International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Capsule Neural Network and Determinantal Point Process (CAPSDPP) based Summarization of Surveillance Videos

Tabiya Manzoor Beigh*[1], V. Prasanna Venkatesan[2], C. Punitha Devi[3]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: Seamless deployment and the low cost of surveillance cameras have benefited various agencies, such as schools, colleges, airports, railway stations, and shopping malls. However, the data generated by these cameras is enormous. Accessing a specific clip requires users to invest time and energy in watching the entire video. Video summarization aims to produce a brief and comprehensive depiction of the essential content of video. The information can be presented using keyframes or video summaries, avoiding redundancy and emphasizing important and varied segments. Constraints such as finite computational capacity and bandwidth restriction limit the availability of resources at the edge. This work proposes a lightweight model based on the Capsule neural network (CapsNet) to summarize surveillance videos. Capsule neural networks are employed to extract spatiotemporal features that capture both motion and visual information. Deep CapsNet features are utilized for shot segmentation. A determinantal point process (DPP) selects diverse keyframes within segmented shots. We assessed the effectiveness of the proposed method with benchmark datasets from the Open Video Project (OVP) and YouTube (YT). Our findings illustrate that the proposed approach surpasses the performance of existing methodologies.

Keywords: *Capsule Neural Network; Determinantal point process; Keyframes; Redundancy; Segmentation; Summarization; Shot; Video surveillanc*

1. Introduction

The rapid development of communication and computational media is generating a significant amount of multimedia data. Video data is generated from various sources, including wireless sensor networks [1] healthcare [2], social media [3], and sports [4]. The data generated from the wireless sensors is transferred to servers promptly. The data can have a lifespan of a week, month, or year, or even be stored indefinitely. Managing this accumulated data is a challenging task that requires large storage systems, making indexing and retrieval difficult. The vast amount of information obtained from these sources can be utilized for various purposes, such as security, smart analysis, and intelligent transportation. However, the challenge lies in the need for storage, computing, and transmitting devices to handle such a massive amount of data. Rather than storing the raw information as it is, it is necessary to process and select subsets of it. Video summarization is a solution that researchers are increasingly focusing on. Video summarization provides a concise summary of the video by highlighting important aspects. Video summarization can be either static or dynamic. Static video summarization provides a summary in keyframes,

which are representative images of the video content. Keyframe selection involves identifying a designated subset of frames from the video set while maintaining the information, content coverage, and diversity of the video. In contrast, dynamic summarization of video provides a short video skim that conveys the information content both statically (through still images) and dynamically (through sound, motion, etc.).

Surveillance cameras are installed in both the public and private domains, capturing videos throughout the day. There is no specific time for video capture or when capturing should stop. However, most of the content in CCTV footage is redundant and of no use to the end user. Users are primarily interested in sections of the video where important events have occurred. In the case of surveillance videos, the videos are either transferred to servers or analysed by humans. However, efficiency and cognitive abilities limit human analysis because humans cannot monitor footage 24/7 and report events perfectly. Therefore, there is a need for a method that can critically analyse and summarize the representative content of the video. Different summarization techniques are used for videos in different genres. For example, in sports videos, the final summary typically includes important events known as highlights, such as wicket falls, sixes, or four-run hits in cricket, or goals and penalties in soccer. In movies, the presence of actors in certain scenes is of utmost importance to viewers. Considering the challenges

[1]*Research Scholar, Department of Computer Science, Pondicherry University, 605014, India*
[1] *ORCID ID: 0000-0001-6358-8161*
[2] *Professor, Department of Banking Technology, Pondicherry University, 605014, India*
[2] *ORCID ID 0000-0002-1444-0918*
[3] *Associate Professor, Department of Banking Technology, Pondicherry University, 605014, India*
ORCID ID: 0000-0002-7917-4526
Corresponding Author email id: taha.beigh@gmail.com

associated with surveillance video summarization, we propose a computationally lightweight technique to address surveillance video summarization. Segmentation of the video shots is done using a Capsule Neural Network, where the CapsNet acts as a feature extractor and segments the shots based on a thresholding scheme. To select keyframes from each video shot, we use determinantal point processes. Our system's capability to recognise important frames in surveillance streams signifies its efficiency, making it an optimal fit for integration into visual surveillance networks.

The primary achievements of this study are:

- A novel framework is created to amalgamate spatiotemporal features produced by Capsule networks for shot segmentation.
- To summarize the video into keyframes, a probabilistic model DPP is applied for diverse subset selection.

The paper is systematically arranged as follows: the upcoming section reviews antecedent studies on surveillance video summarization. Section 3 demonstrates the proposed Capsule Neural Network and Determinantal Point Process (CAPSDPP) based summarization of surveillance videos. Section 4 discusses the results and compares the obtained results with existing techniques. In the final section, the paper is concluded.

2. Related Work

Anshy et al. [5] suggested a method that works on the extraction of a representative frame using Bayesian fuzzy clustering and refined the clustering results using deep CNN. While the technique produced statistically significant results, a major drawback is the lack of testing for real-time applications. In [6], Yasmin et al. proposed a moment-based candidate keyframe selection technique. They applied agglomerative clustering to obtain a set of diverse and informative keyframes, using the motion feature in the video. The constraint associated is the computational cost of extracting the motion vector field. Sadiq et al.[7] proposed an efficient method of keyframe extraction from the CCTV streams. They employed k-means clustering to create different clusters of diverse frames. The difference between cluster centers defined eligibility criteria for a frame to be a keyframe or not. However, the procedure does not address videos of long duration. A keyframe selection method proposed by Pandian et al. [8] adopts a multi-clustering approach to identify the main activities in surveillance videos. Similar activities within the frame were clustered using a Markov chain-based approach. The inter-relationship between activities in the frames is further refined by using adjacency matrix-based clustering. A dual-stream CNN model is utilized to extract multiple attributes [20]. These attributes are clubbed at various levels to generate a more discriminative feature representation. The selection of candidate keyframes is accomplished through the utilization of a neighbourhood peak detection algorithm. Daniel et al. [9] used curve simplification as a summarization technique. In a high-dimensional feature space, video sequences manifest as trajectory curves, enabling intricate pattern recognition by representing video sequences as a trajectory curve. A neural network-based summarization is suggested by Mohammad et al. [10] for the selection of keyframes in surveillance environments. The keyframe selection is based on the memorability and entropy score. The method is designed for resource-constrained devices. A video summarization model for intelligent transportation has been suggested by Balamurugan et al.[11]. The model highlights abnormal events by using the additive event summarization method. A variant of the recurrent neural network is used for the frame verification of classified events. Ambreen et al.[12] used the transfer learning concept to summarize the crowded scenes during the COVID-19 period. The model generates short summaries comprising masked and non-masked persons. The generated summaries helped to detect the violations of mask regulations.

Based on an analysis of the contemporary literature, it is evident that specific approaches can produce quality summaries; however, their excessive computational requirements hinder their practicality for surveillance networks and resource-limited devices. Certain effective methods lack sufficient capability to identify important frames from the live surveillance feed accurately.

3. Proposed Methodology

The proposed work tackles the challenge of video summarization by providing a concise set of keyframes. Keyframe selection is done in such a way that the information in the video is circulated through the keyframes, eliminating redundancy. The sequence of the suggested methodology is given in Fig. 1. It has been observed that for a better and more precise video summary, shot segmentation should be done properly [10]. The first and most important step in our proposed methodology is to perform shot segmentation. To achieve this task, the input comprises video content. A capsule neural network is used to find the feature vectors. Shot segmentation is done using Capsule Neural Network. Capsule Neural networks extract the spatiotemporal features. These feature vectors encompass visual as well as temporal information. The distance between the feature vectors is calculated. An already set threshold is used to segment feature vectors in corresponding shots. Each segmented shot contains a set of frames about a particular shot. The DPPs are used for the selection of keyframes from each shot.

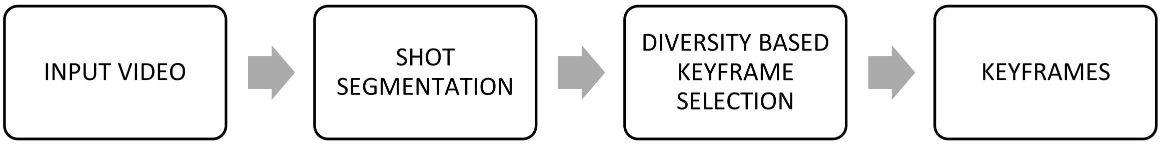


Fig. 1: Workflow of the proposed methodology

3.1 Shot Segmentation Using Capsule Neural Network

An input in the form of videos is given to the Capsule neural network. Capsule neural networks are advanced neural networks that mimic the functioning of different components of the brain to achieve a specific task. Modules and components of the brain are termed capsules in the architecture [13]. The architecture of the Capsule Neural Network used in this work is given in Fig. 2. Capsules are a set of neurons clubbed together to perform computations on the input and encapsulate output information in vector form. An object's existence likelihood is determined by the length of the resultant vector. The orientation of the output vectors provides information about the abstract attributes, including texture, size, color, and position.

Capsule neural network follows the routing agreement mechanism to attain spatial as well as temporal information about the input vectors. In the first step, input vectors are multiplied by the weight matrices to achieve the affine transformation as given in eq.(1).

$$\widehat{U}_{j|i} = W_{ij} U_i \qquad (1)$$

where $U_{j|i}$ is the output vector at level i, W_{ij} is the weight matrix or transformation matrix. U_i is the previous layer input vector. The spatial interactions of features at various levels are encoded by this operation. The weighted sum is obtained by finding the coupling coefficients using the routing agreement principle as given in eq. (2).

$$S_j = \sum_i c_{ij} u_{j|i} \qquad (2)$$

where S_j is the weighted sum of input vectors, C_{ij} are the coupling coefficients and $U_{j|i}$ is the vector at the lower-level capsule. The maximum amount of data from the lower-level capsules is received by higher-level capsules using dot product operation and expectation maximization. Between the weight matrix and the prediction vector calculated by the lower capsule layers, this dot product is created. The capsule holding the highest dot product capsule is designated the parent capsule. The weighted sum of input vectors is passed through a non-linear activation function known as the squash function. As specified in eq. (3), the output vector is normalized between 0 and 1 in terms of length.

$$V_j = \frac{||s_j||}{1+||s_j||} \frac{s_j}{||s_j||} \qquad (3)$$

where V_j is the final output vector. The output generated by the Capsule Neural Network comprises the spatiotemporal features and interframe motion curve. The distinguishing property of spatiotemporal features is that they can simultaneously represent visual as well as motion attributes. In this approach, these extracted spatiotemporal deep features are used for the intelligent segmentation of shots. From the DigitCaps layer of the capsule neural network, 16-dimensional feature vectors are extracted. To find the distance between the deep features, whose formula is given in eq. (4), is done using Euclidean distance.

$$D = \sum_{i=1}^{n} (\bar{Y_i} - Y_i)^2 \qquad (4)$$

After extensive experimentation, the optimal threshold value for distinguishing between identical and distinct shots was determined to be 0.8. Two frames are said to be in the same shot if the Euclidean distance between them is less than or equal to 0.8. If the Euclidean distance exceeds the threshold value of 0.8, it is implied that frames belong to the different shots.

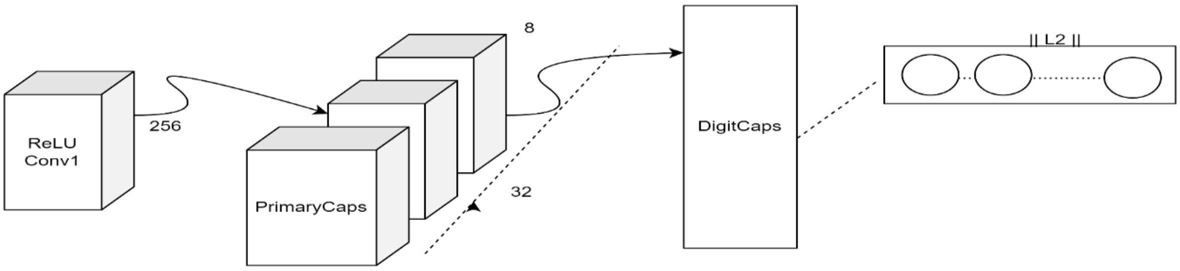


Fig. 2 Architecture of Capsule Neural Network

3.2 Diversity-based Keyframe Extraction using the Determinantal Point Process

Determinantal point processes, DPP, also known as fermion processes were first used to model the Pauli exclusion principle. It specifies that two electrons or particles cannot concurrently occupy identical quantum states.[14]. A DPP specifies a distribution across all the subsets of a ground set that quantifies the element's repulsion or negative correlation with one another. The property of repulsion makes determinantal point processes perfect to model diversity. There are many variants of DPP which include, Sequential DPP [15], continuous DPP[16], and structured DPPs[17]. In this work, Sequential determinantal point processes are used to extract diverse and representative frames from all segmented shots. The drawback of other variants is that they do not consider the sequential order of frames. If similar events happen at two different timestamps, they will report a single event and select keyframes relevant to a single event. In this way, loss of information occurs and the keyframes will convey incomplete content. Sequential Determinantal Point Processes (SeqDPP) can exploit sequential progression inherent in video frames. In the SeqDPP framework, the initial step involves segmenting the video into m disjointed consecutive short sections, ensuring that the union of these segments covers the entire video.

A ground set T comprises all frames N in each shot. The affinity among frames is measured by an N * N positive, semi-definite kernel matrix. The kernel matrix L or the marginal matrix kernel contains all the information for any subset to be $\widetilde{T}$ being included in the power set T.

In this work, SeqDPP encodes the odds of selecting a subset of keyframes from the frame list. The probability of a subset $\tilde{T}$ is detailed in eq. (5).

$$P = \left(\tilde{T} = T; L\right) = \frac{\det\ (L_{\tilde{T}}\ )}{\det(\ L+I)} \qquad (5)$$

where $L_{\bar{T}}$ is the corresponding principal of the sub-matrix corresponding to the subset $\widetilde{T}$ with selected rows and columns in accordance with the indices in the subset $\tilde{T}$, det is the determinant function in matrices which allows the DPPs to follow the policy of repulsion. Consider the subset has two frames fr_i and fr from the same shot as in eq. (6), then we have:

$$P\ (\ T - \{\ i, j\}) \propto L_{ii}\, L_{jj}\ -\ {L_{ij}}^2 \quad (6)$$

If both the frames are similar, then the probability $P\ (\ T - \{\ i, j\}) = 0$, resulting in only one frame in either frame fr_i or fr_j because determinant property in DPP does not allow affinity, resulting in a diverse set of elements. In this case, the probability of having two similar frames in the same set is very low. The final diverse set will be the one having the highest probability.

$$T^* \ =\ argmax_T\ P\big(\ T\ =\ \tilde{T}\ \big) = argmax_{\tilde{T}} \det L_{\tilde{T}} \quad (7)$$

Within the model's context, θ serves as the parameterization for L, covering all relevant parameters. In the case of surveillance video summarization, parameters θ is learned and optimized using maximum likelihood estimation [18] as given in eq. (8):

$$\theta =\ argmax_\theta\ \textstyle\sum_i \log\{\ P\tilde{T}^{(i)*}\ \ \subset T^{(i)};\ L^{(i)}\ (\theta))\} \quad (8)$$

The pseudocode for training Sequential DPP is given below.

```
Pseudocode for training Sequential DPP
Training Sequential DPPs for Keyframe Selection:
Input:   A set of feature vectors representing frames of a video.
           $X= \{x_1, x_2, x_3,\ldots\ldots x_n\}$
           K : Kenel matrix representing pairwise similarities between frames.
Initialize an empty set to store selected keyframes
           SelectedKF = {}
Iterative training
 For each iteration t= 1, 2, 3,…….n:
   Compute the conditional kernel matrix $K_{condition}$ based on previously selected keyframes SelectedKF and the remaining frames.
    Compute the probability vector $P_t$ for selecting the next keyframe using the determinant of the conditional kernel matrix $K_{condition.}$
           $\boldsymbol{P_t(i)\ \propto\ \det K_{condition}}$
Sample the next frame $x_t$ using the probability vector $p_t$
Add the selected keyframe $x_t$ to the SelectedKF.
Output: Selected keyframes SelectedKF.
```

3.3 The Entire Process of Algorithm Execution

Our central aim in this paper is to identify key frames to summarize videos. The functionality primarily arises from the automatic partitioning of the video into distinct shots. After the shot segmentation is concluded and the sequence reaches the transitional area, the input to the algorithm i.e, all the frames from the prior shot are given to DPP. This procedure occurs amidst unnecessary transition effects, thus enhancing real-time performance. It is worthy of further elaboration that the threshold for selecting keyframes is decided by the intermediate value situated between the peak points and the basal levels within each diverse attention curve of DPP.

4. Experimental results

4.1 Datasets

The suggested approach is tested using two benchmark datasets sourced from the OVP database[19] and the YT database. The OVP dataset comprises different video genres including documentary, educational, ephemeral, historical, and lecture. 50 videos are taken from the OVP as they have ground truth summaries. The video may last one or two minutes, with a maximum allowable length of four minutes. The duration of the videos approximates to 75 minutes. The videos carry significant sound data, with frames formatted in RGB. The video dimensions are of 352 × 240 pixels. From the YouTube database, 50 videos belonging to different genres were downloaded. The duration of the videos of the YouTube dataset (YT) varied from 1 to 10 minutes. It includes videos from genres like cartoons, Sports, and news. Each dataset includes five user-generated summaries for every video, with keyframes chosen by the users. The results obtained from both datasets were compared with the existing methods and some user-generated summaries.

4.2 Evaluation metrics

Precision: The ratio of correctly generated keyframes in the summary to the total number of frames. as shown in eq. (9). The accuracy of the technique is determined by using the extracted false keyframes.

$$\text{Precision} = \frac{N_{relevant}}{N_{extracted}} \qquad (9)$$

$N_{relevant}$ signifies the number of keyframes similar in the user-generated summary as well as by the suggested technique.

Recall: It correlates with the likelihood of extraction from every ground truth keyframe as shown in eq. (10)

$$\text{Recall} = \frac{N_{relevant}}{N_{GT}} \qquad (10)$$

N_{GT} articulates the keyframe count in the ground truth summary.

Using these metrics (Precision and Recall), the F-measure is calculated as given in eq. 11.

$$\text{F-measure} = \frac{2*Precision*Recall}{Precision+Recall} \qquad (11)$$

4.3 Results

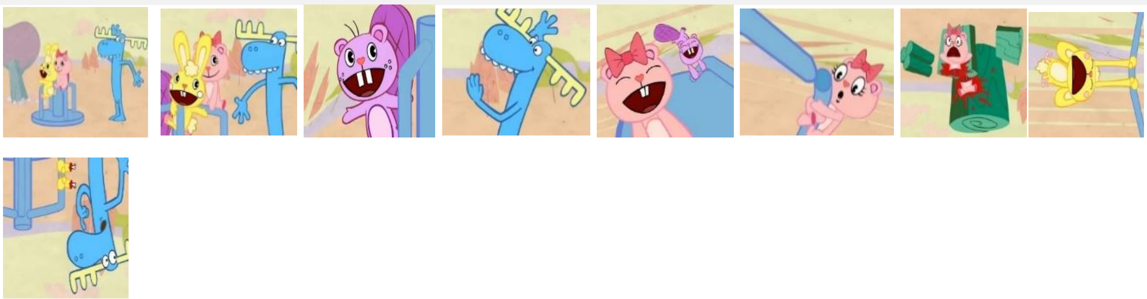
Table 1 provides an overview of sample videos, presenting their F-measure scores for comparison. Table 1 highlights that the suggested work produces better F-measure scores as compared to the existing techniques.

Table 1. F1-measure comparison of the current and existing techniques.

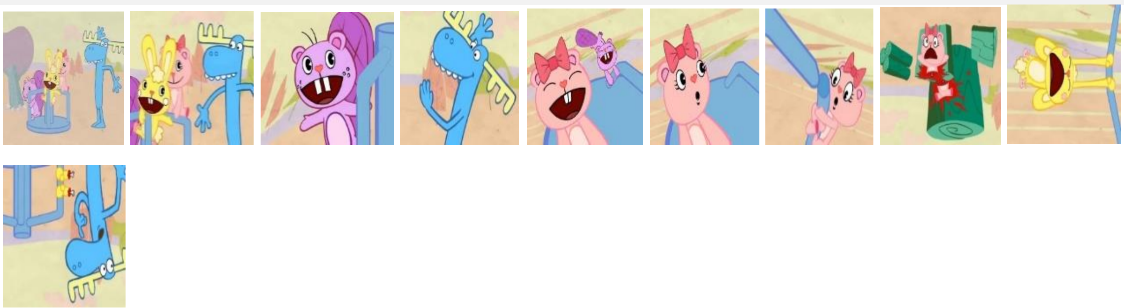
Videos	STIMO [20]	OV [9]	VSUMM[21]	Tanveer et al. [10]	Proposed method
Video 1	60%	66%	58%	78%	80%
Video 5	54%	46%	73%	72%	76%
Video 50	89%	77%	71%	80%	82%
Average F1-score	67%	63%	67.3%	76%	79.3%

The relative performance of the suggested work is visualized in Fig 3. A summary comparison is conducted by the suggested method, existing techniques, and the ground truth summary generated by five different users. The summary generated from the proposed system is almost similar to the one generated by the humans. It is quite visible from the summary that the keyframes selected are informative as well as diverse. For comparative analysis, a video sample is chosen from diverse genres, and the results are compared with alternative methods and summaries generated by users as shown in Table 2. The comparison with other methods shows that the resultant summary is very similar to the one generated by humans

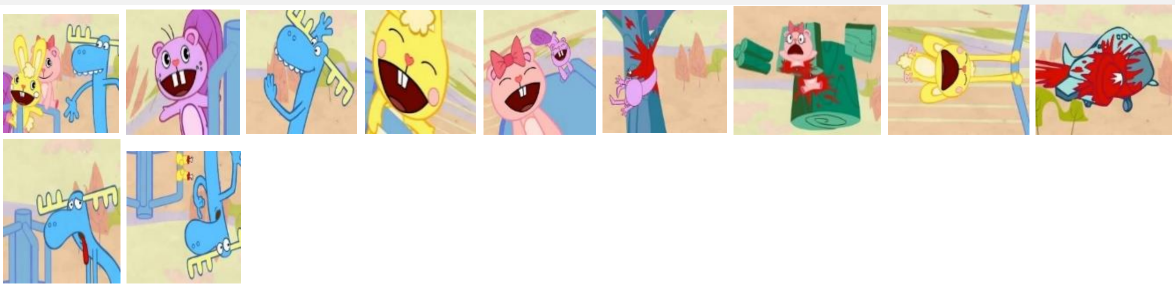
Proposed method



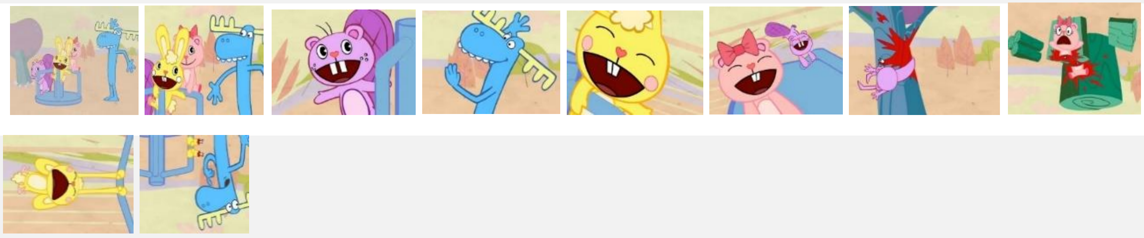
Tanveer et al. [10]



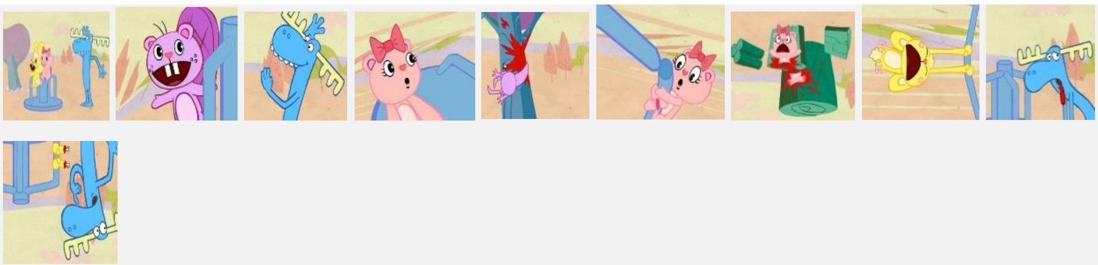
Fei et al [22]



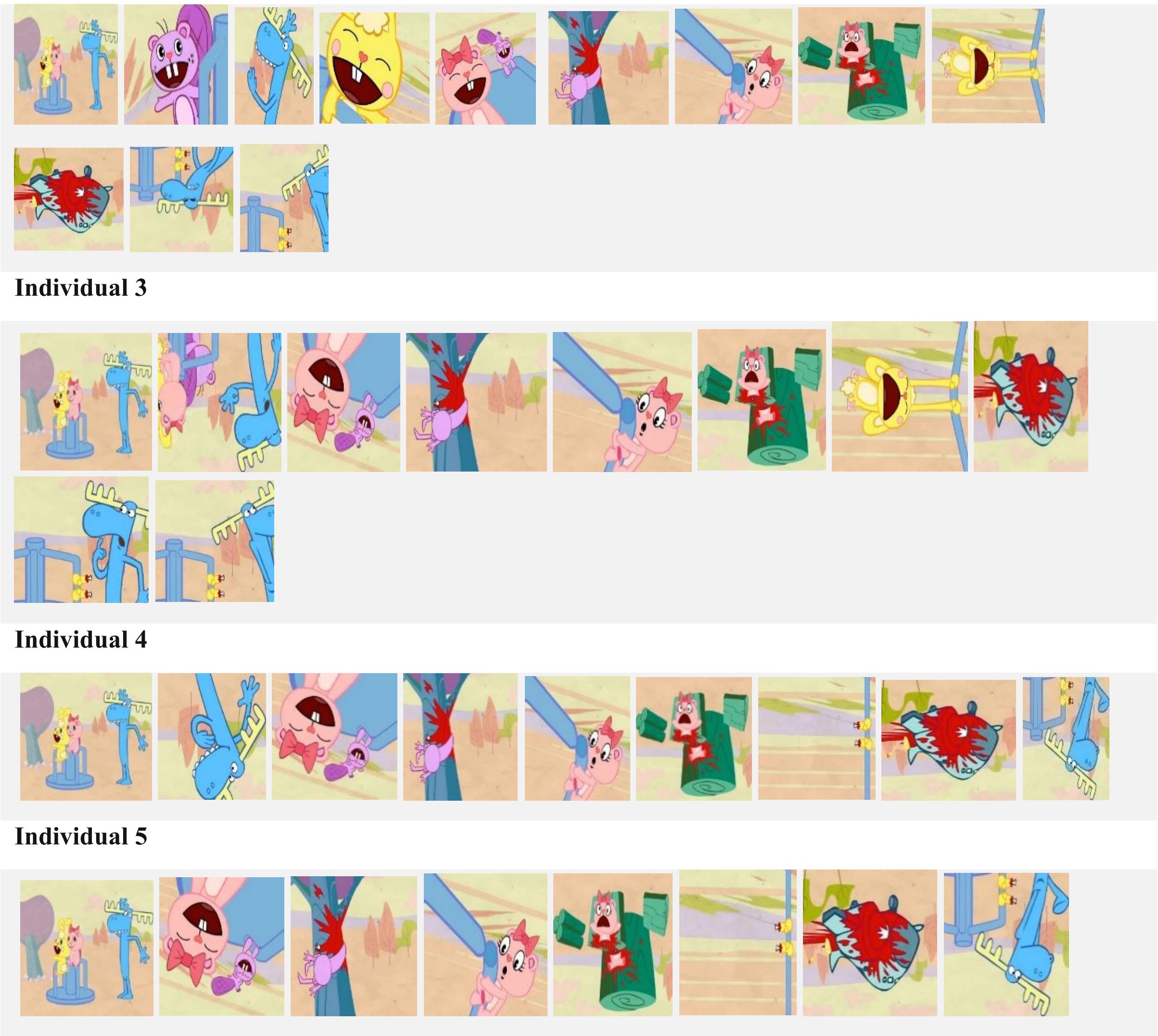
VSUMM [21]



Individual 1



Individual 2



Individual 3

Individual 4

Individual 5

Fig. 3 Comparison of summaries produced by the proposed method, existing techniques, and humans.

Table 2. Average F-measure score generated by the proposed method, existing methods, and user-generated summaries.

Genre	Video No.	Lee et al. [23]	Human Generated Summary	Tanveer et al.[10]	Proposed Method
Cartoon	12	65%	71%	73%	75%
News	88	67%	76%	78%	80%
Home	108	52%	67%	73%	77%
Average F-measure score		61%	71%	74%	77%

5. Conclusion

Convolutional neural networks have produced significant results in image and video processing domains. The inherent property of the CNNs to downsample the spatial information is the main reason for its inability to capture the spatial hierarchies. CapsNets can capture spatial as well as temporal features. In this paper, CapsNets have been used in the summarization of surveillance videos. CapsNet is used to extract the features from the videos. Shot segmentation is done with the help of CapsNet deep features. These features help in the intelligent segmentation of surveillance videos into insightful shots. Determinantal point process is used for the selection of keyframes from segmented shots. Similar frames within the shot will be discarded, eliminating redundancy. The final set of keyframes constitutes the final summary. Assessment of the proposed method includes two benchmark datasets from the OVP database and the YT database. The results reveal that the suggested technique is performing better than existing methodologies. The robustness of the proposed technique in a real-time environment is still a challenge. Furthermore, the proposed technique can be extended and applied in other resource-constrained domains such as the Internet of Things.

Conflict of interest We affirm that no conflicts of interest are present.

References

[1] K. Budati, S. Islam, M. K. Hasan, N. Safie, N. Bahar, and T. M. Ghazal, “Optimized Visual Internet of Things for Video Streaming Enhancement in 5G Sensor Network Devices,” Sensors, vol. 23, no. 11, p. 5072, May 2023, doi: 10.3390/s23115072.

[2] R. Arunachalam, G. Sunitha, S. K. Shukla, S. N. pandey, S. Urooj, and S. Rawat, “A smart Alzheimer’s patient monitoring system with IoT-assisted technology through enhanced deep learning approach,” Knowl Inf Syst, vol. 65, no. 12, pp. 5561–5599, Dec. 2023, doi: 10.1007/s10115-023-01890-x.

[3] D. J. Cassidy et al., “#SurgEdVidz: Using Social Media to Create a Supplemental Video-Based Surgery Didactic Curriculum,” Journal of Surgical Research, vol. 256, pp. 680–686, Dec. 2020, doi: 10.1016/j.jss.2020.04.004.

[4] D. M. Davids, A. A. E. Raj, and C. S. Christopher, “Hybrid multi scale hard switch YOLOv4 network for cricket video summarization,” Wireless Networks, vol. 30, no. 1, pp. 17–35, Jan. 2024, doi: 10.1007/s11276-023-03449-8.

[5] Singh and M. Kumar, “Bayesian fuzzy clustering and deep CNN-based automatic video summarization,” Multimed Tools Appl, vol. 83, no. 1, pp. 963–1000, Jan. 2024, doi: 10.1007/s11042-023-15431-9.

[6] G. Yasmin, S. Chowdhury, J. Nayak, P. Das, and A. K. Das, “Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework,” Neural Comput Appl, vol. 35, no. 7, pp. 4881–4902, Mar. 2023, doi: 10.1007/s00521-021-06132-1.

[7] Muhammad, B. Sadiq, I. Umoh, and H. Bello Salau, “A K-Means Clustering Approach for Extraction of Keyframes in Fast- Moving Videos,” pp. 147–157, Jul. 2020.

[8] A. Pandian and S. Maheswari, “A keyframe selection for summarization of informative activities using clustering in surveillance videos,” Multimed Tools Appl, vol. 83, no. 3, pp. 7021–7034, Jan. 2024, doi: 10.1007/s11042-023-15859-z.

[9] DeMenthon, V. Kobla, and D. Doermann, “Video summarization by curve simplification,” in Proceedings of the sixth ACM international conference on Multimedia - MULTIMEDIA ’98, New York, New York, USA: ACM Press, 1998, pp. 211–218. doi: 10.1145/290747.290773.

[10] K. Muhammad, T. Hussain, and S. W. Baik, “Efficient CNN based summarization of surveillance videos for resource-constrained devices,” Pattern Recognit Lett, vol. 130, pp. 370–375, Feb. 2020, doi: 10.1016/j.patrec.2018.08.003.

[11] G. Balamurugan and J. Jayabharathy, “Abnormal Event Detection using Additive Summarization Model for Intelligent Transportation Systems.” [Online]. Available: www.ijacsa.thesai.org

[12] Sabha and A. Selwal, “Data-driven enabled approaches for criteria-based video summarization: a comprehensive survey, taxonomy, and future directions,” Multimed Tools Appl, vol. 82, no. 21, pp. 32635–32709, Sep. 2023, doi: 10.1007/s11042-023-14925-w.

[13] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” in 31st Conference on Neural Information Processing Systems (NIPS 2017, Long Beach, CA, USA, 2017.

[14] W. Pauli, “The Connection Between Spin and Statistics,” Physical Review, vol. 58, no. 8, pp. 716–722, Oct. 1940, doi: 10.1103/PhysRev.58.716.

[15] B. Gong, W.-L. Chao, K. Grauman, and S. Fei, “Diverse Sequential Subset Selection for Supervised Video Summarization,” in Advances in neural information processing systems , 2014.

[16] R. H. Affandi, E. B. Fox, R. P. Adams, and B. Taskar, “Learning the Parameters of Determinantal Point Process Kernels.”

[17] Kulesza and B. Taskar, “Structured Determinantal Point Processes,” in Neural Information Processing Systems, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:13192203

[18] Kulesza, “Determinantal Point Processes for Machine Learning,” Foundations and Trends® in Machine Learning, vol. 5, no. 2–3, pp. 123–286, 2012, doi: 10.1561/2200000044.

[19] G. Geisler and G. Marchionini, “The open video project,” in The open video project. Proceedings of the Fifth ACM Conference on Digital Libraries, Association for Computing Machinery (ACM), Jun. 2000, pp. 258–259. doi: 10.1145/336597.336693.

[20] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, “STIMO: STIll and MOving video storyboard for the web scenario,” Multimed Tools Appl, vol. 46, no. 1, pp. 47–69, Jan. 2010, doi: 10.1007/s11042-009-0307-7.

[21] S. E. F. De Avila, A. P. B. Lopes, A. Da Luz, and A. De Albuquerque Araújo, “VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method,” Pattern Recognit Lett,

vol. 32, no. 1, pp. 56–68, 2011, doi: 10.1016/j.patrec.2010.08.004.

[22] M. Fei, W. Jiang, and W. Mao, “Memorable and rich video summarization,” J Vis Commun Image Represent, vol. 42, pp. 207–217, Jan. 2017, doi: 10.1016/j.jvcir.2016.12.001.

[23] Yu-Chyeh Wu, Yue-Shi Lee, and Chia-Hui Chang, “VSUM: Summarizing from Videos,” in IEEE Sixth International Symposium on Multimedia Software Engineering, IEEE, pp. 302–309. doi: 10.1109/MMSE.2004.90.