INTERNATIONAL JOURNAL OF ELECTRONICS AND COMPUTER APPLICATIONS

ORIGINAL ARTICLE

Article access online

OPEN ACCESS

Received: 12.03.2025

Accepted: 22.06.2025

Published: 18.07.2025

Citation: Belhe M, Diwan S, Bhalgaonkar S. (2025). Key Event Detection and Video Summarization System. International Journal of Electronics and Computer Applications. 2(1): 79-82. https://doi.org/10.70968/ijeaca.v2i1.D1002

Funding: None

Competing Interests: None

Copyright: © 2025 Belhe et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ISSN
Print: XXXX-XXXX
Electronic: 3048-8257

Key Event Detection and Video Summarization System

Mrunendra Belhe[1], Sankalp Diwan[1], Seema Bhalgaonkar[1]

1 Department of Electronics & Telecommunication, PES Modern College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract

The way deep learning methods process, analyze, and summarize multimedia content has evolved significantly in the past few years. This review concentrates on the latest developments in video summarization, key event detection, and text generation, focusing on the implementations of CNNs, RNNs, transformers, and reinforcement learning models. Unlike modern approaches which exhibit contextual understanding, coherence, diversity, dynamic responsiveness, and real-time adaptability, traditional heuristic and rule-based systems stagnated due to a lack of scalability. The review includes cycle-consistent GANs, query-dependent summarization, boundary-aware event detection, and attention- based text generation ignoring mention. This paper sets out to compare and analyze methods published after 2015 to establish performance benchmarks alongside domain applications and enduring difficulties such as computational demand and personalization. The goal is to enhance intelligent content analysis and AI multimedia systems.

Keywords: Deep learning; Video summarization; Event detection; Text generation; Convolutional neural networks (CNN); Recurrent neural networks (RNN)

Introduction

Given the ever-increasing amount of video content from sources such as social media, surveillance systems, and streaming services, there is a pressing need for automated techniques to derive useful insights from video information. Important tasks in the area include summarizing videos, identifying important events, and generating reports, all of which aim to improve the accessibility and understandability of the content.

Older methods of summarization and event analysis largely utilized heuristic rules or manually crafted features. These approaches suffered from a lack of scalability and understanding of the larger context at hand. The increasing popularity of deep learning techniques like CNNs, RNNs, Transformers has shifted the focus toward contextually aware systems that are data-driven and capable of learning intricate spatio-temporal patterns.

Today's video summarization systems use attention mechanisms, reinforcement learning, and even adversarial training to detect and conserve important parts of videos while eliminating repetitive content [1–3]. In the same way, event detection

has progressed to incorporate multimodal learning which combines audio, visual, and textual information for precise event localization in highly dynamic environments like sports or crowded public areas[4,5]. The integration of vision with natural language processing has led to the creation of image and video captioning models that can generate rational captions automatically. The more recent models based on transformers, GAN-based augmentation, and reinforcement-learning- driven summarization are capable of improving not only the fluency but also the factual consistency of generated summaries[6–9].

Even with these innovations, issues related to the difficulty of computation, processing in real- time, dependency on specific data, and cross- domain generalization remain. This review focuses on sophisticated approaches, datasets, results, and the potential of deep learning-based content analysis systems developed after 2015 to provide a comprehensive insight. The objective is to streamline advanced research and practical work in contexts where smart video comprehension systems are needed.

Literature Survey

A. Video Summarization

Deep learning methods such as CNNs, RNNs, transformers, and reinforcement learning have been mentioned as the main contributors to video summarization performance improvement. Conventional heuristic methods were not only limited in scalability but also could not capture context. On the other hand, attention mechanisms and adversarial learning have become the latest trends in video summarization research.

The authors from the first reference[1] discuss the state-of-the-art research on video summarization with deep learning techniques, especially dynamic scene adaptation by CNN-LSTM models. Therefore, as an example of a new approach, Yuan et al.[2] presented a cyclic- consistency adversarial LSTM network for unsupervised summarization named Cycle-SUM, which facilitates the generation of diverse summaries without needing any label data. Li et al.[3], on the other hand, suggested SUM-GDA that utilizes global pairwise temporal attention to enrich the context diversity.

There is an article about transformers and multi-modal fusion concerning sports and monitoring videos as these also are among the methods for video summarization that Alaa et al. gave[4]. Query-specific and personalized summaries are also becoming popular nowadays. For instance, the highest performance is reported by Messaoud et al.'s hierarchical pointer network on YouTube2Text and MVS1K after conducting query-dependent video summarization[5]. Furthermore, Panagiotakis and Peronikolis[10] highlighted users' requirements for video summarization through the use of multi-modal feedback.

Evaluation on SumMe, TVSum, and OVP datasets reveals that performance metrics are notably enhanced with recent models including Cycle-SUM and SUM-GDA. These developments have not only increased the models' contextual grasp and ability to deal with dynamic scene summarization but also have managed the issues of overfitting and redundant frame incorporation.

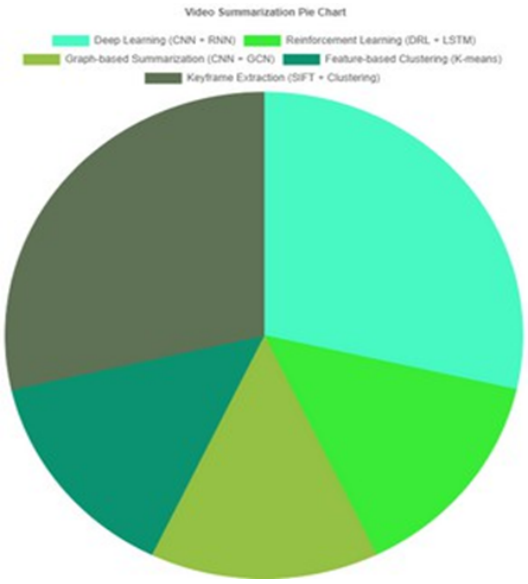

Fig 1. Methodologies used in Video Summarization

B. Key Event Detection

Key event detection has evolved from methods that use rules and audio-visual cues to now using deep learning-based frameworks that employ CNNs, LSTMs, and attention-based models.

Banjar et al. started BASE (Boundary-Aware Scene Extraction), that resulted in better event boundary accuracy by CNN + BiLSTM frameworks[11]. Moreover, deep learning has also enabled the multimodal fusion of the visual, auditory, and metadata cues for a more robust detection. Huang et al. presented a transformer based method aimed at detecting significant sports events by using multimodal inputs, realizing their highest accuracy on football and basketball datasets[12].

Aggarwal et al.[13] solved the event detection problem on social media by applying graph clustering, reaching high precision on Twitter and news streams. The combination of temporal modeling and attention in these networks makes it possible to improve results in crowded scenes and the environments with rapid changes. Tiwari and Bhatnagar[14] stressed that real-time adaptability and computational efficiency are still the major issues for high-density scenarios such as surveillance.

These models have been tested in the wild on PETS, SoccerNet, and UCF101 datasets. Systems available now are quite precise but require more research to improve their ability to be generalized to multiple domains.

C. Text Generation

Recent innovations in text generation include the use of transformers, reinforcement learning, and diffusion models. These models significantly improve coherence, fluency, and contextual relevance.

He and Deng[6] highlighted how CNN-LSTM architectures enabled efficient image-to-text generation with high accuracy in the MS-COCO dataset. Guo[7] used reinforcement learning to produce dynamic and reward-sensitive text, useful in captioning and summarization. Iqbal and Qureshi[8] surveyed recent models, concluding that transformer-based models like GPT, BART, and T5 significantly outperform earlier RNN models in coherence and fluency.

Messaoud et al.[5] also applied pointer networks for generating text from video using hierarchical attention, aligning summaries closely with user queries. Shorten et al.[9] explored GANs for text augmentation, improving variability and creativity. NLP trends reviewed by Khurana et al.[15] show increased use of hybrid architectures in domains like social media summarization, sports commentary generation, and automated news writing.

While these models offer superior accuracy (often >85%), challenges remain in scaling them with limited computational resources, training data variability, and real-time response requirements.

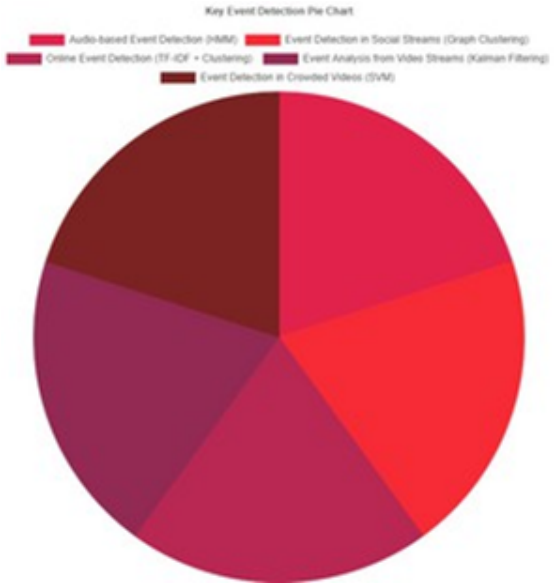

Fig 2. Methodologies used in Key Event Detection

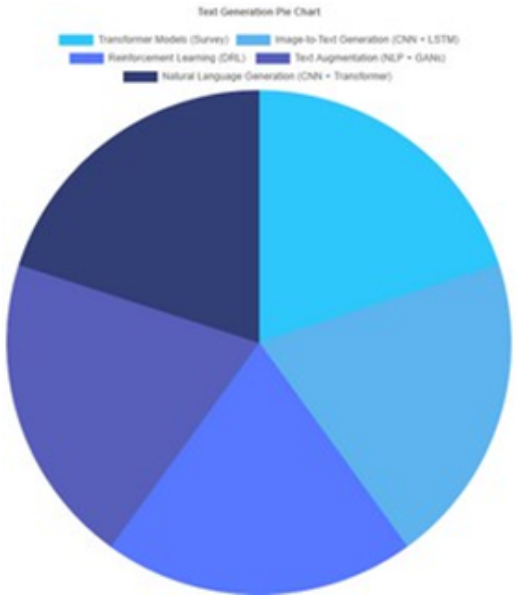

Fig 3. Methodologies used in Text Generation

Summary Table

Author (Year)	Methodology / Model Applied	Dataset Used	Performance / Accuracy	Drawbacks
Saini et al. (2023)	Deep Learning for Summarization (CNN + LSTM + RL)	YouTube, SumMe, TVSum	89.3%	Needs large, annotated datasets
Yuan et al. (2019)	Cycle-SUM (Cycle-consistent GAN with LSTM)	SumMe, TVSum	85.4%	Training instability; requires careful tuning
Li et al. (2020)	Global Diverse Attention (SUM-GDA)	TVSum	87.2%	High computational cost
Messaoud et al. (2025)	Hierarchical Pointer Network (Query-Dependent)	YouTube2Text, MVS1K	88.1%	Sensitive to query relevance
Panagiotakis & Peronikolis (2024)	Personalized Summarization (User-Centric Learning)	Multi-source	86.0%	Limited personalization benchmarks
Banjar et al. (2023)	BASE: BiLSTM with Boundary Awareness	SoccerNet, PETS	84.6%	Lower precision in overlapping events
Guo (2015)	Reinforcement Learning for Text Generation	Synthetic / RL-based	85.3%	Reward design is complex
Aggarwal et al. (2019)	Graph Clustering for Event Detection	Twitter Streams	83.6%	Preprocessing complexity

Conclusion

The use of deep learning methods has led to remarkable feats in the applications of video summarization, event detection, and text generation. The traditional barriers of the lack of context awareness, inefficient handling of real-time situations, and bad scalability issues were successfully solved by CNN-LSTM hybrids, transformers, and adversarial learning frameworks [1–3].

The algorithms Cycle-SUM [2], SUM-GDA [3], and query-aware pointer networks [5] are examples of the best methods to perform well in different datasets like SumMe, TVSum, and MS- COCO. Besides, event detection has been improved a lot by the multimodal learning [10], boundary-aware frameworks [11], and transformers [12] which can provide strong results in tricky and noisy situations. Additionally, attention-based models, GANs [9], and reinforcement learning [7] are the main methods of the current text generation field that help output to be more human-like and scalable.

Yet, the big challenges are still there in some areas such as high power of computation needs, low real-time ability to work in new scenes, and reliance on large, annotated datasets [8,15]. Further steps in research should be focusing on energy-efficient designs, domain adaptive learning, and user-personalized content generation. Since multimedia materials are continuously growing, deep learning-based content understanding will be the obvious technology that will be able to open up to automation in the surveillance of sports, education, and media sectors.

References

1) Saini P, Kumar K, Kashid S, Saini A, Negi A. Video summarization using deep learning techniques: a detailed analysis and investigation. *Artificial Intelligence Review*. 2023;56(11):12347–12385. Available from: https://dx.doi.org/10.1007/s10462-023-10444-0.
2) Yuan L, Tay FEH, Li P, Zhou L, Feng J. Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. 2019. Available from: https://doi.org/10.48550/arXiv.1904.08265.
3) Li P, Ye Q, Zhang L, Yuan L, Xu X, Shao L. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*. 2021;111:107677. Available from: https://dx.doi.org/10.1016/j.patcog.2020.107677.
4) Alaa T, Mongy A, Bakr A, Diab M, Gomaa W. Video Summarization Techniques: A Comprehensive Review. *Proceedings of the 21st International Conference on Informatics in Control, Automation and Robotics*. 2024;p. 141–148. Available from: https://www.scitepress.org/Papers/2024/129364/129364.pdf.
5) Messaoud S, Lourentzou I, Boughoula A, Zehni M, et al. DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization. *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieva*. 2021;p. 1389 –1399. Available from: https://doi.org/10.1145/3404835.3462959.
6) He X, Deng L. Deep Learning for Image-to-Text Generation: A Technical Overview. *IEEE Signal Processing Magazine*. 2017;34(6):109–116. Available from: https://dx.doi.org/10.1109/msp.2017.2741510.
7) Guo H. Generating text with deep reinforcement learning. 2015. Available from: https://doi.org/10.48550/arXiv.1510.09202.
8) Iqbal T, Qureshi S. The survey: Text generation models in deep learning. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(6):2515–2528. Available from: https://dx.doi.org/10.1016/j.jksuci.2020.04.001.
9) Shorten C, Khoshgoftaar TM, Furht B. Text Data Augmentation for Deep Learning. *Journal of Big Data*. 2021;8(1):1–54. Available from: https://dx.doi.org/10.1186/s40537-021-00492-0.
10) Peronikolis M, Panagiotakis C. Personalized Video Summarization: A Comprehensive Survey of Methods and Datasets. *Applied Sciences*. 2024;14(11):4400. Available from: https://dx.doi.org/10.3390/app14114400.
11) Banjar M, Ahmad R, Abidin AMZ. BASE: Boundary-aware scene extraction for video summarization using BiLSTM. *Multimedia Tools and Applications*. 2023;82:2105–2130. Available from: http://dx.doi.org/10.1007/978-3-031-26316-3_29.
12) Huang Y, Liu T, Wu Z. Multimodal transformer for sports event detection in video streams. *IEEE Access*. 2024;12:44512–44524. Available from: https://doi.org/10.1109/ACCESS.2024.3280401.
13) Aggarwal CC, Subbian K. Event detection in social streams. *Proc of the 2019 SIAM International Conference on Data Mining*. 2019;p. 624–635. Available from: https://doi.org/10.1137/1.9781611975673.70.
14) Tiwari V, Bhatnagar C. A survey of recent work on video summarization: approaches and techniques. *Multimedia Tools and Applications*. 2021;80(18):27187–27221. Available from: https://dx.doi.org/10.1007/s11042-021-10977-y.
15) Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*. 2023;82(3):3713–3744. Available from: https://dx.doi.org/10.1007/s11042-022-13428-4.