International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 02 | Feb 2024 www.irjet.net p-ISSN: 2395-0072

Exploring Various Techniques for Video Summarization

Ajinkya Somawanshi, Devang Shirodkar, Vinayak Yadav, Krushna Tawri, Prof. Rakhi Punwatkar

[1234] *UG Student, Dept. of computer Engineering, Zeal college of engineering, Maharashtra, India*
[5]*Professor, Dept. of computer Engineering, Zeal college of engineering, Maharashtra, India*

Abstract - *Video summarization is a critical task in multimedia analysis, especially in today's digital world, where the volume of video data is vast. Deep learning methods have been widely explored for this purpose, but they often suffer from inefficiencies in processing long-duration videos. This paper addresses the challenge of unsupervised video summarization by proposing a novel approach that selects a sparse subset of video frames to optimally represent the input video. The key idea is to train a deep summarizer network using a generative adversarial framework, comprising an autoencoder LSTM network as the summarizer and another LSTM network as the discriminator. The summarizer LSTM is trained to select video frames and decode the obtained summarization to reconstruct the input video. At the same time, the discriminator LSTM aims to distinguish between the original video and its reconstruction. The adversarial training between the summarizer and discriminator, along with regularization for sparsity, enables the network to learn to generate optimal video summaries without the need for labeled data. Evaluation of multiple benchmark datasets demonstrates competitive performance compared to fully supervised state-of-the-art approaches, showcasing the effectiveness of the proposed method in unsupervised video summarization.*

Key Words: **Event summarization · Critical information in videos · Surveillance systems · Video analysis · Multimedia analysis · Deep learning · Unsupervised learning · Autoencoder LSTM · Long short-term memory network (LSTM)**

1. INTRODUCTION

In today's digital age, videos have become one of the most influential and prevalent forms of multimedia, connecting with users quickly and effectively. The widespread availability of high-speed internet and affordable storage has led to an explosion of video data generation, with platforms like YouTube, Netflix, and social media hosting vast amounts of visual content. However, this abundance of video data presents challenges in terms of storage, bandwidth, and human resources required for analysis.

Video summarization (VS) has emerged as a crucial technique to address these challenges by condensing lengthy videos into concise representations while preserving key information. The primary objective of VS is to analyze videos by removing unnecessary frames and preserving keyframes, thus facilitating efficient browsing and structured access to video content. Automatic VS (AVS) powered by Artificial Intelligence (AI) is a rapidly growing research area, enabling the automatic summarization of lengthy videos without human intervention.

The applications of VS span various domains, including surveillance, education, entertainment, and medical diagnostics. From monitoring and tracking to creating movie trailers and enabling video search engines, the practical use cases of video summaries are diverse and far-reaching. Additionally, VS plays a vital role in reducing frame redundancy, thereby optimizing storage requirements and computational time.

This paper focuses on the problem of unsupervised video summarization, where the goal is to select a sparse subset of frames that minimizes the representation error between the original video and its summary. We propose a novel approach based on a generative adversarial framework, combining an autoencoder LSTM network as the summarizer and another LSTM network as the discriminator. By training these networks adversarially, we aim to produce optimal video summarizations without the need for labeled data.

In this paper, we present an overview of our proposed approach to unsupervised video summarization and discuss its application in various domains. We also delve into the technical details of our methodology, including the use of deep learning architectures such as CNNs and LSTMs for feature extraction and the implementation of a generative adversarial network for optimization. Through experimental evaluation of benchmark datasets, we demonstrate the effectiveness of our approach in generating high-quality video summaries.

Overall, this paper contributes to the ongoing research in video summarization by presenting a novel unsupervised approach that leverages deep learning and generative adversarial techniques to produce compact and informative video summaries across diverse domains.

2. RELATED WORK

This section reviews related: (i) problem formulations of video summarization; (ii) approaches to supervised and unsupervised video summarization; (iii) deep learning approaches; (iv) work using the generative adversarial framework in learning; (v) Unsupervised approaches; and (vi) Attention-Based Approaches.

Problem Formulations: Traditional approaches like video synopsis and montages condense video content by tracking moving objects or merging keyframes into summary images, often overlooking the preservation of temporal motion layouts. Additionally, hyperlapses offer alternative techniques for temporal manipulation. However, recent efforts have focused on storyboard generation, which represents a subset of representative video frames, albeit without leveraging deep learning methods extensively.

Supervised vs. Unsupervised Summarization: The supervised methods rely on human-annotated keyframes for training, optimizing frame selectors to minimize loss with respect to ground truth annotations. Conversely, unsupervised methods utilize heuristic criteria for keyframe selection, with transfer learning showing promise but posing challenges in ensuring domain correlations. Notably, the performance of unsupervised methods has been dominant in scenarios where obtaining human annotations is impractical, such as in military or nursing home environments.

Deep Architectures: Deep learning, particularly Long Short-Term Memory (LSTM) networks, has been employed for keyframe selection, both in forward and reverse temporal directions. Recurrent auto-encoders have also been used for representing annotated temporal intervals in highlights. While LSTM-based models like vsLSTM and dppLSTM focus on structured prediction and diversity enhancement, unsupervised generative adversarial learning models like SUM-GAN offer a novel approach by incorporating variational auto-encoder LSTMs and regularization techniques tailored for video summarization.

Generative Adversarial Networks (GANs): GANs, typically used in image problems, have emerged as a novel approach to video summarization. They extend prior methods by incorporating a new variational auto-encoder LSTM and suitable regularization for frame selection. Unlike previous approaches that primarily rely on discriminators to provide learning signals, the proposed GAN-based models integrate frame selectors, enhancing the overall summarization process.

Unsupervised Approaches: Unsupervised techniques dominate the field, with clustering-based methods and dictionary learning being prevalent for key frame identification. Clustering algorithms group visually similar frames or shots into clusters, with group centers serving as representative keyframes. Similarly, dictionary learning leverages base vectors in the model to reconstruct the visual content of the original video, effectively identifying key frames or shots.

Attention-Based Approaches: Attention-based LSTM frameworks leverage low-level features such as motion and face to capture user attention, facilitating a deeper understanding of complex viewer attention mechanisms. By modeling attention cues derived from user interaction, these frameworks can extract key shots that align with user preferences, contributing to more effective video summarization strategies.

3. DEEP LEARNING - BASED VIDEO SUMMARIZATION

Deep learning (DL) has emerged as a powerful paradigm within machine learning, offering various network structures and applications across domains such as cybersecurity, natural language processing, bioinformatics, robotics, and medical information processing. In the context of video summarization (VS), DL methods have shown remarkable effectiveness and versatility, encompassing supervised, weakly supervised, unsupervised, and reinforcement learning approaches.

3.1 Supervised Learning-Based Video Summarization

Supervised learning techniques in VS involve learning from labeled data to predict future outcomes, although acquiring well-defined datasets can be costly and challenging due to the need for domain knowledge and the vast diversity of online content. Supervised models are typically categorized as classification or regression models, utilizing algorithms like linear classifiers, k-nearest neighbors, support vector machines, decision trees, and random forests. Deep belief networks (DBNs), deep neural networks (DNNs), and convolutional neural networks (CNNs) are prominent DL techniques employed in supervised video summarization, each offering unique capabilities in feature extraction and classification. For instance, DBNs leverage a deep architecture of stacked restricted Boltzmann machines (RBMs) for feature extraction and classification, while DNNs enhance model accuracy through multiple hidden layers. CNNs, on the other hand, excel in extracting high-level features from video frames through convolutional and pooling layers.

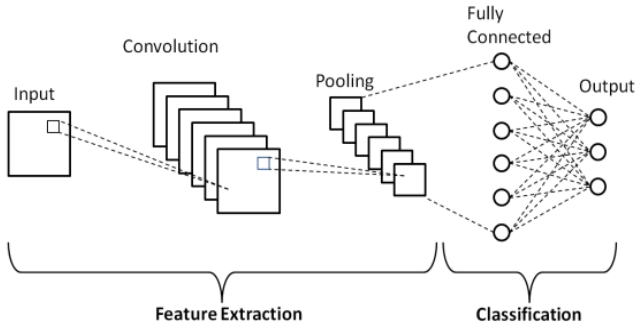

Fig. 3.1 Basic Architecture of CNN

3.2 Weakly Supervised Learning-Based Video Summarization

Weakly supervised learning strikes a balance between supervised and unsupervised approaches, requiring only a small amount of labeled data for training. Methods in this category, such as weakly supervised reinforcement learning, combine networks like the Video Classification Sub-Network (VCSN) and Summary Generation Sub-Network (SGSN) to construct meaningful video summaries while minimizing the need for extensive labeling.

3.3 Unsupervised Learning-Based Video Summarization

Unsupervised learning techniques in VS operate without labeled data, relying on clustering, association, and dimensionality reduction methods like principal component analysis (PCA), k-means clustering, and singular value decomposition (SVD). Generative adversarial networks (GANs) have emerged as a robust unsupervised learning framework for video summarization, enabling the generation of informative summaries through adversarial training between a generator and discriminator network.

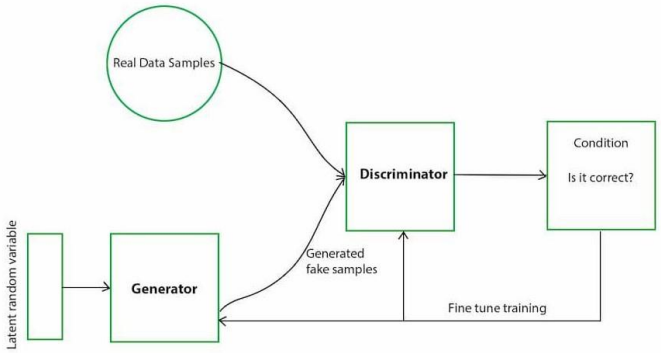

Fig. 3.3 General Working of Generative Adversarial Networks

3.4 Reinforcement Learning-Based Video Summarization

Reinforcement learning (RL) approaches in VS involve sequential decision-making processes, where an agent learns to maximize rewards through trial and error. RL-based methods leverage hierarchical LSTM networks, 3D spatiotemporal U-Nets, and diverse reward functions to generate comprehensive and representative video summaries while adapting to varying video content and lengths.

4. REVIEW OF VAE AND GAN

Variational Autoencoder (VAE) is a directed graphical model that defines a posterior distribution over observed data given an unobserved latent variable. It involves two key components: the encoder, which maps the input data to a latent space, and the decoder, which reconstructs the input data from the latent space. Learning in VAE is achieved by minimizing the negative log-likelihood of the data distribution, which involves two terms: the reconstruction loss and the Kullback-Leibler divergence term.

Generative Adversarial Network (GAN) is a neural network framework consisting of two competing subnetworks: a 'generator' network (G) and a 'discriminator' network (D). The generator network generates data mimicking an unknown distribution, while the discriminator network discriminates between generated samples and true observations. The objective of GANs is to find a generator that fits the true data distribution while maximizing the probability of the discriminator making a mistake. This is formulated as a minimax optimization problem, where D is trained to maximize the probability of correct sample classification (true vs generated), and G is simultaneously trained to minimize $\log(1 - D(\hat{x}))$.

5. Main component of our approach

Our approach consists of two main components: the summarizer and the discriminator recurrent networks. The summarizer uses a selector LSTM (sLSTM) to pick out a subset of frames from the input video, which have been preprocessed using CNN's deep features. The selected frames are then encoded into a deep feature using an encoder LSTM (eLSTM). The sLSTM generates importance scores to guide the selection process, which are used to weigh the input sequence of frame features before it is fed into the eLSTM. The final component of the summarizer is a decoder LSTM (dLSTM), which takes the encoded deep feature as input and reconstructs a sequence of features corresponding to the input video.

The role of the discriminator is to distinguish between the original video and the reconstructed video, assigning them to the 'original' and 'summary' classes, respectively. It employs a classifier LSTM (cLSTM) with a binary classification output. The training process involves adversarial training, where the cLSTM is trained to accurately recognize reconstructed sequences as 'summary' while confusing reconstructed sequences with 'original' ones.

Our training approach utilizes four loss functions: LGAN represents the augmented GAN loss, while Lreconst denotes the reconstruction loss for the recurrent encoder-decoder. Additionally, we incorporate an additional frame selector, governed by a prior distribution, which produces the encoded representation and the reconstructed feature sequence. The adversarial training of cLSTM is regulated to ensure high accuracy in recognizing the reconstructed summary as 'summary' while causing confusion between the reconstructed summary and 'original' videos.

In summary, our method involves adversarial training between the summarizer (comprising sLSTM, eLSTM, and dLSTM) and the discriminator (cLSTM) until the discriminator can no longer distinguish between the reconstructed summaries and the original videos

6. Training of sLSTM, eLSTM, and dLSTM

In this section, we will explain how we approach learning two sets of parameters: (i) Summarizer parameters which represent the sLSTM, eLSTM, and dLSTM, denoted as {θs, θe, θd}, and (ii) GAN parameters that define dLSTM and cLSTM, denoted as {θd, θc}. It's worth noting that θd are shared parameters between the summarizer and GAN.

As shown in Fig. 3, our training process revolves around four loss functions: LGAN, Lreconst, Lprior, and Lsparsity. Our generative-adversarial training introduces an additional frame selector s_p, which is governed by a prior distribution, such as a uniform distribution. Sampling input video frames with s_p yields a subset that is passed to eLSTM, generating the encoded representation e_p. Subsequently, dLSTM reconstructs a video sequence x_p. We utilize x_p to regulate the learning of the discriminator, ensuring that cLSTM accurately identifies x_p as the 'summary' class while confusing x with the 'original' class. Lprior is imposed by the prior distribution over e.

We formulate an adversarial learning algorithm that iteratively optimizes three objectives:

1. Minimizes (Lreconst + Lprior + Lsparsity) to learn {θs, θe}.

2. Minimizes (Lreconst + LGAN) to learn θd.

3. Maximizes LGAN to learn θc.

We define Lreconst and LGAN as follows:

Reconstruction loss Lreconst: Instead of relying on the Euclidean distance between input and decoded output, we base Lreconst on the hidden representation in cLSTM, particularly the output of its last hidden layer, φ(x), for input x. Given that x has passed through the frame selector s and eLSTM, resulting in e, we define Lreconst as the expectation of the log-likelihood log p(φ(x)|e).

Loss of GAN LGAN: Our aim is to train the discriminator (cLSTM) to classify reconstructed feature sequences x_hat as 'summary' and original feature sequences x as 'original'. To regulate this training, we additionally enforce that cLSTM learns to classify randomly generated summaries x_p as 'summary', where x_p is reconstructed from a subset of video frames randomly selected by sampling from a uniform prior distribution.

Given these definitions of Lreconst and LGAN, along with the specification of Lsparsity, the training parameters θs, θe, θd, and θc are updated using the Stochastic Gradient Variational Bayes estimation, adapted for recurrent networks. Algorithm 1 summarizes the steps of our training approach. Note that it employs capital letters to denote a mini-batch of the corresponding variables referenced in the preceding text.

7. Variants of our Approach

This section provides details on our regularization strategies used in the learning process. We have employed three types of regularization, each contributing to a specific aspect of the summarization process.

The first type, "Summary-Length Regularization", penalizes the selection of a large number of key frames in the summary. It is defined as:

$$L_{sparsity} = (1/2M) * \sum_{t=1}^{M} (s_t - \sigma)^2$$

Here, M denotes the total number of video frames, and σ is a hyper-parameter representing the expected percentage of frames to be selected in the summary. Our approach is referred to as SUM-GAN when employing this regularization.

The second type, "Diversity Regularization" aims to ensure the selection of frames with high visual diversity to mitigate redundancy in the summary. We have used two standard definitions for diversity regularization: (i) Determinantal Point Process (DPP), and (ii) Repelling regularizer (REP). Our DPP-based regularization is defined as:

$$L_{dpp\_sparsity} = -\log(P(s))$$

Here, P(s) represents the probability assigned to the selection indicator s by DPP. Our approach is referred to

as SUM-GANdpp when employing this regularization. For the repelling regularization, we define:

$L_{rep\_sparsity} = (1/M(M-1)) \sum_t \sum_{t' \neq t} (e_t^T e_{t'})/(||e_t|| * ||e_{t'}||)$

This variant of our approach is referred to as SUM-GANrep.

The third type, "Keyframe Regularization", is tailored for the supervised setting where ground-truth annotations of key frames are provided during training. This regularization enables fair comparisons with recently proposed supervised methods. Here, we consider importance scores as 2D softmax outputs {s_t} instead of scalar values. The sparsity loss is defined as the cross-entropy loss:

$L_{sup\_sparsity} = (1/M) \sum_t$ cross-entropy$(s_t, \hat{s}_t)$

Here, $\hat{s}_t$ represents the ground-truth importance score for frame t. Our approach is referred to as SUM-GANsup when employing this regularization.

8. Data Sets used in VS

This section provides an overview of various datasets typically used for Video Summarization (VS) evaluation, along with different evaluation methodologies. The following datasets are commonly used for VS evaluation:

1. TVSum: This dataset consists of 50 videos covering various categories, such as news, tutorials, user-generated content, and documentaries. Each video ranges from 2 to 11 minutes and has been annotated by 20 individuals based on frame relevance and ratings.

2. SumMe: This dataset includes 25 movies with durations ranging from 1 to 6 minutes, covering diverse topics such as holidays, events, and games. Annotations from 15 to 18 users are available for critical portions of each film.

3. CoSum: This dataset contains 51 videos with a total length of 4444 films ranging from 11 to 25 minutes. Each video spans roughly 147 minutes, covering various topics.

4. Thumk1K: This dataset comprises videos from YouTube on topics such as skydiving, bridge crossings, sports, and cultural landmarks.

5. Open Video Project (OVP): This is a collection of 50 videos annotated with five different user keyframe sets, covering educational, transitional, scientific, comedic, and other content.

6. YouTube Platform: This dataset offers a collection of 50 videos, with the YouTube Highlight dataset containing 100 videos. Annotations are made using Amazon Mechanical Turk (AMT) technology, with individual selections lasting around 5 seconds.

7. UCF101: This dataset comprises real-time activity videos obtained from YouTube, consisting of 101 categories for action recognition. Videos are grouped into 25 categories, with each group containing four to seven action videos.

8. MSR-VTT: This dataset combines popular video search queries with videos from a commercial video search engine. It offers 10,000 online video clips with a total duration of 41.2 hours.

9. LoL: This dataset comprises 218 long videos with durations between 30 and 50 minutes. LoL annotations originate from YouTube channels featuring society highlights.

In addition, there are several other datasets that are less frequently used, such as Videos in the Wild (VTW), FVPSum, UCLA, MED-Summaries, YouCook2, and COIN datasets. Each of these datasets offers unique characteristics and annotations for VS evaluation.

9. Performance Measure

The following sections provide an overview of various methods used for evaluating Video Summarization (VS):

9.1 Static VS Evaluation:

Initially, evaluation relied on detailed criteria such as frame relevance, repetitive or missing information, and instructional importance. However, this method can be time-consuming and lacks reproducibility. To overcome these limitations, some studies assess the quality of generated summaries using objective metrics like commitment and image reconstruction capacity. For example, Chasanis et al. evaluated the rate of generated summaries using the commitment standard, while Liu et al. considered the reconstruction capacity of images. Additionally, comparative analyses of user summaries have been presented. The F-score evaluation results for static summaries on datasets like TVSum, SumMe, OVP, YouTube, and VSUMM are provided.

9.2 Dynamic VS Evaluation:

Initially, user-created datasets were employed for evaluation, followed by the use of F-score. An evaluation process introduced concurrently with the SumMe dataset utilized predefined criteria similar to the BBC. Other methods utilize the Matthews correlation coefficient or rely on a single ground-truth summary instead of multiple user summaries for evaluation. The F-score evaluation results for dynamic summaries on datasets like TVSum, SumMe, and YouTube are presented.

Observations based on F-score evaluations:

- The TVSum dataset is widely used, with the Convolutional Neural Network Bi-Convolutional Long Short Term Memory Generative Adversarial Network method achieving an F-score of 69.0% for static summaries.

- The SumMe dataset is the second most used, with the Deep Attentive Preserving method achieving an F-score of 45.5% for static summaries.

- Multi Convolutional Neural Network outperformed on the Open Video Project dataset with an F-score of 82.0% for static summaries.

- The Multi-edge optimized LSTM RNN for video summarization approach achieved the best F-score of 85.8% for static summaries on the YouTube dataset.

- For static summaries using the VSUMM dataset, the Multi-edge optimized LSTM RNN for video summarization approach achieved the highest F-score of 92.4%.

- For dynamic summaries, the Convolutional Neural Network Bi-Convolutional Long Short Term Memory Generative Adversarial Network method achieved an F-score of 72.0% on the TVSum dataset.

- The Dilated Temporal Relational-Generative Adversarial Network method performed well on the SumMe dataset, achieving an F-score of 51.4% for dynamic summaries.

- The unsupervised learning-based Cycle-SUM method outperformed with an F-score of 77.3% for generating dynamic summaries.

10. Results

We evaluate our approach on four datasets: SumMe, TVSum, Open Video Project (OVP), and YouTube.

1) SumMe: This dataset comprises 25 user videos capturing various events like cooking and sports. Video lengths range from 1.5 to 6.5 minutes, and frame-level importance scores are provided.

2) TVSum: Consisting of 50 YouTube videos from 10 categories, the TVSum dataset features diverse content and varying video lengths (1 to 5 minutes).

3) OVP: We use the same 50 videos as in previous studies. These videos span various genres and have lengths ranging from 1 to 4 minutes.

4) YouTube: This dataset comprises 50 videos collected from websites, featuring content like cartoons, news, and sports, with durations ranging from 1 to 10 minutes.

Evaluation Setup: We employ the keyshot-based metric proposed in previous research. Precision and recall are defined based on the temporal overlap between generated and user-annotated keyshots. The F-score, the harmonic mean of precision and recall, serves as the evaluation metric. We follow established procedures to convert frame-level scores to key frames and key shot summaries and vice versa across all datasets.

Implementation Details: To ensure fair comparison, we use the output of the pool5 layer of the GoogLeNet network as the feature descriptor for each video frame. Our framework includes a two-layer LSTM with 1024 hidden units for the discriminator LSTM (cLSTM), and two two-layer LSTMs with 2048 hidden units each for eLSTM and dLSTM. We adopt a decoder LSTM that reconstructs the feature sequence in reverse order. Parameters of eLSTM and dLSTM are initialized with those of a pre-trained recurrent autoencoder model on original video feature sequences. Adam optimizer with default parameters is used for training.

Baselines: Given the generative structure of our approach, we cannot entirely replace subnetworks with baselines. Thus, besides variations of our approach defined in Section 6, we also evaluate other baselines.

Quantitative Results: The model with additional frame-level supervision, SUM-GANsup, outperforms unsupervised variants. Variations with explicit regularization, such as SUM-GANdpp and SUM-GANrep, perform slightly better than SUM-GAN. SUM-GANdpp generally outperforms SUM-GANrep. Training with combined losses from VAE and GAN improves accuracy.

Comparison with State of the Art: Our unsupervised SUM-GANdpp model outperforms all unsupervised approaches across all datasets, nearly 5% better than state-of-the-art unsupervised methods on SumMe. SUM-GANsup outperforms supervised methods in all datasets except OVP.

Comparison with Shallow Features: We evaluate our model with shallow features used in previous studies and find our model consistently outperforms the state of the art, even when shallow features perform better than deep features in some cases.

Qualitative Results: We illustrate the temporal selection pattern of different approaches using an example video, showing selected frames and frame-level importance scores. Despite small variations, all approaches cover temporal regions with high frame-level scores, with most failure cases occurring in videos with very slow motions and no scene changes.

11. Challenges

Video summarization (VS) faces several challenges due to the hierarchical structure of videos, which include frames, shots, and scenes.

Multimodal Nature: Videos consist of various modalities such as images, audio, text, and rotating images, making summarization more complex than other types of content. Utilizing high-level features from different video categories poses challenges in generating summaries effectively.

Spatio-temporal Dependencies: Designing architectures to capture spatio-temporal dependencies is complex. Representing high-dimensional video features, whether shallow or deep, to convey vast amounts of information poses significant challenges.

User Subjectivity: Each video can generate multiple summaries based on user preferences, making it difficult for a single summarizer to meet all users' needs without interaction and customization. Query-specific VS requires understanding both the visual data and textual queries, making it user-specific.

Generation of Importance Scores: Determining the significance of frames or segments varies among individuals and depends on various factors such as summary classification, context, and video type.

Evaluation of Summaries: Lack of a single qualitative and quantitative evaluation metric complicates evaluation. Various metrics like F-score, precision, recall, and accuracy are used, along with viewer ratings based on criteria like informativeness, coverage, and ranking.

Application-based Challenges: VS applications span various domains like sports, surveillance, and user-generated content. Each domain poses unique challenges, such as the complexity of sports videos or the variability and quality issues of user-generated content.

Storage and Computation: DL-based VS requires large annotated datasets for learning, leading to storage and computational challenges. Annotating large-scale datasets, especially surveillance videos, is challenging due to their diversity and complexity.

Data Mining and Information Retrieval: Extracting meaningful information from raw videos is complicated due to their unstructured nature. Video data mining faces challenges in considering both perceptual and semantic content for information retrieval.

12. CONCLUSIONS

In conclusion, the exploration of various techniques for video summarization has revealed a diverse landscape of approaches, each with its strengths and limitations. From keyframe extraction and clustering to deep learning-based methods, researchers have made significant strides in enhancing the efficiency and effectiveness of video summarization processes. The choice of technique often depends on the specific requirements of the application, such as real-time processing, content understanding, or user preferences.

While traditional methods offer simplicity and computational efficiency, deep learning approaches, particularly those leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated superior performance in capturing complex temporal dependencies and semantic information. However, these methods may pose challenges in terms of computational complexity and the need for extensive labeled data.

As the field continues to evolve, future research directions could focus on hybrid approaches that leverage the strengths of both traditional and deep learning techniques. Additionally, addressing challenges related to interpretability, scalability, and the development of standardized benchmarks will contribute to the broader adoption and evaluation of video summarization methods.

In summary, the exploration of diverse techniques for video summarization underscores the importance of considering the specific requirements and constraints of the application. By combining the strengths of various approaches, researchers can pave the way for more robust and adaptable video summarization systems in the future.

REFERENCES

[1] García, J. Gallardo, A. Mauricio, J. López, and C. Del Carpio, “The Vid2Seq: LargeScale Pretraining of a Visual Language Model for Dense Video Captioning” in Proc. Int. Conf. Artif. Neural Netw. Cham, Switzerland: Springer, Sep. 2023 , pp. 635–642.

[2] S. S. Kar and S. P. Maity, “Advances in sports video summarization: An applicationbased on Cricket videos” IEEE Trans. Biomed. Eng., vol. 65, no. 3, pp. 608–618, Mar. 2022.

[3] R. Chern C. I. Serrano, V. Shah, and M. D. Abràmoff, “A Semantic Text Summarization of Long Videos” Int. J. Telemed. Appl., vol. 2018, pp. 1–14, Oct. 2021

[4] Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, “Cycle-sum: Cycleconsistent adversarial lstm networks for

unsupervised video summarization,'' in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 9143–9150.

[5] Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, ''Exploring global diverse attention via pairwise temporal relation for video summarization,'' Pattern Recognit., vol. 111, Mar. 2021, Art. no. 107677.

[6] B. Zhao, M. Gong, and X. Li, ''Hierarchical multimodal transformer to summarize videos,'' Neurocomputing, vol. 468, pp. 360–369, Jan. 2022.

[7] B. Zhao, X. Li, and X. Lu, ''HSA-RNN: Hierarchical structure-adaptive RNN for video summarization,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7405–7414.

[8] Mahasseni, M. Lam, and S. Todorovic, ''Unsupervised video summarization with adversarial LSTM networks,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 202–211.

[9] D C. L. Giles, G. M. Kuhn, and R. J. Williams, ''Dynamic recurrent neural networks: Theory and applications,'' IEEE Trans. Neural Netw., vol. 5, no. 2, pp. 153–156, Mar. 1994