

# Leveraging 3D convolutional networks for effective video feature extraction in video summarization

Bhakti Deepak Kadam<sup>1,2</sup>, Ashwini Mangesh Deshpande<sup>1</sup>

<sup>1</sup>Department of Electronics and Telecommunication Engineering, MKSSS's Cummins College of Engineering for Women, Pune, India

<sup>2</sup>Department of Electronics and Telecommunication Engineering, SCTR's Pune Institute of Computer Technology, Pune, India

---

## Article Info

### Article history:

Received Apr 9, 2024

Revised Sep 13, 2024

Accepted Sep 30, 2024

---

### Keywords:

3D convolution

Deep neural networks

Feature representation

Pretrained networks

Video summarization

---

## ABSTRACT

Video feature extraction is pivotal in video processing, as it encompasses the extraction of pertinent information from video data. This process enables a more streamlined representation, analysis, and comprehension of video content. Given its advantages, feature extraction has become a crucial step in numerous video understanding tasks. This study investigates the generation of video representations utilizing three-dimensional (3D) convolutional neural networks (CNNs) for the task of video summarization. The feature vectors are extracted from the video sequences using pretrained two-dimensional (2D) networks such as GoogleNet and ResNet, along with 3D networks like 3D Convolutional Network (C3D) and Two-Stream Inflated 3D Convolutional Network (I3D). To assess the effectiveness of video representations, F1-scores are computed with the generated 2D and 3D video representations for chosen generic and query-focused video summarization techniques. The experimental results show that using feature vectors from 3D networks improves F1-scores, highlighting the effectiveness of 3D networks in video representation. It is demonstrated that 3D networks, unlike 2D ones, incorporate the time dimension to capture spatiotemporal features, providing better temporal processing and offering comprehensive video representation.

*This is an open access article under the [CC BY-SA](#) license.*



---

## Corresponding Author:

Bhakti Deepak Kadam

Department of Electronics and Telecommunication Engineering

MKSSS's Cummins College of Engineering for Women

Pune, Maharashtra, India

Email: bhakti.kadam@cumminscollege.in

---

## 1. INTRODUCTION

Video feature extraction is a foundational aspect of computer vision, designed to overcome the challenges presented by the intricate and extensive nature of video data. Its purpose is to facilitate analyses that are not only more efficient but also more interpretable and accurate. Effective comprehension of videos demands representation at multiple levels, necessitating an appropriate video representation. Effective video processing relies on video feature extraction for the following reasons: (i) reducing the complexity of video data for more manageable analysis, (ii) simplifying the interpretation of underlying information for both humans and deep learning models, (iii) condensing meaningful information while retaining essential characteristics, (iv) enhancing the model's ability to generalize patterns for accurate predictions, and (v) efficient utilization of computational resources by focusing on relevant aspects.

Video feature extraction involves selecting and/or combining variables to generate feature vectors. Feature vectors of a video are numerical representations that capture various attributes and characteristics of the video content in a structured format. These vectors can encapsulate various visual, spatial, temporal, motion, and audio attributes of the video content. These vectors facilitate effective analysis, understanding, and processing of videos in numerous applications. Feature extraction efficiently reduces the volume of data that needs processing while maintaining accuracy in representing videos.

The combination of video segmentation and feature extraction is instrumental in mitigating computational overhead by streamlining the preprocessing across all frames in the video. When analyzing videos for computer vision tasks, a variety of features play a pivotal role. The different video features utilized for video understanding are illustrated in Figure 1. These diverse features contribute to a holistic comprehension of video content, facilitating various applications within the field of computer vision [1].

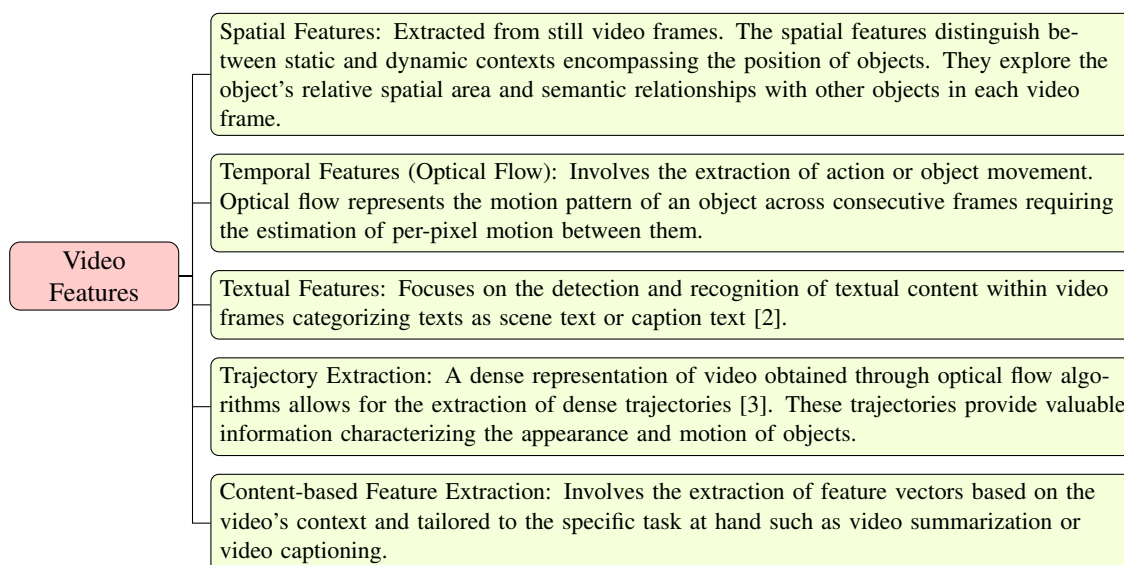


Figure 1. Types of video features

Features can be extracted using both classical methods that rely on local, hand-crafted features and advanced techniques involving deep neural networks, as detailed in section 2. This research explores the utilization of three-dimensional convolutional neural networks (3D CNNs) to enhance video representations. A comparative analysis is conducted on conventional video summarization and query-focused video summarization techniques. Our contributions are as follows: i) the video features are extracted utilizing pretrained 3D CNNs (C3D and inflated 3D (I3D)); ii) the specific baseline algorithms for generic and query-focused video summarization are chosen and the F1-scores for generated video presentations are computed; and iii) the performance is evaluated by comparing pretrained two-dimensional (2D) and 3D convolutional networks for feature extraction in terms of calculated F1-scores.

Given its numerous advantages, feature extraction stands as a fundamental process in numerous research applications, such as:

- Video classification: it is the task of assigning one or more global labels to the video. The proper extraction of features from the input video leads to the prediction of accurate frame labels that describe the entire video [4].
- Action recognition: the action recognition in videos aims to infer the actions of one or persons in the video. The spatial and long-range temporal feature extraction is necessary for human activity or action recognition [5].
- Video understanding: is the task of recognition and localization of different actions or events occurring in the video. As the localization is in both spatial and temporal dimensions, this task requires spatiotemporal feature extraction [6].

- Video captioning: is the task of generating automatic captions for a video. This leads to efficient information retrieval from the video in the form of text. As captioning is the textual description of the video, it needs extraction of more complex features [7].
- Simultaneous localization and mapping (SLAM): is a method used for autonomous vehicles that develops a map and localizes the vehicle in the same map [8]. In SLAM, spatial and motion features need to be extracted and matched for localization and obstacle detection.
- Video summarization: is a process of generating a temporally condensed version of the input video. Video representations at multiple levels are necessary for spatiotemporal modelling due to long durations of videos [9]–[11].

These applications make video feature extraction a valuable research topic for study. The structure of the paper is as follows: section 2 discusses the related work on video feature extraction. Section 3 elaborates on the use of 3D convolutional networks employed for extracting video features. The experimental result and analysis are discussed in section 4 and section 5 provides conclusions.

## 2. RELATED WORK

This section explores the existing video feature extraction techniques employed in summarization methodologies in literature. Video summarization and feature extraction represent longstanding research areas in computer vision. Video feature vectors can be extracted using classical vision techniques focusing on hand-crafted features as well as deep neural networks [1]. The classification of feature extraction techniques is provided in Figure 2.

With advancements in deep learning, video feature extraction has also leveraged these technologies. Key trends propelling the field forward include the integration of multimodal information, the development of self-supervised learning techniques, and the exploration of novel architectures such as transformers. The deep learning based feature extraction techniques have outperformed the classical vision techniques. These models are effectively utilized in various research domains [1]. The merits of deep learning based techniques include:

- Extraction of complex and abstract features by feature engineering: feature engineering deals with the extraction of features from natural data. The spatiotemporal models utilize state-of-the-art feature engineering models to extract more complex features from videos.
- Feature extraction for unstructured data: deep neural networks can handle unstructured data better than hand-crafted features by training on various abstract features.
- Unsupervised feature learning: the process of labelling the available data is expensive and time-consuming. This process is more challenging when it is extended for videos. The traditional techniques do not perform well on unsupervised data, but spatiotemporal models can be efficiently used with unlabelled data.
- High-quality results: the semantic relationships between objects and their motion patterns are also explored while extracting the features using modern machine vision techniques. This leads to improvement in the quality of results in different computer vision tasks.

Most of the summarization methods employ 2D CNNs, GoogleNet, and Residual Network (ResNet) to extract video features. GoogleNet, also known as Inception V1, was presented in 2014 [12]. ResNet, introduced in 2015, brought forth the ResNet architecture [13]. In video summarization frameworks, GoogleNet and ResNet pretrained on the ImageNet dataset [14] are widely employed for feature extraction from input video sequences. 2D CNNs face several challenges in video feature extraction due to their limitations in handling temporal information:

- Lack of temporal awareness: 2D CNNs process each frame independently, missing temporal relationships crucial for understanding motion and events.
- Handling motion: they struggle with dynamic content and the complexity of integrating optical flow.
- Spatiotemporal features: they capture only spatial features, lacking the rich spatiotemporal context needed for tasks like action recognition.
- Multi-modal integration: combining visual features with audio and text is challenging without inherent temporal modeling.

This study proposes the use of 3D CNNs for video feature extraction to overcome these limitations.

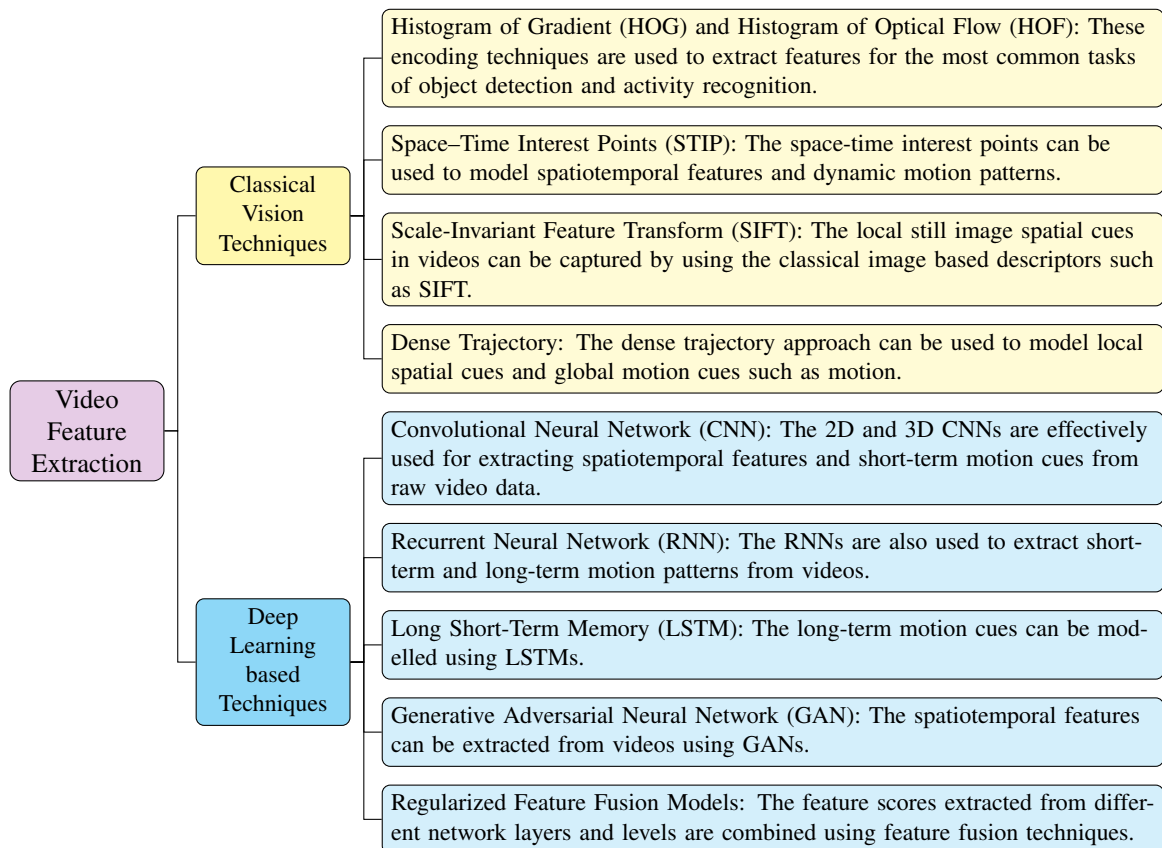


Figure 2. Classification of video feature extraction techniques

### 3. METHOD

This section presents the merits of utilizing 3D convolution for video comprehension, along with the application of 3D CNNs to capture features from video sequences in summarization algorithms. A video consists of many segments, with each segment comprising shots, and these shots are composed of sequences of frames. For a comprehensive understanding of the videos, it is necessary to learn feature representations at different levels. To extract feature vectors at different levels, the video is divided into small, non-intersecting shots. After segmenting the video, features are extracted using pretrained 3D convolutional networks. Figure 3 illustrates the extraction of video features using pretrained 3D CNN [15].

#### 3.1. 2D and 3D convolution

The fundamental difference between 2D and 3D convolution lies in the dimensionality of the input data that each processes. Generally, 2D convolution is employed on two-dimensional data, like images. This convolutional process entails moving a 2D kernel/filter across the input image, conducting element-wise multiplications, and subsequently aggregating the results. The convolution and pooling are performed spatially in 2D CNNs [15]. As a result, it does not model the temporal information. Figure 4 illustrates the distinction between 2D and 3D convolution. When 2D convolution is employed on an image, it yields another image as shown in Figure 4(a). Similarly, applying 2D convolution to multiple images (treating them as distinct channels) also produces an image as the output as indicated by Figure 4(b).

3D convolution is designed for three-dimensional data, such as video sequences or volumetric data. The 3D kernel is not only applied across height and width but also extends to the depth dimension (or time, in the context of videos). This convolutional process traverses the complete volume of the input data. The convolution and pooling are performed spatiotemporally in 3D CNNs. As a result, it preserves the temporal information outputting a volume as shown in Figure 4(c).

The advantages of 3D convolution over 2D convolution are prominent in tasks involving spatiotemporal data, such as video processing. The benefits of 3D CNNs include:

- Spatial-temporal features: it integrates spatial and temporal features simultaneously for a comprehensive data representation, crucial for video sequences.
- Temporal information capture: it effectively captures temporal information by considering the time dimension, which is essential for video analysis and action recognition.
- Natural extension for video analysis: it extends CNN capabilities for video understanding by inherently considering the temporal dimension.
- Unified framework for video processing: it provides a unified approach for processing both spatial and temporal dimensions, simplifying architecture compared to separate 2D and 1D processing units.
- Volumetric understanding: it enables the modelling of volumetric data, offering comprehensive spatial and temporal understanding, beneficial for 3D medical imaging and other volumetric data tasks.

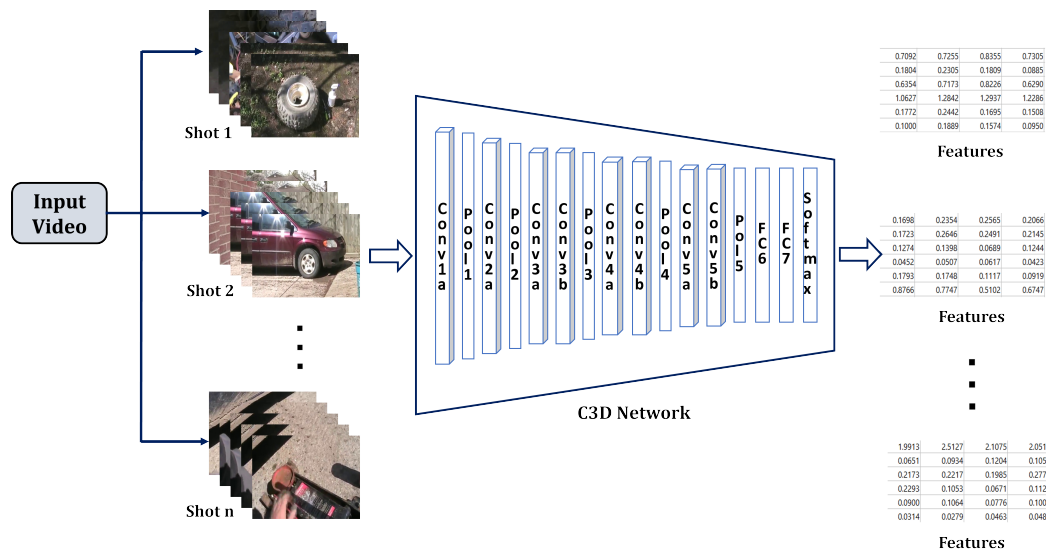


Figure 3. Extracting video features using 3D CNN

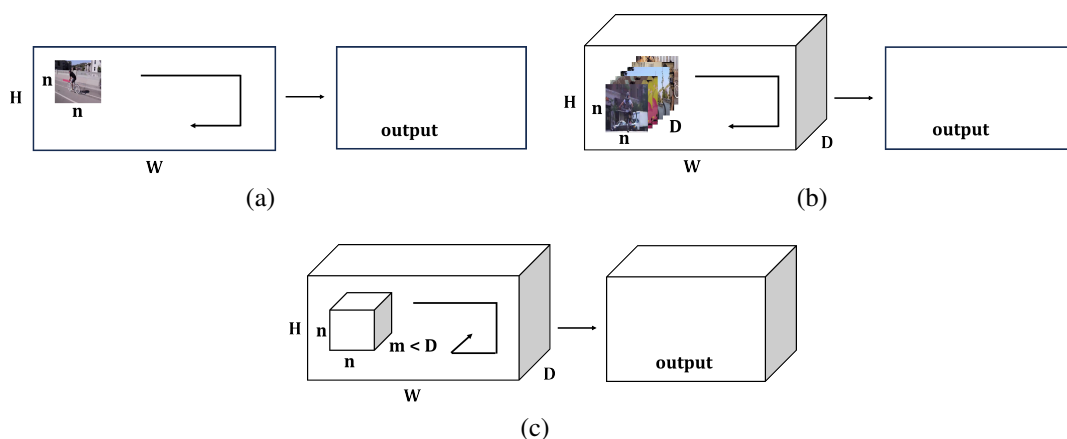


Figure 4. Comparison between 2D and 3D convolution [15]: (a) 2D convolution with an image, (b) 2D convolution with multiple frames, and (c) 3D convolution on a video

### 3.2. 3D CNNs: C3D and I3D

In video summarization methods proposed by researchers in the literature, GoogleNet is the frequently chosen deep network for feature vector extraction. GoogleNet [12] is 2D CNN pretrained on ImageNet dataset [14]. Recently, 3D CNNs, C3D, and I3D are also employed for video feature extraction. The spatiotemporal feature extraction using a 3D CNN was proposed by researchers in 2015 [15]. C3D is a deep 3D CNN with a homogeneous architecture containing  $3 \times 3 \times 3$  convolutional kernels followed by  $2 \times 2 \times 2$  pooling at each layer. The C3D model offers generic feature extraction. It provides a compact representation of video segments, generating a 4096 element vector from a 16-frame input. The model's homogeneous architecture, featuring small kernel sizes  $3 \times 3 \times 3$ , ensures fast and efficient inference, enabling optimized implementations on embedded platforms.

The I3D model, introduced by researchers in 2017, is a two-stream I3D ConvNet that extends 2D CNN principles into the 3D domain [16]. By inflating 2D filters and pooling kernels into 3D, the I3D model aims to capture spatiotemporal features from videos, leveraging successful architectures and parameters from ImageNet. Key features include the adaptation of 2D filters to 3D, expansion of the receptive field in space and time, and the use of two 3D streams for enhanced performance.

## 4. RESULTS AND DISCUSSION

This section discusses the various baseline summarization methods selected for the study, the experimentation conducted, and the results obtained from these experiments. It provides an overview of summarization techniques, experimental setup including the datasets, and the evaluation metric employed to assess the performance of the summarization methods along with the performance comparisons.

### 4.1. Summarization methods

The effectiveness of 3D CNNs in video feature extraction is demonstrated through the examination of two summarization frameworks: conventional video summarization and query-focused video summarization. An overview of summarization methods is provided.

#### 4.1.1. Conventional video summarization methods under consideration

Conventional or generic video summarization involves generating a concise video summary by automatically selecting keyframes or keyshots representing the most important content necessary for understanding the video. This type of summarization is generally content-driven, relying on the visual information within the video to determine what should be included in the summary. The methods under consideration are:

- Diversity-representativeness reward deep summarization network (DR-DSN): a deep summarization network [17] proposed for estimating the likelihood of individual video frames and generating the video summary.
- Video attention summarization network (VASNet): a summarization method [9] combining a soft self attention and two-layer regressor network.
- Positional encoding with global and local multi-head attention for summarization (PGL-SUM): integration of positional encoding with global and local multi-head attention [18] for calculating importance scores of frames.
- Summarization generative adversarial network with attention autoencoder (SUM-GAN-AAE): a supervised summarization technique leveraging the combination of adversarial learning with attention mechanism [19] for summarizing videos.
- Concentrated attention summarization (CA-SUM): a summarization network employing [11] concentrated attention considering uniqueness and diversity of video frames.
- Deep summarization network with reinforcement learning (DSR-RL): a recurrent summarization network [20] incorporating self attention mechanism and reinforcement learning.

#### 4.1.2. Query-focused video summarization methods under consideration

Query-focused video summarization generates the video summary based on specific input queries by the user. This type of summarization is context-driven, relying on viewer queries, making it more personalized than conventional summarization. The methods under consideration are:

- Three-player adversarial network (TPAN): a generative adversarial network with three players [21] operating on three sets of query-conditioned summaries to generate query-focused video summaries.

- Mapping network (MapNet): a mapping network [10] that investigates the correlation between video shots and queries.
- Hierarchical variational network (HVN): a novel architecture, hierarchical variational network [22] designed to capture long-range temporal dependencies based on queries with its multi-level variational block.
- Query-relevant segment representation module with global attention module (QSRM-GAM): a two-stage approach, consisting of the query-relevant segment representation module and global attention module [23] proposed for video summarization, taking into account user interests.
- Convolutional hierarchical attention network (CHAN): a pioneering model, convolutional hierarchical attention network [24] to employ local and global self-attention for query-focused video summarization.

## 4.2. Experimental setup

An experiment is carried out to extract the spatiotemporal features from video sequences using pre-trained C3D [15] and I3D [16] networks. The C3D network is trained on the Sports-1M dataset [25]. Motion features are obtained in the RGB and flow formats. RGB features are extracted from video frames utilizing the I3D model [16], pretrained on Kinetics 400 dataset [26], in conjunction with PWC-Net [27]. Flow features, on the other hand, are extracted using I3D network with recurrent all-pairs field transforms (RAFT) [28]. All experiments are conducted on a computer equipped with an NVIDIA RTX 3060 GPU.

### 4.2.1. Datasets

TVSum [29] and SumMe [30] are the publicly available benchmark datasets employed for generic video summarization. The SumMe dataset [30] comprises 25 videos spanning various genres like sports, holidays, and cooking. In contrast, the TVSum dataset [29] includes 50 YouTube videos across 10 categories such as documentary, educational, and egocentric. Both datasets come with multiple user annotations, including user-selected keyframes and shot-level importance scores.

For query-based video summarization, the benchmark dataset used is the query-focused video summarization (QFVS) dataset [31]. The QFVS dataset [31] includes four egocentric consumer-grade videos recorded in uncontrolled everyday scenarios, each lasting 3 to 5 hours and featuring a diverse range of events. For each video and query pair, the dataset includes four query-based summaries, consisting of one oracle summary and three user-generated summaries.

### 4.2.2. Evaluation metric

Video representations for sequences in the above-mentioned datasets are generated using C3D, I3D (RGB), and I3D (flow) networks, and F1-scores are computed. The F1-score assesses the similarity between the ground truth summary (user summary) and the generated machine summary [32]. It is the harmonic mean of precision and recall. This metric is the most commonly used approach for measuring the performance of summarization frameworks.

## 4.3. Results

The experimental results are presented in Tables 1 and 2. The results provide a comparative analysis of the performance of various video representation techniques for conventional and query-focused summarization methodologies under study.

Table 1. Comparative analysis of feature extraction techniques in generic summarization methods assessed on TVSum and SumMe datasets

Method	F1-score with GoogleNet		F1-score with C3D		F1-score with I3D (RGB)		F1-score with I3D (Flow)	
	SumMe	TVSum	SumMe	TVSum	SumMe	TVSum	SumMe	TVSum
DR-DSN [17]	42.1	58.1	55.8	65.7	55.3	65.4	55.5	65.6
VASNet [9]	49.7	61.4	51.4	63.8	60.2	62.6	60.8	62.3
PGL-SUM [18]	57.1	62.7	55.7	65.3	60.3	64.8	60.1	64.2
SUM-GAN-AAE [19]	48.9	58.3	52.3	62.7	52.1	62.6	51.7	62.3
CA-SUM [11]	51.1	61.4	61.4	66.4	61.9	65.1	60.2	64.9
DSR-RL [20]	50.3	61.4	54.8	64.5	51.7	63.4	55.3	62.8

Table 2. Comparative analysis of feature extraction techniques in query-focused summarization methods assessed on QFVS dataset

Method	Features	Result (F1-scores)
TPAN [21]	ResNet152 + C3D	46.05
MapNet [10]	ResNet152 + C3D	47.20
CHAN [24]	ResNet	46.94
HVN [22]	C3D	48.87
QSRM-GAM [23]	I3D	49.20
CHAN [24]	C3D	51.43
CHAN [24]	I3D	50.78

### 4.3.1. Results with conventional video summarization methods

Table 1 provides the performance comparison of conventional summarization frameworks in terms of F1-scores. The F1-scores reported for GoogleNet are retrieved from corresponding papers. Table 1 and Figure 5 indicate that the F1-scores for C3D and I3D video representations show a significant improvement over GoogleNet. Additionally, Figures 5(a) and 5(b) depict a rising trend in F1-scores for the SumMe and TVSum datasets, respectively. The results show that using 3D CNNs for video feature extraction has enhanced the F1-scores on the SumMe dataset, with improvements of 32.5% for DR-DSN, 3.4% for VASNet, 6.9% for SUM-GAN-AAE, 20.1% for CA-SUM, and 8.9% for DSR-RL. Similarly, on the TVSum dataset, the F1-scores have increased by 13% for DR-DSN, 3.9% for VASNet, 4.1% for PGL-SUM, 7.5% for SUM-GAN-AAE, 8.1% for CA-SUM, and 5% for DSR-RL.

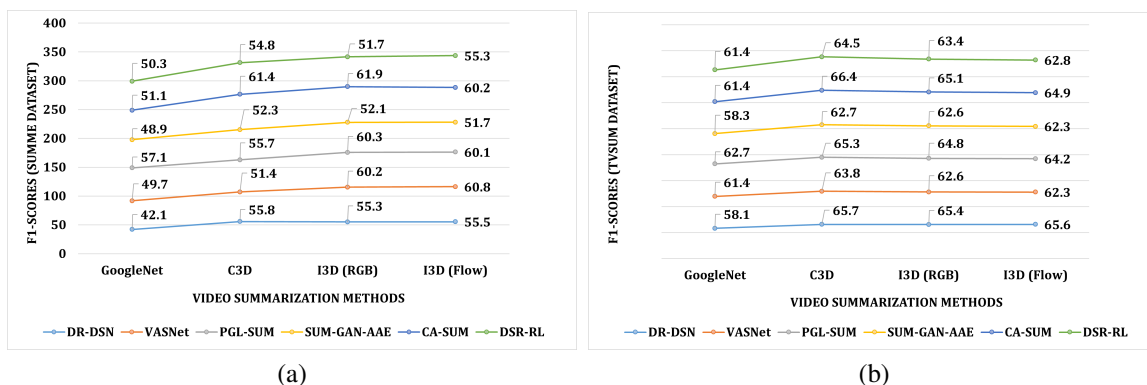


Figure 5. Performance comparison of feature extraction techniques assessed on (a) SumMe and (b) TVSum dataset

### 4.3.2. Results with query-focused video summarization methods

Table 2 provides the comparative analysis of query-based summarization frameworks in terms of F1-scores. The majority of summarization methods utilize ResNet for extracting video features. ResNet [13] is a pretrained 2D CNN trained on the ImageNet dataset. The F1-scores presented in Table 2 indicate that video representations obtained with C3D result in enhanced F1-scores.

## 4.4. Discussion

Previous studies have shown that GoogleNet and ResNet are well-established models that excel at extracting spatial features from individual video frames using inception modules, which are composed of multiple parallel convolutional filters of varying sizes. These models effectively capture intricate spatial details within each frame, making them highly suitable for image-based tasks. However, their focus on spatial features alone limits their ability to fully capture the temporal dynamics inherent in video sequences.

3D CNNs extend the capabilities of conventional 2D convolutions by integrating the time dimension, allowing for the simultaneous analysis of both spatial and temporal aspects of video data. This integration is crucial for tasks involving video sequences, where understanding motion and changes over time is as important



as recognizing spatial features within individual frames. This study investigates the use of 3D CNNs for effective video feature extraction in summarization. It is demonstrated that by capturing spatiotemporal features, 3D CNNs offer a more comprehensive representation of video content, resulting in notable improvements in performance for video summarization. Although 3D CNNs have advanced video feature extraction, there is scope for further research and development. Future research in video feature extraction can explore several promising directions like hybrid models that combine the spatial strengths of 2D CNNs with the temporal capabilities of 3D CNNs and multi-modal video analysis.

## 5. CONCLUSION

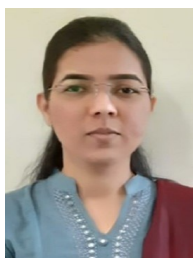
In this paper, a comparative investigation of video feature extraction using 3D CNNs, focusing on their applications in generic and query-specific video summarization is conducted. This study examines the classical and deep learning based feature extraction techniques, highlighting the advantages of deep learning approaches. The majority of existing video summarization techniques commonly rely on 2D CNNs, such as GoogleNet and ResNet, for feature extraction. It is demonstrated that 3D CNNs, such as C3D and I3D, are more effective for video feature extraction in both generic and query-specific video summarization compared to traditional 2D CNNs. By evaluating F1-scores for various summarization methods, it is concluded that 3D CNNs significantly improve performance due to their ability to capture both spatial and temporal features. This underscores the superiority of 3D CNNs in providing a more comprehensive understanding of video content, marking a notable advancement in video feature extraction and summarization techniques.





## REFERENCES

- [1] M. Suresha, S. Kuppa, and D. S. Raghukumar, "A study on deep learning spatiotemporal models and feature extraction techniques for video understanding," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 2, pp. 81–101, 2020, doi: 10.1007/s13735-019-00190-x.
- [2] A. Mirza, O. Zeshan, M. Atif, and I. Siddiqi, "Detection and recognition of cursive text from video frames," *Eurasip Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 1–19, 2020, doi: 10.1186/s13640-020-00523-5.
- [3] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013, doi: 10.1007/s11263-012-0594-8.
- [4] Y. Xian, B. Korbar, M. Douze, L. Torresani, B. Schiele, and Z. Akata, "Generalized few-shot video classification with video retrieval and feature generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8949–8961, 2021, doi: 10.1109/TPAMI.2021.3120550.
- [5] Q. Wu, Q. Huang, and X. Li, "Multimodal human action recognition based on spatio-temporal action representation recognition model," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16409–16430, 2023, doi: 10.1007/s11042-022-14193-0.
- [6] C. Liu, X. Wu, and Y. Jia, "A hierarchical video description for complex activity understanding," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 240–255, 2016, doi: 10.1007/s11263-016-0897-2.
- [7] S. Sah, T. Nguyen, and R. Ptucha, "Understanding temporal structure for video captioning," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 147–159, 2020, doi: 10.1007/s10044-018-00770-3.
- [8] R. Liu *et al.*, "Exploiting radio fingerprints for simultaneous localization and mapping," *IEEE Pervasive Computing*, vol. 22, no. 3, pp. 38–46, 2023, doi: 10.1109/MPRV.2023.3274770.
- [9] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *14th Asian Conference on Computer Vision*, 2019, vol. 11367 LNCS, pp. 39–54, doi: 10.1007/978-3-030-21074-8\_4.
- [10] Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, "Deep reinforcement learning for query-conditioned video summarization," *Applied Sciences (Switzerland)*, vol. 9, no. 4, p. 750, 2019, doi: 10.3390/app9040750.
- [11] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames," in *ICMR 2022 - Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 407–415, doi: 10.1145/3512527.3531404.
- [12] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2016, vol. 2016-December, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [14] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
- [15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.
- [16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp. 6299–6308, doi: 10.1109/CVPR.2017.502.
- [17] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1, pp. 7582–7589, doi: 10.1609/aaai.v32i1.12255.





- [18] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining global and local attention with positional encoding for video summarization," in *Proceedings - 23rd IEEE International Symposium on Multimedia (ISM)*, 2021, pp. 226–234, doi: 10.1109/ISM52913.2021.00045.
- [19] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *MultiMedia Modeling: 26th International Conference (MMM)*, 2020, vol. 11961 LNCS, pp. 492–504, doi: 10.1007/978-3-030-37731-1\_40.
- [20] A. Phaphuangwittayakul, Y. Guo, F. Ying, W. Xu, and Z. Zheng, "Self-attention recurrent summarization network with reinforcement learning for video summarization task," in *Proceedings - IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428142.
- [21] Y. Zhang, M. Kampffmeyer, X. Liang, M. Tan, and E. P. Xing, "Query-conditioned three-player adversarial network for video summarization," *arXiv preprint arXiv:1807.06677*, 2019, doi: 10.48550/arXiv.1807.06677.
- [22] P. Jiang and Y. Han, "Hierarchical variational network for user-diversified & query-focused video summarization," in *Proceedings of the 2019 ACM International Conference on Multimedia Retrieval (ICMR)*, 2019, pp. 202–206, doi: 10.1145/3323873.3325040.
- [23] S. Nalla, M. Agrawal, V. Kaushal, G. Ramakrishnan, and R. Iyer, "Watch hours in minutes: summarizing videos with user intent," in *European Conference on Computer Vision*, 2020, vol. 12539 LNCS, pp. 714–730, doi: 10.1007/978-3-030-68238-5\_47.
- [24] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang, "Convolutional hierarchical attention network for query-focused video summarization," in *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 07, pp. 12426–12433, doi: 10.1609/aaai.v34i07.6929.
- [25] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.
- [26] W. Kay *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- [27] D. Sun, X. Yang, M. Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943, doi: 10.1109/CVPR.2018.00931.
- [28] Z. Teed and J. Deng, "Raft: recurrent all-pairs field transforms for optical flow," in *Computer Vision—ECCV 2020: 16th European Conference, 2020*, pp. 402–419, [Online]. Available: [https://doi.org/10.1007/978-3-030-58536-5\\_24](https://doi.org/10.1007/978-3-030-58536-5_24).
- [29] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "TVSum: summarizing web videos using titles," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June-2015, pp. 5179–5187, doi: 10.1109/CVPR.2015.7299154.
- [30] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *European Conference on Computer Vision*, 2014, vol. 8695 LNCS, no. PART 7, pp. 505–520, doi: 10.1007/978-3-319-10584-0\_33.
- [31] A. Sharghi, J. S. Laureland, and B. Gong, "Query-focused video summarization: dataset, evaluation, and a memory network based approach," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2127–2136, doi: 10.1109/CVPR.2017.229.
- [32] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, vol. 2019-June, pp. 7596–7604, doi: 10.1109/CVPR.2019.00778.

## BIOGRAPHIES OF AUTHORS



**Bhakti Deepak Kadam**     is a research scholar in MKSSS's Cummins College of Engineering for Women, Pune. She received her BE degree in E&TC and M.Tech degree in electronics engineering from Savitribai Phule Pune University in 2011 and 2014, respectively. Her current research interests include video processing, computer vision, and deep learning. She can be contacted at email: [bhakti.kadam@cumminscollege.in](mailto:bhakti.kadam@cumminscollege.in).



**Ashwini Mangesh Deshpande**     is an associate professor in the Electronics and Telecommunication Department, at MKSSS's Cummins College of Engineering for Women, Pune, India. Her research interests include image and video processing, computer vision, deep learning, and satellite image processing. She has 40 research papers in reputed journals and conferences. She has acted as a PI in various Government-funded projects. She can be contacted at email: [ashwini.deshpande@cumminscollege.in](mailto:ashwini.deshpande@cumminscollege.in).