JOURNAL OF ACADEMIC RESEARCH FOR HUMANITIES
ISSN 2663-192x

ORCID of the Journal: https://orcid.org/0009-0000-0723-9485
DOI Number of the Paper: https://zenodo.org/records/14494115
Edition Link: Journal of Academic Research for Humanities JARH, 4(3) Oct-Dec 2024
Link of the Paper: https://jar.bwo-researches.com/index.php/jarh/article/view/529
HJRS Link: Journal of Academic Research for Humanities JARH (HEC-Recognized for 2023-2024)

CONDENSING VIDEO CONTENT: DEEP LEARNING ADVANCEMENTS AND CHALLENGES IN VIDEO SUMMARIZATION INNOVATIONS

Corresponding & Author 1:	**FARYAL SHAMSI,** Lecturer, Department of Computer Science, Sukkur IBA University, Sindh, Pakistan, Email: faryal.shamsi@iba-suk.edu.pk.
Author 2:	**IRUM SINDHU**, Lecturer, Department of Computer Science, Sukkur IBA University, Sindh, Pakistan, Email: irum.sindhu@iba-suk.edu.pk.

Paper Information

Citation of the paper:
(Jarh) Shamsi, F., & Sindhu, I. (2024). Condensing Video Content: Deep Learning Advancements and Challenges in Video Summarization Innovations. In *Journal of Academic Research for Humanities*, 4(*4*), 113–124.

QR Code for the Paper:

Subject Areas for JARH:
1 Humanities
2 Computer Science

Timeline of the Paper at JARH:
Received on: 22-09-2023.
Reviews Completed on: 09-12-2024.
Accepted on: 15-12-2024.
Online on: 15-12-2024.

Recognized for BWO-R:
HJRS HEC Journal Recognition System

Published by BWO Researches INTL.:
BWO RESEARCHES (PVT) LTD INTERNATIONAL

DOI Image of the paper:
DOI 10.5281/zenodo.14494115

Abstract

The exponential increase in video uploads on platforms like YouTube, exceeding 500 hours per minute, presents critical challenges in indexing, retrieval, and navigation. Existing methods heavily depend on user-generated metadata, often misaligned with the actual content. To address these challenges, we conducted a systematic review of video summarization techniques employing deep learning. An initial pool of 300 research articles was screened using strict quality criteria, resulting in a final selection of 44 studies. Articles were included if they focused on video summarization, employed deep learning approaches, utilized video datasets for evaluation, and were published in English or Urdu between 2019 and 2024 in peer-reviewed journals or conference proceedings. Papers were excluded if they lacked evaluations, used non-English/Urdu datasets, or were published before 2019. This review synthesizes recent advancements, highlights practical applications, and discusses relevant datasets, offering valuable insights for researchers and practitioners seeking to enhance automated video indexing and retrieval on social networking platforms.

Keywords: YouTube, Evaluations, Systematic, Indexing, Platforms.

Introduction

With the increasing number of social media websites, video content is extensively increasing across the internet. YouTube is the most popular social networking website and must manage more than 500 hours of video content every minute. Indexing, retrieval, navigation and searching for this huge collection of video data is an enormous challenge (Shamsi et al., 2019). Furthermore, video indexing techniques often depend heavily on user-provided textual annotations, such as titles and descriptions. However, this metadata is frequently inaccurate or misaligned with the actual content of the video, limiting its reliability for effective indexing and retrieval. The user who uploads the video has complete freedom to provide captions and descriptions, which may intentionally be misleading for personal motives, potentially confusing or misguiding the audience. For instance, a user while uploading the video might label a video as "Breaking News: Major Earthquake Hits City", while the content is unrelated and instead promotes a product or personal agenda. Such misleading captions exploit viewer curiosity, resulting in inaccurate indexing and a poor user experience. Another example could be a video titled "Shocking Secret to Weight Loss Revealed!", but the actual content is a vlog about someone's vacation with no mention of weight loss. This type of mislabeling not only misguides the audience but also undermines the reliability of video search and recommendation algorithms. In the given scenario, certainly, some automation is required to comprehend the video contents to manage the video files correctly. There are various video summarization methods proposed in yesteryears. Furthermore, various review papers have also been published in the context of video summarization, but the literature still appears to be somewhat scattered, concerning the domain of research. Many review papers reference outdated literature, where the methodologies proposed are now outdated by newer fields such as deep learning. This paper specifically summarizes the video summarization techniques exploiting deep learning techniques. Moreover, the employment practices of each model with some relevant datasets are also featured in the paper. The systematic literature review given in this paper is significant for the new researchers interested in the field of deep learning pursuing multimedia content (Sindhu & Shamsi, 2023b), especially video data. The given details can be used to mend the problems concerned with video indexing and retrieval by social network platforms (Sindhu & Shamsi, 2023a). The paper provides a comprehensive review of deep learning-based text summarization, focusing on evolution, methodologies, and challenges in the field. It categorizes techniques into extractive, abstracted, and hybrid approaches, highlighting their strengths and limitations. Input representation methods, such as embedding and attention mechanisms, are explored for their role in improving model performance. The study also reviews various training strategies and architectures, including transformers and pre-trained models. Additionally, it evaluates datasets used for training and testing, emphasizing their diversity and scalability.

Research Methodology

The systematic literature review process followed in this study is outlined in Table I. The methodology is inspired by the widely accepted systematic review framework (Mujtaba et al., 2019), which is considered a standard in the field. To identify primary studies, five databases (as listed in Table I) were targeted, and a detailed search strategy was developed, as shown in Table II. This phase involved defining the study selection criteria, formulating search keywords and queries, and conducting a quality assessment of the retrieved studies. The inclusion and exclusion criteria used for quality assessment are presented in Table III. The overall review process is illustrated in the PRISMA Flow Diagram (Figure 1). Furthermore, this review categorizes video summarization

techniques, highlights the datasets used for evaluation, and provides critical appraisal alongside future research directions, as detailed in Tables IV, V, and VI, respectively.

Search Strategy: This review includes most of the studies that used Video Summarization as their source of data for deep learning using different video summarization techniques Thus, various search keywords are formulated to retrieve the related literature from two reliable and high-quality academic databases (Shamsi & Sindhu, 2021) as follows.

- Web of Science
- Elsevier/ Scopus
- IEEE Explore
- ACM Digital Library
- Science Direct
- Springer Link

Table 1

Search Strategy

Search and Review Strategy
I. Target Well-Known Academic Databases II. Selection and Grouping of appropriate Keywords with respect to topic III. Target specific types of publication with respect to year and quality IV. Query Formation to get maximum number of relvant literature

The authors prepared a list of several relevant keywords to search the relevant literature on different techniques available for video summarization using Deep learning from the selected databases. Table II shows the keywords with their corresponding group of synonyms, used to perform queries. Each keyword within the group is paired using the OR operator, whereas the groups are paired using the AND operator to form a search query. Table II presents the application of the query on the article title, abstract, and keywords to identify relevant journal and conference articles from the targeted bibliographic databases, covering publications from January 2019 to January 2024. The last row of Table. II shows how keywords from different groups are concatenated to form a query that was executed in two bibliographic databases.

Retrieved Studies

The search query is shown in Table. II, when applied to the selected five bibliographic databases, retrieved more than 300 studies. Figure. I show the number of detailed search results from each database. The search records from each database against the search query were stored in the citation manager software. Subsequently, the duplicate studies across various databases were removed and only distinct copies of each primary study were stored in Endnote.

Screening and Selection Criteria

After removing the duplicate records, the remaining 199 studies were screened based on the title, abstract, and keywords of the retrieved articles using the study inclusion and exclusion criteria (see Table III) by authors. For any discrepancies, majority voting was used to include or exclude the article. Table. II demonstrates the article title abstract, and the keywords-based screening process included only 44 articles out of 72 screened articles and excluded the remaining articles.

Data Extraction

Data extracted from the 44 selected primary studies presents a critical review of these four aspects. These aspects are tabulated in Table. And comprised the following four aspects: (1) methods of Video Summarization, (2) various characteristics of the datasets, (3) Results (4) deep learning approaches and (5) Limitations. The findings of the Systematic literature review process are discussed in the next Section.

Table 2

Query Formation with Keyword Groups and Criteria

GROUP G1 Keywords: **Video**	video, visual, movie, film, clip
GROUP G2 ywords: **Summary**	Summary, Summarization, synopsis, precis, condensation, compress, description, story, narrative, narration, chronical
GROUP G3 ords: **Deep Learning**	deep learning, artificial neural network, convolutional neural network, recurrent neural network, low short-term memory, generative adversarial neural network, generative, pretrained, transformer, ANN, CNN, RNN, LSTM, GAN, GPT
arch Criteria **C1:** ublication Type	Journal, Conference
arch Criteria **C2:** rea of Research	Computer Science
arch Criteria **C3:** ars of Publication	2019 -2024
arch Criteria **C4:** Language	English
	Search Query: (Use of OR within G1) AND (OR within G2) AND (OR within G3) + C1 + C2 + C3 + C4

Video Summarization Methods

Video summarization methods proposed in the last five years vary from each other concerning the purpose of summarizing the video and the way of presenting summaries. Some studies aim to summarize the video to save memory space (Shamsi, Nazeer, Memon, & Mangrio, 2017), some identify a specific event or set of events whereas some come up with an automated summary of the video replacing human labour. Transformers and Generative Adversarial Networks (GAN) are found to be the most effective techniques. (Khan, Hussain, Khan, Khan, & Baik, 2024) has used transformers but (Y. Zhang, Kampffmeyer, Zhao, & Tan, 2019) and (Fu, Tai, & Chen, 2019) used GANs and has achieved the most effective results so far.

Table 3

Inclusion & Exclusion Criteria

S.No	Inclusion Criteria
1	The article should have video summarization as one of the main topics
2	The article should have used deep learning as main approach
3	The article should have used some video datasets for evaluation of results
4	The article should have reported result and evaluation criteria.
5	The article must have been published between 2019 and 2024.
6	The article should be a conference proceeding or journal publication
7	The article should have used videos in Urdu or English Language
8	The article should be written in English Language
S.No	**Exclusion Criteria**
1	The paper has been published in languages other than English.
2	The article uses datasets on language other than English or Urdu.
2	The paper has been published before 2019
3	The paper does not use Deep Learning.

1. Monolithic Vs Pluralistic Video Summarization

Video Summarization A video is composed of images called frames. In supervised learning for video summarization, each frame (or keyframe) is labelled with summarization information which can be utilized to extract the overall summary of a complete video. Various video summarization techniques rely on visual features only and use the key-frames to summarize the video content. Such type of summaries no doubt generates a representative compressed version of the original video, in terms of display. Negi, Kumar, & Saini, 2024) & (Gunuganti, Yeh, Wang, & Norouzi, 2024) are examples of such monastic video summarization techniques. Pluralistic Video Summarization A long sports video, such as a cricket match or a movie contains a significant amount of textual and audio content such as scorecards, player information or a musical sequence describing some important part of the video. For the summarization of such types of videos, only visual features of the video may not be sufficient to generate a representative summary. For these videos, multi-modal features of video must be considered for summarization purposes. Studies like (Xie, Chen, Zhao, & Lu, 2024), (Pang, Nakashima, Otani, & Nagahara, 2023), (Y. Zhang, Liu, Zhu, & Kang, 2022) & (A.-A. Liu et al., 2019) are found to be pluralistic video summarization techniques.

2. Supervised Summarization of Videos

Use of Transformer

A recent paper (Khan et al., 2024) proposes a Vision Transformer (ViT)-assisted deep pyramidal refinement network for video summarization, highlighting the challenges of extracting representative features and refining them for accurate summaries. The method uses a dense prediction transformer and multi-scale feature refinement to predict frame importance scores and generate keyframe-based video summaries. Experimental results on the TV Sum and SumMe datasets show their superior performance, achieving F1 scores of 62.4% and 51.9%, respectively. The paper introduces an innovative approach by incorporating Vision Transformers for feature extraction, which effectively improves video summarization. However, while the method shows slight improvements over existing techniques, real-world applicability could benefit from further evaluation on a wider variety of video types and settings. A transformer-based method called the spatiotemporal vision transformer (STVT) (Hsu, Liao, & Huang, 2023) consists of three components: an embedded sequence module,

a temporal inter-frame attention (TIA) encoder, and a spatial intra-frame attention (SIA) encoder. The model learns inter-frame correlations among non-adjacent frames and intra-frame attention for better frame importance representations. This approach outperforms state-of-the-art methods on the SumMe and TV Sum datasets, offering more human-like video summaries. The STVT method successfully combines inter-frame and intra-frame attention mechanisms, addressing key challenges in video summarization by considering both long-range dependencies and detailed frame-level importance.

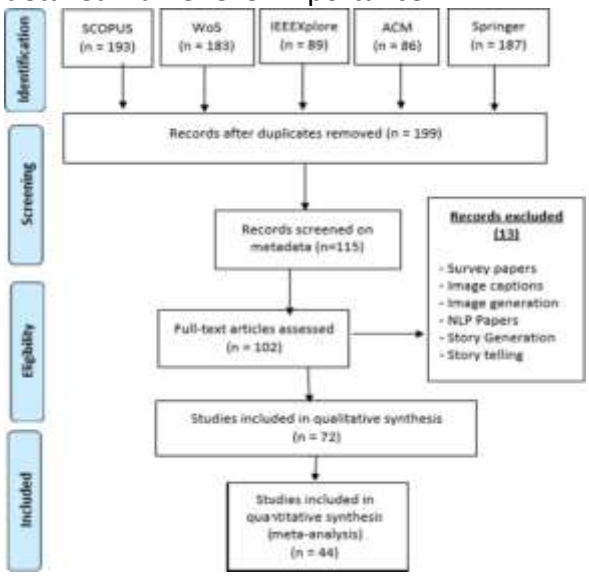

Figure. I

PRISMA Diagram - Study Selection Process

A topic-aware video summarization (Zhu, Zhao, Hua, & Wu, 2023), aimed at creating multiple summaries from the same video based on diverse user interests. Unlike traditional methods that produce a single, objective summary, this approach focuses on generating summaries tailored to specific topics such as characters, scenery, or animals. To support this task, the authors developed Topics, a new dataset of 136 videos annotated with topic labels and frame-level importance scores, incorporating visual, audio, and textual modalities for richer content representation. They also propose a multimodal Transformer-based model that simultaneously predicts topic labels and generates topic-specific summaries by adaptive fusing multimodal features. Experimental results show that the method effectively generates high-quality, topic-aware summaries, addressing diverse user preferences. Furthermore, the multimodal Transformer model's ability to leverage multiple data modalities enhances the robustness and accuracy of summary generation. However, the framework's reliance on multimodal data may present challenges for scalability and computational efficiency, especially in scenarios with limited modality-specific data (e.g., audio-free or text-scarce videos). Additionally, while topic Sum represents a significant improvement over existing datasets, its size and scope may not fully capture the variability in real-world video contexts. Expanding the dataset with more diverse and longer videos could further validate the model's applicability in broader scenarios.

Use of Auto-encoders

A multi-modal self-supervised learning framework for video summarization (H. Li et al., 2023) addresses the challenge of small-scale annotated datasets. The method explores semantic consistency between video and text at both coarse-grained and fine-grained levels, along with frame recovery, to obtain meaningful video representations. Experiments demonstrate the effectiveness of the approach in achieving superior performance in rank correlation and F-score compared to existing methods. The proposed multi-modal self-supervised approach effectively addresses the challenge of data scarcity in video summarization, offering a promising alternative to traditional deep learning methods that rely on large, annotated datasets. While the results are promising, the scalability and generalization of the approach to different types of video content beyond the newly collected dataset would require further investigation.

Use of CNN, RNN or LSTM

An innovative deep learning strategy of Capsule Network has been used by (Huang & Wang, 2019) whereas (J. Wang et al., 2019)

exploits RNN for the successful completion of video summarization tasks. Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTM networks are well-suited to classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. The (Ji et al., 2020) uses the same technique. A convolution neural network or CNN is a widely used deep learning technique that works on structured arrays of data such as images. Convolution neural network contains several stacked layers each one after another, each layer is capable of comprehending shapes that may be sophisticated in nature. (Chu & Liu, 2019), (J. Wang et al., 2019) and (Elfeki & Borji, 2019) use CNN for summarizing the given video. However, (Lal et al., 2019) & (Y. Zhang et al., 2019) combine both approaches of LSTM and CNN.

Use of GAN

(Y. Zhang et al., 2019) & (Fu et al., 2019) are GAN-based Methods for Video Summarization. (Gao et al., 2020) & (L. Yuan et al., 2019) performed unsupervised summarization, on videos from various categories using a Graph Neural Network with an active learning graph and Cycle-consistent Adversarial Bidirectional-LSTM architecture respectively. On the other hand (Milbich, Bautista, Sutter, & Ommer, 2017) exploit Convolutional Neural Network for summarization specifically on sports videos with better performance. The object of this study was to summarize the video based on similar postures, rather than creating a reflective summary of the complete contents of videos.

Other Methods

A Boundary-Aware framework for Summary Clip Extraction (BASE) (Q. Li, Chen, Xie, & Han,2023) presents a method designed to enhance video summarization in event-focused videos, aiming to extract clips representing significant events with accurate boundaries and completeness. BASE overcomes the challenges of boundary detection and clip completeness by introducing a distance-based importance signal, referred to as progress information. This signal enables the model to detect boundary transitions more precisely, while also preserving the integrity of event-related clips. The framework learns the semantic shifts along video frames and utilizes this information to generate non-overlapping summary clips. The approach outperforms existing methods by effectively balancing boundary accuracy and event completeness. The BASE framework successfully addresses the trade-off between boundary accuracy and clip completeness in event-centric video summarization by incorporating progress information. This distance-based signal improves boundary detection and overcomes limitations found in previous methods that relieve binary classification or similarity metrics. By integrating both temporal and spatial features, the model improves its overall precision and adaptability. However, processing both types of information simultaneously may increase computational demands, which could be a concern for real-time use.

Table 4

Video Summarization Method Mapping

Method	Supervised	Unsupervised	Reinforcement
CNN	(Lal, Duggal, & Sreedevi, 2019) , (Y. Zhang et al., 2019) (Chu & Liu, 2019), (J. Wang et al., 2019) and (K. (Elfeki & Borji, 2019)		
RNN/LSTM	(Huang & Wang, 2019) , (J. Wang et al., 2019), (K. (Lal et al., 2019), and (Y. Zhang et al., 2019)		
Spatio-Temporal LSTM		(Hu, Hu, Wang, Xiong, & Zhong, 2022)	
Auto-encoders	(H. Li, Ke, Gong, & Drummond, 2023)		
Attention Model	(Wu, Song, Wang, & Zhang, 2024), (Ji, Jiao, Pang, & Shao, 2020), (Y.-T. Liu, Li, Yang, Chen, & Wang, 2019) and (Lal et al., 2019)		
Transformers	(Khan et al., 2024), (Hsu et al., 2023), (Zhu et al., 2023) and (Zhu et al., 2023)		
GAN	(Y. Zhang et al., 2019), (Z. (Fu et al., 2019), (Gao, Yang, et al., 2020) and (L. Yuan, Tay, Li, & Feng, 2019)		
YOLO, ResNet, VGG		(Negi et al., 2024) and (Gunuganti et al., 2024)	
Network Addressing		(Xie et al., 2024) and (Zang, Yu, et al., 2023)	
Diffusion Models			

3. Unsupervised Summarization of Videos

Use of YOLO, ResNet, and VGG

To minimize the requirement of Benchmarks and Ground-Truth availability supervised techniques are being used in the context of Video Summarization with Deep Learning. (Negi et al., 2024) proposes a deep learning-based video summarization technique to extract keyframes efficiently, leveraging YOLOv5 for object detection, ResNet-50 and VGG-16 for feature extraction, and PCA for dimensionality reduction. Using K-means clustering, Silhouette scores, and Pearson Correlation Coefficient (PCC), the method refines frame selection, demonstrating superior recall performance on industrial surveillance datasets. Another author (Gunuganti et al., 2024) proposes graph-based modelling and attention mechanisms to predict the importance of frames and generate concise video summaries. The model employs graph representation to learn relationships between frames and an attention mechanism to capture the significance of these relationships. The encoder is trained adversarial using a graph-based generator and discriminator, eliminating the need for human annotations. Experimental results on benchmark datasets (TVSum, SumMe) show that the proposed model outperforms existing methods, demonstrating its ability to generate high-quality video summaries with competitive performance on various evaluation metrics.

Table 5

Dataset Mapping Review

Name	Used by
TVSum and TVSum	(Khan et al., 2024),(Hsu et al., 2023), (Wu et al., 2024), (J. Wang et al., 2019), (Huang & Wang, 2019), (Fu et al., 2019), (Apostolidis, Metsai, Adamantidou, Mezaris, & Patras, 2019), (He et al., 2019), (Zhao, Li, & Lu, 2019), (Harakannanavar et al., 2022) and (Ghauri et al., 2021)
TVSum	(Y. Zhang et al., 2019)
YouTube	(Fu et al., 2019), (Zhao et al., 2019)
VUSUM and RAI	(Huang & Wang, 2019)
TopicSum	(Zhu et al., 2023)

Network Addressing

A Knowledge-Aware Multimodal Network (KAMN) for video summarization (Xie et al., 2024) integrates visual, audio, and implicit semantic knowledge from external databases. The model uses a knowledge-based encoder to enhance frame representations and a fusion module to effectively combine multimodal features. Experiments show that KAMN outperforms existing supervised methods and performs competitively in unsupervised settings, demonstrating superior video summarization quality. A Deep Non-Local Video Summarization Network addressing the limitations of traditional methods by capturing long-term dependencies between video frames is proposed by (Zang et al., 2023). Unlike recurrent models like LSTM, DN-VSN uses non-local blocks to process frames in parallel, reducing information loss across frames. The network also considers video distribution factors when calculating shot-level scores, improving summary quality. Experimental results demonstrate that DN-VSN outperforms existing unsupervised methods on several benchmark datasets, making it an effective solution for video summarization. Another paper (Pang et al., 2023) introduces the Motion-Assisted Reconstruction Network (MAR-Net) for unsupervised video summarization, which effectively integrates appearance and motion information. MAR- -Net uses a Bidirectional Modality Encoder (BiME) to model dependencies through bidirectional attention and a Video Context Navigator (VCN) to enhance semantic consistency between video contexts. Experimental results show MAR- -Net outperforms existing methods. (Zhong, Lin, Lu, Fares, & Ren, 2022) presents the Deep Semantic and Attentive Network for Video Summarization (DSAVS), an unsupervised framework that selects the most semantically representative video summary without requiring frame-level labels. By minimizing the distance between video and text representations, DSAVS addresses the

challenge of capturing long-range temporal dependencies through a self-attention mechanism.

3.*Deep Reinforcement Learning for Summarization of Videos*

Diffusion Models

Deep reinforcement learning is gaining popularity for solving problems where the environment is complex or uncertain. As video content is also diverse in nature and the feature set is different when dealing with videos from various domains. Deep reinforcement learning has performed far better in video summarization problems. Deep reinforcement learning-based algorithms are mostly used for summarizing open-domain videos and sometimes on surveillance videos. (Yu et al., 2024) proposes the Diffusion Model of Feature Fusion (DMFF), an unsupervised video summarization method designed to address the instability issues of GAN-based methods and the challenges of reward function formulation in reinforcement learning.

Shot-level Semantics

The RCL framework (Y. Zhang et al., 2022) addresses two key issues - feature representation and inefficient context modelling. The framework combines an Optimized Coding Module (OCM) based on GRU for concise shot representations, with contrastive learning to enhance feature discrimination. Additionally, the Dissimilarity-Guided Attention Graph (DGAG) aggregates features based on semantic dissimilarity to improve context modelling. Reinforcement learning with shot-level semantics is exploited by (Y. Yuan & Zhang, 2022) to address challenges in video summarization, such as user subjectivity and redundant information. The method employs an encoder-decoder model, where a convolutional neural network (CNN) extracts features, and a bidirectional LSTM decoder assigns probability weights to keyframe selection, preserving spatiotemporal dependencies. To mitigate user bias, a shot-level semantic reward function is designed, ensuring more representative summaries. The approach is evaluated on four benchmark datasets—SumMe, TVSum, CoSum, and VTW—and outperforms existing methods. (Mathews et al., 2023) presents an unsupervised reinforcement learning framework for summarizing ultrasound videos, and improving triage in emergency departments and telemedicine. Using attention-based encoders and a Bi-LSTM network, the method classifies and segments key landmarks from videos. Tested on lung ultrasound data, it achieved over 80% precision, 44% F1 score, and a 77% reduction in storage, benefiting telemedicine's bandwidth and storage efficiency.

Table 6

Video Summarization Literature Summary

Study	Methodology	Future Direction
(Khan et al., 2024)	Dense Transformer + multi-scale feature refinement	Increases computational complexity
(Hsu et al., 2023)	Spatiotemporal vision transformer (STVT)	High Cost of longer Videos
(Zhu et al., 2023)	Multimodial Transformer	Useless for audio or text free videos
(Lal et al., 2019)	MerryGoRoundNet+LSTM	Need better accuracy
(Ji et al., 2020)	Attentive Adversarial Seq2Seq model	Obtain Contextual Information
(Y. Zhang et al., 2019)	DTR-GAN +LSTM	-
(Elfeki & Borji, 2019)	actionness estimation ranking	Need better accuracy
(Chen et al., 2019)	GAN + Attention Model	Embed multi-modal information such as audio
(Zhao et al., 2019)	RNN: LSTM	Shot boundary Detection
(Huang & Wang, 2019)	Capsule Net, Attention, LSTM Model	Shot Boundary Detection
(Rochan & Wang, 2019)	F: V– >S Adversarial Mapping	Applicable for small scale videos
(Chu & Liu, 2019)	Spatio-temporal Neural Net, LSTM	More features for better performance
(Apostolidis et al., 2019)	SUM-GAN model + stepwise labels learning process	Best unsupervised algorithm
(He et al., 2019)	unsupervised attentive conditional GAN	-
(Y.-T. Liu et al., 2019)	Hierarchical Multi-Attention Network (H-MAN)	Claims better performance the RNN
(H. Wang et al., 2019)	LSTM + SMN Stacked Memory Network	-

Semantic Similarity Reward

Semantic similarity is a reward criterion, ensuring that the summary generated is semantically aligned with the original video

proposed by (Mathews et al., 2023). The approach combines a video classification sub-network (VCSN) for extracting semantic representations and a summary generation sub-network (SGSN) trained via reinforcement learning. The method achieves state-of-the-art performance on multiple evaluation metrics. The incorporation of semantic similarity as a reward is a significant advancement in video summarization, bridging the gap between unsupervised and weakly supervised techniques.

Weakly Supervised Models

Furthermore, e (Afzal & Tahir, 2021) & (Chen et al., 2019) have used deep reinforcement learning techniques using visual features but (Yaliniz & Ikizler-Cinbis, 2021) has reported comparatively better performance on open-domain reinforcement learning but still needs improvement. Furthermore, (Messaoud et al., 2021) identify that the F-1 score is not a suitable metric for the evaluation of video summarization in their case, the justification applies to the prior studies with lower performance as well. More suitable evaluation metrics are used by (Boqing, 2021) & (W. Zhang et al., 2019) such as BLEU4, CIDEr, METEOR and ROGUE-L as described in Section IV-B.

Table 7

Study	Min. Accuracy	Max. Accuracy	Overall
(Khan et al., 2024)	51.9	62.4	57.15
(Wu et al., 2024)	48.0	58.8	53.4
(Ji et al., 2020)	45.5	63.6	54.55
(Elfeki & Borji, 2019)	40.1	56.3	48.2
(Chen et al., 2019)	43.6	58.4	51
(Zhao et al., 2019)	43.7	59.2	51.45
(Huang & Wang, 2019)	46	58	52
(Rochan & Wang, 2019)	48	56.1	52.05
(Chu & Liu, 2019)	47.3	58	52.65
(Apostolidis et al., 2019)	47.8	58.4	53.1
(He et al., 2019)	47.2	59.4	53.3
(T. Liu et al., 2019)	51.8	60.4	56.1
(J. Wang et al., 2019)	58.3	64.5	61.4

Results

The findings suggest that unsupervised methods of video summarization perform far better when a specific category of video is targeted. On the other hand, if the same technique is used on broad categories of videos, then performance is degraded.

Video Summarization Datasets: Video summarization relies heavily on the availability of high-quality datasets. While some researchers develop their datasets for convenience and tailored evaluation, these datasets may be perceived as biased. Publicly available benchmark datasets are preferred as they ensure reproducible results and fair comparisons across different methods.

SumMe Dataset: The SumMe dataset comprises 25 videos, each accompanied by approximately 15 human-generated summaries, resulting in over 375 summaries. This dataset, which includes ground-truth values, is widely used as a benchmark for evaluating video summarization techniques, making it an essential resource for new researchers.

***TVSum Dataset*:** The TVSum dataset (Title-based Video Summarization) serves as another benchmark for evaluating video summarization methods. It contains 50 videos with 1,000 annotations and is freely available online.

YouTube [small]: This dataset consists of 50 videos, with both colour and sound information. The genres span documentaries, educational content, ephemeral films, historical footage, and lectures. Video durations range from 1 to 4 minutes, with a total aggregate duration of 75 minutes.

TopicSum: This dataset is a text-based dataset designed for topic-focused extractive summarization. It contains multiple documents

grouped by topics, each accompanied by human-written summaries. While primarily textual, the dataset's principles can inspire topic-based approaches in multimodal tasks, such as video summarization.

Evaluation Metrics

Evaluation metrics such as Accuracy, F-Score, BLEU, CIDEr, METEOR, and ROUGE-L are essential for assessing the quality of video summarization. Accuracy measures the proportion of correctly predicted or selected keyframes or scenes compared to the ground truth, providing a straightforward measure of how well the summary aligns with human expectations. F-Score balances precision (how many selected frames are relevant) and recall (how many relevant frames are selected), offering a harmonic means to evaluate the quality of the summary. BLEU, often adapted for video summarization, evaluates the overlap of n-grams (e.g., frame sequences or shot descriptions) between the generated summary and a reference summary, with BLEU-4 focusing on 4-grams for more comprehensive analysis. CIDEr measures the semantic similarity between generated summaries and multiple references by weighing n-grams with TF-IDF, capturing the consensus among human evaluators. METEOR aligns generated summaries with references through exact matches, stemming, and paraphrasing, emphasizing recall and nuanced matching. ROUGE-L evaluates summaries based on the longest common subsequence (LCS) between generated and reference summaries, ensuring that both content and sequence coherence are captured.

Potential Gaps

While benchmark datasets are instrumental in advancing research, they are not without biases. Many datasets emphasize specific types of videos (e.g., user-generated content or educational videos), leaving other genres (e.g., cinematic or sports videos) underrepresented. Additionally, cultural and linguistic biases may arise if datasets are curated from a narrow demographic, limiting their global applicability.

Conclusion

The exponential increase in video content, driven by the proliferation of social media platforms, has created an urgent need for effective video summarization techniques. These methods are designed to distil video content into concise representations, capturing the essential segments while preserving the context and key events. Despite notable progress in this field, challenges related to dataset biases, scalability, and adaptability persist.

Dependency on Datasets

Benchmarks such as SumMe and TVSum dominate the evaluation landscape. While these datasets provide a valuable starting point, their limitations, including restricted diversity, short video durations, and cultural biases, their applicability to broader contexts.

i. Dominance of LSTM Architectures

Many reviewed studies employ Long Short-Term Memory (LSTM) networks, which excel at capturing temporal dependencies in videos. However, reliance on a single deep learning architecture underscores a lack of exploration into newer, potentially more effective models, such as Transformer-based architecture.

ii. Performance Discrepancies

The effectiveness of summarization techniques varies significantly across datasets and video types. This suggests that many methods are highly dataset-specific, lacking generalizability to diverse real-world scenarios.

Recommendations

i. Augmenting datasets

Incorporate longer, more diverse videos from underrepresented genres and languages.

ii. Dynamic benchmarks

Develop datasets that evolve, incorporating emerging video types and annotation schemes.

iii. Cross-dataset evaluations

Encourage testing algorithms across multiple datasets to mitigate overfitting to a single benchmark.

Actionable Recommendations for Future Research

i. Diversification of Datasets

Develop and adopt benchmarks that encompass a wider range of video genres, lengths, and cultural contexts to ensure robust, generalizable solutions.

ii. Exploration of Novel Architectures

Beyond LSTMs, emerging models such as Transformers and graph-based neural networks should be explored for their potential to enhance temporal and contextual understanding in video summarization.

iii. Focus on Real-World Applicability

Techniques should be evaluated not only on academic benchmarks but also in real-world applications, such as social media indexing, where scalability and computational efficiency are crucial.

iv. Cross-Dataset Validation

To address the problem of overfitting, future studies should validate their methods across multiple datasets, ensuring their adaptability and effectiveness in diverse scenarios.

References

Ahmed, M., & Sh, M. (2021). Design, usage and impact of virtual university mobile LMS application on students learning of virtual university of Pakistan. International Journal of Advanced Trends in Computer Science and Engineering, 10(3), 1837-1843.

Fu, T.-J., Tai, S.-H., & Chen, H.-T. (2019). Attentive and adversarial learning for video summarization. In 2019 winter conference on applications of computer vision (wacv) (pp. 1579–1587).

Gao, J., Yang, X., Zhang, Y., & Xu, C. (2020). Unsupervised video summarization via relation-aware assignment learning. IEEE Transactions on Multimedia.

Ghauri, J. A., Hakimov, S., & Ewerth, R. (2021). Supervised video summarization via multiple feature sets with parallel attention. In 2021 ieee an international conference on multimedia and expo (icme) (p. 1-6s). doi: 10.1109/ICME51207.2021.9428318

Gunuganti, J., Yeh, Z.-T., Wang, J.-H., & Norouzi, M. (2024). Unsupervised video summarization with adversarial graph-based attention network. Journal of Visual Communication and Image Representation, 104200.

Harakannanavar, S. S., Sameer, S. R., Kumar, V., Behera, S. K., Amberkar, A. V., & Puranikmath, V. I. (2022). Robust video summarization algorithm using supervised machine learning. Global Transitions Proceedings, 3(1), 131–135.

He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Guan, H. (2019). Unsupervised video summarization with attentive conditional generative adversarial networks. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 2296–2304).

Hsu, T.-C., Liao, Y.-S., & Huang, C.-R. (2023). Video summarization with spatiotemporal vision transformer. IEEE Transactions on Image Processing, 32, 3013-3026. doi: 10.1109/TIP.2023.3275069

Hu, M., Hu, R., Wang, Z., Xiong, Z., & Zhong, R. (2022). Spatiotemporal two-stream lstm network for unsupervised video summarization. Multimedia Tools and Applications, 81(28), 40489–40510.

Huang, C., & Wang, H. (2019). A novel key-frames selection framework for comprehensive video summarization. IEEE Transactions on Circuits and Systems for Video Technology, 30(2), 577–589.

Ji, Z., Jiao, F., Pang, Y., & Shao, L. (2020). Deep attentive and semantic preserving video summarization. Neurocomputing, 405, 200–207.

Khan, H., Hussain, T., Khan, S. U., Khan, Z. A., & Baik, S. W. (2024). Deep multi-scale pyramidal features network for supervised video summarization. Expert Systems with Applications, 237, 121288.

Lal, S., Duggal, S., & Sreedevi, I. (2019). Online video summarization: Predicting the future to better summarize the present. In 2019 winter conference on applications of computer vision (wacv) (pp. 471–480).

Li, Q., Chen, J., Xie, Q., & Han, X. (2023). Video summarization for event-centric videos. Neural Networks, 161, 359–370. Li, Z., & Yang, L. (2021). Weakly supervised deep 10 reinforcement learning for video summarization with semantically meaningful reward. In Proceedings of the ieee/cvf winter conference on applications of computer vision (pp. 3239–3247).

Liu, A.-A., Shao, Z., Wong, Y., Li, J., Su, Y.-T., & Kankan halli, M. (2019). Lstm-based multi-label video event detection. Multimedia Tools and Applications, 78, 677–695.

Liu, Y.-T., Li, Y.-J., Yang, F.-E., Chen, S.-F., & Wang, Y.-C. F. (2019). Learning hierarchical self-attention for video summarization. In 2019 ieee international conference on image processing (icip) (pp. 3377–3381).

Mathews, R. P., Panicker, M. R., Hareendranathan, A. R., Chen, Y. T., Jaremko, J. L., Buchanan, B., . . . Mathews, G. (2023). Unsupervised multi-latent space rl framework for video summarization in ultrasound imaging. IEEE Journal of Biomedical and Health Informatics, 27(1), 227-238. doi: 10.1109/JBHI.2022.3208779

Messaoud, S., Lourentzou, I., Boughoula, A., Zehni, M., Zhao, Z., Zhai, C., & Schwing, A. (2021). Deepqamvs:

Query-aware hierarchical pointer networks for multi-video summarization. In (p. 1389-1399).

Minhas, S., Hussain, T., Ghani, A., Sajid, K., & Pakistan, L. (2021). Exploring students online learning: A study of Zoom application. Gazi University Journal of Science, 34(2), 171-178.

Milbich, T., Bautista, M., Sutter, E., & Ommer, B. (2017). Unsupervised video understanding by the reconciliation of posture similarities. In Proceedings of the ieee international conference on computer vision (pp. 4394–4404).

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., . . . Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. Expert Systems with Applications, 116, 494-520. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417418306110

doi: https://doi.org/10.1016/j.eswa.2018.09.034

Negi, A., Kumar, K., & Saini, P. (2024). Object of interest and unsupervised learning-based framework for an effective video summarization using deep learning. IETE Journal of Research, 70(5), 5019–5030.

Pang, Z., Nakashima, Y., Otani, M., & Nagahara, H. (2023). Contrastive losses are natural criteria for unsupervised video summarization. In Proceedings of the ieee/cvf winter conference on applications of computer vision(pp. 2010–2019).

Rochan, M., & Wang, Y. (2019). Video summarization by learning from unpaired data. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition (pp. 7902–7911).

Shamsi, F., Nazeer, M. I., Memon, R. A., & Mangrio, M. I. (2017). Reflections of practical implementation of the academic course analysis and design of algorithms taught in the universities of Pakistan. Sukkur IBA Journal of Computing and Mathematical Sciences, 1(2), 31–38.

Shamsi, F., Sher, M. D., & Shaikh, S. (2019). Content-based automatic video genre identification. International Journal of Advanced Computer Science and Applications, 10(6).

Shamsi, F., & Sindhu, I. (2021). Improving DBLP efficiency through social media mining. Journal of Information & Communication Technology (JICT), 15(1).

Sindhu, I., & Shamsi, F. (2023a). Adverse use of social media by higher secondary school students: A case study on meta social network platforms. International" Journal of Academic Research for Humanities", 3(4), 205–216.

Sindhu, I., & Shamsi, F. (2023b). Prediction of IMDb movie score & movie success by using Facebook. In 2023 international multi-disciplinary conference in emerging research trends (imcert) (Vol. 1, pp. 1–5).

Wang, G., Wu, X., & Yan, J. (2024). Progressive reinforcement learning for video summarization. Information Sciences, 655, 119888.

Wang, J., Wang, W., Wang, Z., Wang, L., Feng, D., & Tan, T. (2019). Stacked memory network for video summarization. In Proceedings of the 27th ACM International Conference on Multimedia (pp. 836–844).

Wu, G., Song, S., Wang, X., & Zhang, J. (2024). Reconstructive network under contrastive graph rewards for video summarization. Expert Systems with Applications, 250, 123860.

Xie, J., Chen, X., Zhao, S., & Lu, S.-P. (2024). Video summarization via knowledge-aware multimodal deep networks. Knowledge-Based Systems, 293, 111670.

Yaliniz, G., & Ikizler-Cinbis, N. (2021). Using independently recurrent networks for reinforcement learning based unsupervised video summarization. Multimedia Tools and Applications, 80(12), 17827–17847.

Yu, Q., Yu, H., Sun, Y., Ding, D., & Jian, M. (2024). Unsupervised video summarization based on the diffusion model of feature fusion. IEEE Transactions on Computational Social Systems.

Yuan, L., Tay, F. E. H., Li, P., & Feng, J. (2019). Unsupervised video summarization with cycle-consistent adversarial LSTM networks. IEEE Transactions on Multimedia, 22(10), 2711–2722.

Yuan, Y., & Zhang, J. (2022). Unsupervised video summarization via deep reinforcement learning with shot-level semantics. IEEE Transactions on Circuits and Systems for Video Technology, 33(1), 445–456.

Zang, S.-S., Yu, H., Song, Y., & Zeng, R. (2023). Unsupervised video summarization using deep non-local video summarization networks. Neurocomputing, 519, 26–35.

Zhang, W., Wang, B., Ma, L., & Liu, W. (2019). Reconstruct and represent video content for captioning via reinforcement learning. IEEE transactions on pattern analysis and machine intelligence, 42(12), 3088–3101.

Zhang, Y., Kampffmeyer, M., Zhao, X., & Tan, M. (2019). Dtrgan: Dilated temporal relational adversarial network for video summarization. In Proceedings of the ACM Turing celebration conference-china (pp. 1–6).

Zhang, Y., Liu, Y., Zhu, P., & Kang, W. (2022). Joint reinforcement and contrastive learning for unsupervised video summarization. IEEE Signal Processing Letters, 29, 2587–2591.

Zhao, B., Li, X., & Lu, X. (2019). Property-constrained dual learning for video summarization. IEEE transactions on neural networks and learning systems, 31(10), 3989–4000. 11

Zhong, S.-H., Lin, J., Lu, J., Fares, A., & Ren, T. (2022). Deep semantic and attentive network for unsupervised video summarization. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(2), 1–21.

Zhu, Y., Zhao, W., Hua, R., & Wu, X. (2023). Topic-aware video summarization using a multimodal transformer. Pattern Recognition, 140, 109578.