$Coll(X, Y) = \sum_{f \in f_X \cap f_Y} Area\left(box\left(X^f\right) \cap box\left(Y^f\right)\right) \tag{10}$$

where $Area\left(box\left(X^f\right) \cap box\left(Y^f\right)\right)$ determines the intersection area that is common to both tube X and Y in the spherical frame f.

$$STime(X, Y) = StartingTime(X) - StartingTime(Y) \tag{11}$$

c. Inclusion rate (I_R):

It is the rate of the number of high priority activities retained in the synopsis video compared to the total number of high priority activity in the raw input video.

$$I_R = \frac{n(P_A)}{N(P_A)} \times 100 \tag{12}$$

where $n(P_A)$ is the number of high priority activities included in the synopsis video and $N(P_A)$ is the total number of high priority activities in the raw input video.

Fig. 4 illustrates the action recognition for prioritizing important actions. Tab. 3 presents the analysis of individual costs used in the process of generating a personalized synopsis video. Fig. 5 illustrates the analysis of optimization algorithms for the proposed work. AO [27] and GPC [30] provide minimum collision and temporal inconsistency while SA [1] and DDAO [29] provide minimum synopsis length and show the observer a specified number of objects per frame respectively. SA [1] provides better results for the preservation of maximum important actions and AO [27] converges faster than other optimization algorithms. Fig. 6 presents the personalized synopsis video generated by using AO [27] with minimum collision.

(a) Pedestrian running (b) Pedestrian walking

Figure 4: Action recognition for prioritizing important actions

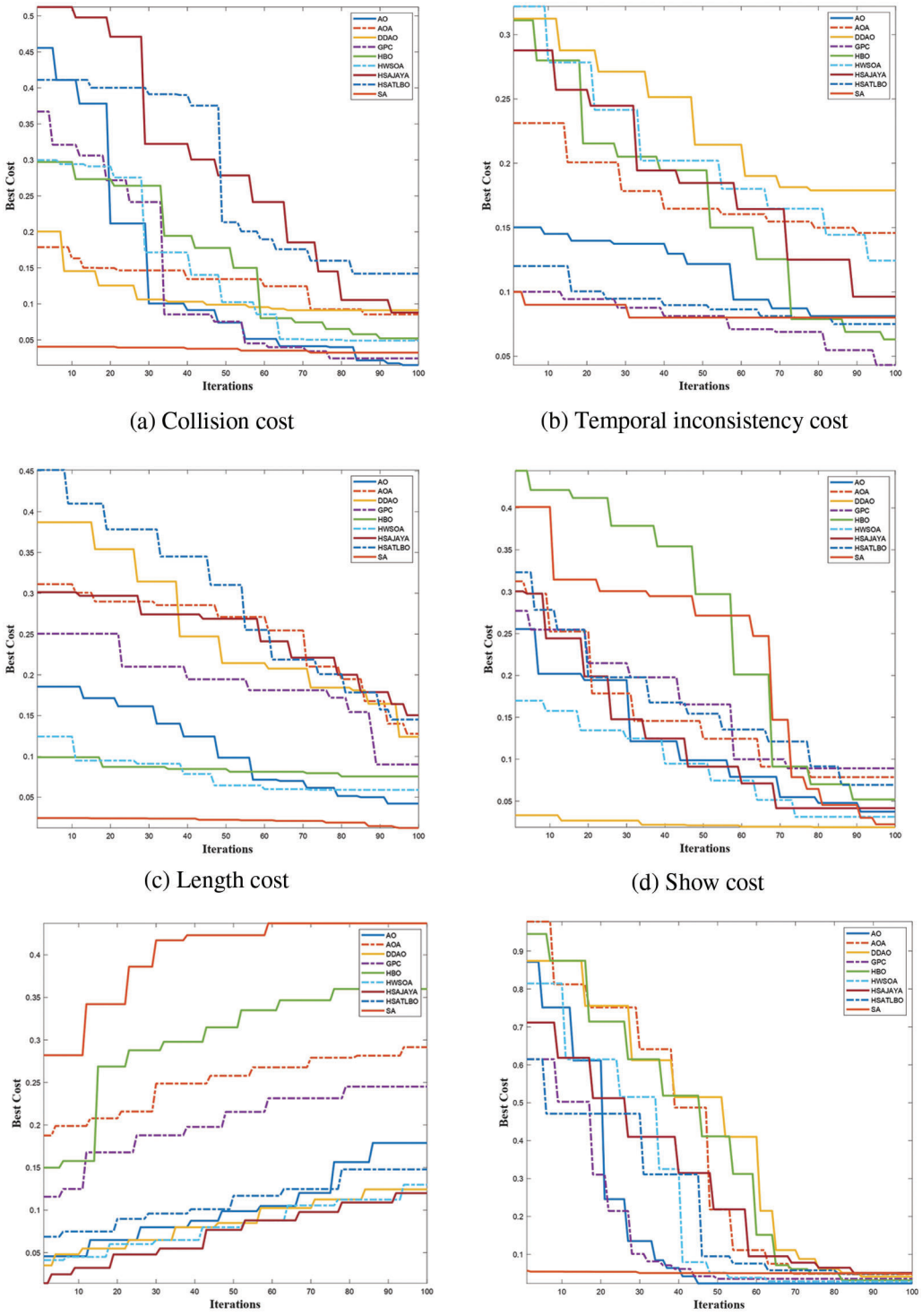
Table 3: Analysis of individual cost in the objective function

	Algorithm used	Activity loss	Collision	Temporal inconsistency	Length	Show	Action recognition
Proposed work	AO [27]	0	0.0150	0.0812	0.0421	0.0374	0.1789
	AOA [28]	0	0.0854	0.1458	0.1277	0.0785	0.2914
	DDAO [29]	0	0.0912	0.1789	0.1240	0.0187	0.1243
	GPC [30]	0	0.0243	0.0432	0.0901	0.0891	0.2451
	HBO [31]	0	0.0521	0.0631	0.0754	0.0519	0.3599
Existing work	HWSOA [32]	0	0.0490	0.1243	0.0589	0.0312	0.1298
	HSAJaya [5]	0	0.0880	0.0963	0.1505	0.0415	0.1199
	HSATLBO [4]	0	0.1420	0.0750	0.1452	0.0693	0.1478
	SA [1]	0	0.0323	0.0800	0.0122	0.0222	0.4370

Tab. 4 presents the performance metrics for the generated personalized synopsis video with duration D in minutes. SA [1] generates synopsis length with a shorter duration whereas, GPC [30] and AO [27] provide better results for non chronology rate and collision rate respectively. SA [1] provides better results for inclusion rates as 60.00%, 68.57%, 82.86%, 85.71%, and 97.14% for the synopsis video of duration for the five cases such as 3, 6, 9, 12, and 15 min respectively.

Table 4: Performance metrics of the generated personalized synopsis video

	Algorithms used	D (mm:ss)	N_R	C_R	$I_R(\%)$				
					3 min	6 min	9 min	12 min	15 min
Proposed work	AO [27]	18:54	0.1378	0.0271	57.14	62.86	74.29	80.00	91.43
	AOA [28]	20:02	0.1989	0.0793	31.43	34.29	40.00	45.71	62.86
	DDAO [29]	19:57	0.2017	0.1279	34.29	37.14	42.86	54.29	68.57
	GPC [30]	19:41	0.0712	0.0322	42.86	45.71	48.57	62.86	74.29
	HBO [31]	19:22	0.0915	0.0745	48.57	51.43	54.29	68.57	80.00
Existing work	HWSOA [32]	19:04	0.1463	0.0491	51.43	54.29	62.86	80.00	85.71
	HSAJaya [5]	21:04	0.1412	0.1024	17.14	20.00	28.57	31.43	40.00
	HSATLBO [4]	20:29	0.1245	0.1645	28.57	31.43	34.29	40.00	54.29
	SA [1]	18:09	0.1345	0.0475	60.00	68.57	82.86	85.71	97.14



(a) Collision cost (b) Temporal inconsistency cost

(c) Length cost (d) Show cost

(e) Recognized action cost (f) Convergence characteristics

Figure 5: Analysis of optimization algorithms

(a) Frame No: 14789 (b) Frame No: 19947 (c) Frame No: 20012

Figure 6: Personalized synopsis video generated by using AO [27] with minimum collision

5 Conclusion

This work introduces a personalized video synopsis framework for generating non-spherical video synopsis from spherical surveillance videos. The advantage of the proposed work is that FOMO is eliminated while watching the spherical videos. Here, an object grouping algorithm is introduced that identifies and groups objects based on the area of occurrence and then divides the objects into four groups. In addition, a personalized tube rearrangement algorithm was proposed. This algorithm aims to perform a tube shift of an object within the viewer's point of view. The action recognition module further reduces the synopsis length by prioritizing important actions. Experimental results and analysis show that the proposed framework offers a potential improvement in collision, temporal consistency, and show cost over the state-of-art video synopsis approach. It is also observed that the convergence rate of the prior art method is slow compared to the proposed framework. Finally, a hybrid optimization framework based on the analysis of the results performed can be considered for future work to condense spherical surveillance video.

Data Availability Statement: The dataset generated and analyzed during this study are available from the corresponding author on reasonable request.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Pritch, A. Rav-Acha and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[2] A. Mahapatra, P. K. Sa, B. Majhi and S. Padhy, "MVS: A multi-view video synopsis framework," *Signal Processing: Image Communication*, vol. 42, pp. 31–44, 2016.

[3] A. Ahmed, S. Kar, D. P. Dogra, R. Patnaik, S. Lee *et al.,* "Video synopsis generation using spatio-temporal groups," in *Proc. ICSIPA*, Kuching, Malaysia, pp. 512–517, 2017.

[4] S. Ghatak, S. Rup, B. Majhi and M. N. Swamy, "An improved surveillance video synopsis framework: A HSATLBO optimization approach," *Multimedia Tools and Applications*, vol. 79, no. 7, pp. 4429–4461, 2020.

[5] S. Ghatak, S. Rup, B. Majhi and M. N. Swamy, "HSAJAYA: An improved optimization scheme for consumer surveillance video synopsis generation," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 2, pp. 144–152, 2020.

[6] K. Namitha and A. Narayanan, "Preserving interactions among moving objects in surveillance video synopsis," *Multimedia Tools and Applications*, vol. 79, no. 43, pp. 32331–32360, 2020.

[7] T. Aitamurto, A. S. Won, S. Sakshuwong, B. Kim, Y. Sadeghi *et al.,* "From fomo to jomo: Examining the fear and joy of missing out and presence in a 360 video viewing experience," in *Proc. CHI Conf. on Human Factors in Computing Systems*, Yokohama, Japan, pp. 1–14, 2021.

[8] Y. C. Su, D. Jayaraman and K. Grauman, "Pano2Vid: Automatic cinematography for watching 360degree videos," in *Proc. Asian Conf. on Computer Vision*, Taipei, Taiwan, pp. 154–171, 2016.

[9] Y. C. Su and K. Grauman, "Making 360 video watchable in 2D: Learning videography for click free viewing," in *Proc. CVPR*, Honolulu, HI, USA, pp. 1368–1376, 2017.

[10] H. N. Hu, Y. C. Lin, M. Y. Liu, H. T. Cheng, Y. J. Chang *et al.,* "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *Proc. CVPR*, Honolulu, HI, USA, pp. 1396–1405, 2017.

[11] Y. Yu, S. Lee, J. Na, J. Kang and G. Kim, "A deep ranking model for spatio-temporal highlight detection from a 360° video," in *Proc. AAAI*, Hilton New Orleans Riverside, New Orleans, Louisiana, USA, 32, 2018.

[12] S. Lee, J. Sung, Y. Yu and G. Kim, "A memory network approach for story-based temporal summarization of 360 videos," in *Proc. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1410–1419, 2018.

[13] S. A. Ahmed, D. P. Dogra, S. Kar, R. Patnaik, S. C. Lee *et al.,* "Query-based video synopsis for intelligent traffic monitoring applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3457–3468, 2019.

[14] A. Mahapatra and P. K. Sa, "Video synopsis: A systematic review," in *High Performance Vision Intelligence*, Noida, India, pp. 101–115, 2020.

[15] K. B. Baskurt and R. Samet, "Video synopsis: A survey," *Computer Vision and Image Understanding*, vol. 181, pp. 26–38, 2019.

[16] S. Ghatak and S. Rup, "Single camera surveillance video synopsis: A review and taxonomy," in *Proc. ICIT*, Bhubaneswar, India, pp. 483–488, 2019.

[17] S. Priyadharshini and A. Mahapatra, "360degree user-generated videos: Current research and future trends," in *High Performance Vision Intelligence*, Noida, India, pp. 117–135, 2020.

[18] B. Ray, J. Jung and M. C. Larabi, "A low-complexity video encoder for equirectangular projected 360 video content," in *Proc. ICASSP*, Calgary, AB, Canada, pp. 1723–1727, 2018.

[19] 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*, Montreal, Quebec, Canada, 28, pp. 1–14, 2015.

[20] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. ICIP*, Beijing, China, pp. 3645–3649, 2017.

[21] W. Fraser and C. C. Gotlieb, "A calculation of the number of lattice points in the circle and sphere," *Mathematics of Computation*, vol. 16, no. 79, pp. 282–290, 1962.

[22] G. Huang, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4700–4708, 2017.

[23] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. Pattern Recognition*, Cambridge, UK, 3, pp. 32–36, 2004.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] W. C. Lo, C. L. Fan, J. Lee, C. Y. Huang, K. T. Chen *et al.,* "360 video viewing dataset in head-mounted virtual reality," in *Proc. Multimedia Systems*, Taipei, Taiwan, pp. 211–216, 2017.

[26] X. Chen, A. T. Z. Kasgari and W. Saad, "Deep learning for content-based personalized viewport prediction of 360-degree VR videos," *IEEE Networking Letters*, vol. 2, no. 2, pp. 81–84, 2020.

[27] L. Abualigah, D. Yousri, M. A. Elaziz, A. A. Ewees, M. AA Al-qaness *et al.,* "Aquila optimizer: A novel meta-heuristic optimization algorithm," *Computers & Industrial Engineering*, vol. 157, pp. 1–37, 2021.

[28] F. A. Hashim, K. Hussain, E. H. Houssein, M. S. Mabrouk and W. Al Atabany, "Archimedes optimization algorithm: A new metaheuristic algorithm for solving optimization problems," *Applied Intelligence*, vol. 51, no. 3, pp. 1531–1551, 2021.

[29] H. N. Ghafil and K. Jarmai, "Dynamic differential annealed optimization: New metaheuristic optimization algorithm for engineering applications," *Applied Soft Computing*, vol. 93, no. 4598, pp. 1–33, 2020.

[30] S. Harifi, J. Mohammadzadeh, M. Khalilian and S. Ebrahimnejad, "Giza pyramids construction: An ancient-inspired metaheuristic algorithm for optimization," *Evolutionary Intelligence*, vol. 14, no. 4, pp. 1743–1761, 2021.

[31] Q. Askari, M. Saeed and I. Younas, "Heap-based optimizer inspired by corporate rank hierarchy for global optimization," *Expert Systems with Applications*, vol. 161, no. 3, pp. 1–41, 2020.

[32] Y. Che and D. He, "A Hybrid whale optimization with seagull algorithm for global optimization problems," *Mathematical Problems in Engineering*, vol. 2021, no. 4, pp. 1–31, 2021.

[33] P. Perez, M. Gangnet and A. Blake, "Poisson image editing," *Proc. SIGGRAPH*, vol. 22, no. 3, pp. 313–318, 2003.